Evaluation guidelines for machine learning tools in the chemical sciences

Andreas Bender, Nadine Schneider, Marwin Segler, W. Patrick Walters, Ola Engkvist and Tiago Rodrigues

Abstract | Machine learning (ML) promises to tackle the grand challenges in chemistry and speed up the generation, improvement and/or ordering of research hypotheses. Despite the overarching applicability of ML workflows, one usually finds diverse evaluation study designs. The current heterogeneity in evaluation techniques and metrics leads to difficulty in (or the impossibility of) comparing and assessing the relevance of new algorithms. Ultimately, this may delay the digitalization of chemistry at scale and confuse method developers, experimentalists, reviewers and journal editors. In this Perspective, we critically discuss a set of method development and evaluation guidelines for different types of ML-based publications, emphasizing supervised learning. We provide a diverse collection of examples from various authors and disciplines in chemistry. While taking into account varying accessibility across research groups, our recommendations focus on reporting completeness and standardizing comparisons between tools. We aim to further contribute to improved ML transparency and credibility by suggesting a checklist of retro-/prospective tests and dissecting their importance. We envisage that the wide adoption and continuous update of best practices will encourage an informed use of ML on real-world problems related to the chemical sciences.

The need for new methods capable of accelerating scientific discoveries has opened the door to the re-emergence of machine learning (ML) — computer programs that use statistical models to learn patterns in data and ultimately to make predictions based on those patterns. Supported by advances in algorithms, computing hardware[1], open source libraries (for example, scikit-learn[2], PyTorch[3], TensorFlow[4] and others), data and storage capacity, ML tools are now able to identify complex patterns that have eluded expert intuition for several decades[5–7]. Although ML tools are not conceived as or expected to work as complete substitutes for expert intuition, their performance is, at the very least, comparable to that of human practitioners in some cases. For example, it can be applied to disease diagnostics[8,9], where image analysis is an established concept (although not without its caveats)[10,11]. This capacity enables ML as suitable for productive use and, most importantly, provides plausible decision-making support to some of the grand challenges in science and engineering[12]. Chemistry is also taking part in this (r)evolution[13–21]; its digitalization is ongoing and multifaceted. Some examples include the prediction of synthetic routes[22,23], the de novo design of molecules[24–28], the optimization of batteries/energy storage[29,30], the identification of biological modes of action[31,32] and protein structures[33,34], and the prediction of physicochemical properties[35–37].

Grasping the impact of ML remains difficult. However, it is reshaping how fundamental and translational research are executed in drug discovery, materials science, nanotechnology, environmental sciences and others. When engaging in ML research, one must consider three key aspects — data, representations and algorithms — that need to be reported and evaluated in detail. All of these components entail limitations, suggesting that the field must further mature. Therefore, ML algorithms should not be treated as a panacea, and evaluating these computational tools objectively is warranted. Evaluating models through a minimum set of harmonized and meaningful experiments/metrics may enable a new era in ML research. Focusing chiefly on supervised learning in this Perspective, we provide a broad and diverse collection of examples from different disciplines in chemistry and authors where our recommendations were or were not applied. We also acknowledge that several deserving studies that employed good ML practices are not mentioned.

Many established (materials science[38], biology[39,40] and microscopy[41]) and emerging (nanotechnology)[42,43] experimental fields in the chemical and physical sciences have already taken steps towards harmonization in reporting practices. These reporting standards are now in the author guidelines of reputed journals, such as those from the American Chemical Society[44]. Building on prior considerations[45], a similar discussion for ML in chemistry and an expanded checklist of proposed retrospective evaluations is included here[46]. Specifically, we discuss the relevance of specific supporting experiments and analyse 'dos and don'ts'. By recommending widely applicable guidelines, we aim to improve expectations from ML publications focused on novel concepts, benchmarks or novel discoveries in the chemical sciences. Keeping practicality, accountability and ease of execution in mind, our recommendations are divided into different categories — data/code reporting; retrospective evaluations; comparison to baseline; prospective evaluation and model interpretation — and are intended to improve quality, transferability and reuse of methods in each of the three ML study types. We envisage that implementing a guide for minimum reporting standards will help to manage expectations, inform decisions/comparisons and enable innovative solutions to the grand challenges in chemistry. ML practitioners, experimental chemists, reviewers and journal editors will thus be brought together onto common ground, closing what we believe to be an outstanding gap in communication that has been hindering further advances of ML in chemistry.

Data sets

Most of the time spent developing a ML method involves collecting and curating relevant information to answer a specific research question. Therefore, collaborating with domain experts (or being one yourself) to formulate relevant research questions is critical. Understanding the shortcomings and uncertainties will help to ensure the data are properly prepared for ML modelling. A less informed community may not appreciate the importance and time required for this step, but can perhaps easily realize that a ML model is only as valuable as its underlying data allows. Accurate predictions for previously unseen events and the generation of sensible research hypotheses are dependent on prior information and data patterns. Therefore, substantial time investment for this data preparation stage is highly recommended and ultimately necessary. In this regard, relevant end points (for example, biological activity) are often not modelled directly — sometimes because of unavailable data — but rather through a proxy that is experimentally more accessible (for example, colloidal aggregation as a proxy for generalized promiscuous behaviour in bioactivity screens)[47,48]. The potential caveats include the oversimplification of a complex problem and the assumption that the selected proxy maps onto the actual research question when that is not necessarily true[49,50]. Furthermore, uncertainty quantification is often neglected in experiments and modelling in the chemical sciences. Not accounting for experimental errors[51] can lead to their propagation and an unrealistic model performance assessment. The benefits of ML then become unclear, which is especially true when the predictive accuracy improves by only a few per cent relative to competing models[52]. Furthermore, data ranges (and the distribution of the data) are also important to consider, especially if the goal is creating a regression model. Moderate model uncertainty can lead to uninformative predictions in narrow experimental ranges. For example, a log range of two for bioactivity measurements with an error/uncertainty of 0.5 log units leads to a prediction covering half of the possible range. When data are highly clustered, and the individual clusters possess different properties, the performance of a predictive model might appear to be high. However, the model can only distinguish between the clusters present, and is unable to model the intra-cluster trends of the property considered. A high ratio of qualitative data (for example, >10 µM or <0.1 µM) is also difficult to include in a regression model. These are outstanding issues without a clearcut solution but it is important to acknowledge that, depending on the data in hand, low resolution and misleading models may be obtained.

Benchmarks

The continuous interest in developing benchmarks and accounting for uncertainty suggests that the community is engaged in finding suitable solutions to consider error estimates in comparative studies. For example, SAMPL blind challenges and Kaggle competitions, among others[53], provide viable means of comparing the accuracy of different tools on high-quality data relevant to chemistry and/or drug development. Over the past decade, the community has developed several public benchmark data sets to evaluate the performance of ML methods[54]. We encourage the use of such resources as a catalyst for developing ML tools, but note that even quality benchmark data sets have limitations and their use may not reflect the real-world performance of ML models (TABLE 1). Specifically, a benchmark cannot detail the general utility of ML models, owing to several unavoidable caveats, such as limited task design, scope, data de-contextualization and community misuse. These important aspects limit the use of benchmarks for an absolute model evaluation[55]. For example, the MOSES benchmark was established to assess, among other things, how well a generative model can match the distribution of the training data. However, trivial changes can dramatically increase the apparent diversity of a set of molecules while still matching the training data distribution[56]. Despite reports[57–59] demonstrating that the DUD-E data set is not suitable, many authors continue to use it as an ML benchmark. This misuse is discouraged, as DUD-E was originally developed to assess the performance of molecular docking methods and not ML. Rather, we recommend the use of dedicated benchmarks when available. The MoleculeNet data set, distributed with the DeepChem package, contains numerous assay artefacts and, in some cases, unrealistic dynamic ranges for the considered properties. Yet, it is commonly used for the evaluation of predictive models[60]. More recently, FS-Mol was generated as a benchmarking solution for ML tasks with small bioactivity data sets, which are ubiquitous in drug discovery[61].

Code and data availability. Data extracted from proprietary sources or public databases, such as ChEMBL or PubChem, are the cornerstone for modelling and their quality is an essential pre-requisite[51,62–64]. We recommend that data sets be made available for scrutiny and re-use, especially when they are used in studies that involve benchmarking a certain tool. It is also important to mention which transformations were applied to the original data set and why any data have been excluded. Aside from a machine-readable format (for example, *.txt or *.csv), training data should also be accompanied by source code, a brief instruction for code usage and example files, including the outputs for one or more test cases. We also recommend listing all packages and versions installed in the environment, as different versions can influence reproducibility of third-party work. All this information should be publicly available and stored in GitHub, GitLab or a similar repository, which can help mitigate the difficulty in describing complex algorithms in the Methods section. Implementing this simple practice should partly address reproducibility concerns[65] in feature engineering, employed data and algorithms. It also promotes trust and allows a swifter development of methods from an existing codebase. Models can be containerized and freely distributed for easier access to experimentalists. These models can be shared via cloud technologies, web applications or applications for local execution. Although some chemistry journals have established standards for the inclusion of source code and data in their articles, this practice is far from universal. In some instances, it might be acceptable to omit source data for proprietary reasons, but any motivations and/or conflicts of interest should be disclosed[66]. In those cases, we recommend making comparisons with other accessible tools (see below) or (also) using public benchmark data sets for a more appropriate contextualization. In this context, we have also recently elected to build two models when copyrighted data could not be disclosed[67]. Here, corporate data were presented in a descriptive manner, including data ranges and measured end points, but no structural information was released. For comparison, a twin model was built using publicly available data, allowing the readership to gauge the utility of the ML method. Admittedly, this is a delicate

Table 1 | **Selected benchmark data sets and tasks**

Data sets	Description	Limitations	Application(s)	Recommended use in ML
MOSES[157]	Metrics: SMILES validity, uniqueness, novelty, filters, fragment/scaffold similarity, similarity to nearest neighbour, internal diversity, FCD and property distribution computed for 30k generated entities Contains 4.6 million ZINC molecules Baseline models: CharRNN, VAE, AAE, JTN-VAE, LatentGAN and non-neural baselines	Strict filtering resulting in narrow property space Not representative of real/historic chemistry Charged molecules removed in curation (bias towards non-protonatable molecules)	de novo design: distribution learning	Yes Compare de novo molecules to training set, particularly property distributions
GuacaMol[54]	Metrics: Distribution Learning: SMILES validity, uniqueness, novelty, KL divergence of physicochemical properties and FCD Optimization: Suite of benchmark tasks Contains ChEMBL24 molecules Baseline models: random, best of data set, GAs, MCTS, LSTM, VAE, AAE, ORGAN	ChEMBL contains molecules from differing domains such as small molecules, antibiotics, small peptides, etc. Quality of benchmark is not contained in the codebase	de novo design: distribution learning	Yes Compare de novo molecules to training set Assess molecular optimization tasks
GDB-13[158]	Contains 975 million computationally enumerated small organic molecules	Data set size and computational expense Not representative of historic chemistry	de novo design: distribution learning	Yes For tool comparison, sample 2 billion de novo molecules and calculate coverage
MoleculeNet[60]	Collection of data sets and baseline models for several property prediction domains (quantum chemical, physics, biophysics, bioactivity)	Includes assay artifacts Properties with unrealistic dynamic ranges	Property prediction	Yes Manage enthusiasm with regards to marginal gains over 'state-of-the-art'
FS-Mol[61]	Designed for prediction tasks with <100 tested small molecules 32–5,000 examples for each of the 5,120 targets	End point is IC/EC_{50}, which is dependent on protein concentration	Bioactivity prediction in low data regimes	Yes
DUD-E[159]	Total of 22,886 ligands with binding affinity and docked pose against 102 protein targets 50 decoys per ligand	Decoys are computationally generated with no ground truth Data de-contextualization and bias issue Not designed for ML applications	Originally designed for molecular docking	No Decoys may be trivially identified by simple models due to topological differences
USPTO[160]	~4 million text-mined chemical reactions from US patent office	Successful reactions only Noise introduced by errors in text-mining Large bias to certain reaction types	Computer aided-synthesis prediction	Yes Evaluation is difficult as USPTO is not ground truth Many successful reaction pathways may exist Should assess or correct for reaction bias
GEOM[161]	33 million conformers for 430,000 small molecules organic molecules	Real conformational distributions are conditional (in vacuum, water, chloroform, bioactive, etc.), may not translate to prospective application	Conformer prediction/ generation	Yes Despite high computational accuracies, conformers and distributions are not experimentally confirmed If model includes final forcefield minimization, should include baseline with random coordinates minimized by same forcefield
PDBbind[162]	Curation of 19,443 protein-ligand complexes and binding affinities	Small data set size Sparsely populated and bias (same targets) Some discrepancies in binding mode, ignored symmetry of homodimers, poor electron density in solvent channels, etc.	Property prediction using 3D data	Yes Address bias by data set splitting via sequence identity or scaffold similarity If model is multi-input (e.g. protein features and ligand features) then ablation study removing each element should be conducted

Table 1 (cont.) | **Selected benchmark data sets and tasks**

Data sets	Description	Limitations	Application(s)	Recommended use in ML
DOCKSTRING[163]	Full matrix of 265k ligands docked against 58 targets (scores and poses) Python package to generate new data Metrics: r^2, enrichment factor, docking-based objective functions Suite of benchmark tasks	Data generated in silico	de novo design Transfer learning Virtual screening	Yes
ExCAPE[164]	Curated set of activities for 1 million compounds and 1,667 targets More than 70 million binary SAR data points	Sparse matrix. Numeric affinity values only present for a fraction of compounds.	(Sparse) bioactivity prediction	Yes

AAE, adversarial autoencoder; CharRNN, character-level recurrent neural network; EF, enrichment factor; FCD, Fréchet ChemNet distance; GA, genetic algorithm; JTN-VAE, junction tree variational autoencoder; LatentGAN, latent vector-based generative adversarial network; LSTM, long short term memory; MCTS, Monte Carlo tree search; ML, machine learning; ORGAN, objective-reinforced generative adversarial network; r^2, coefficient of determination; VAE, variational autoencoder.

situation where comparable and relevant data may not be in the public domain for all cases. For these reasons a generally applicable recommendation cannot be made.

The evaluation of new ML tools is often inconsistent because there is no consensus on which benchmarks and data sets to use. Ideally, publicly available data sets are representative of real-world data, which one can only aspire to obtain with widespread effort and continuous updates. As the field progresses, it is important that computational and experimental groups work together to develop benchmarks that accurately reflect how methods will perform when applied prospectively.

Retrospective evaluation

Retrospective evaluation studies are accessible, cost-effective and should be standard practice. Challenging the potential utility of predictive ML is needed and is carried out on data that have not taken part in model building (external test data) — but for which outputs are known (FIG. 1). To that end, cross-validations can be conducted in meaningful ways that differ according to the available training set (reviewed in REFS[68,69]). They ultimately serve as a workhorse to estimate uncertainty[70] through performance metrics (TABLE 2). To better frame the domain of applicability of a ML model, we also recommend reporting the distribution of modelled properties in the training set. For example, the property distribution for molecules in a training set might not resemble the distribution for molecules encountered in the future — known as data shift or the domain-of-applicability problem. As such, retrospective studies can serve as a baseline for the performance of a ML model. We note that retrospective studies mimic rediscovery rather than a novel discovery[71], with the latter often being the prime goal in prospective usage.

In supervised learning, where the ground truth is known, the training data are partitioned into *k* folds for cross-validation. The number of folds is usually 5 or 10 and one fold is iteratively held out to find appropriate hyperparameters[72–76]. Multiple data-partitioning schemes may be employed, such as random, (un)stratified, leave-one-out, cluster- or time-split[77]. The time-split scheme has been reported as a more realistic assessment of prospective predictiveness in drug discovery efforts[77,78], and thus might be preferred when timestamp data are available. Cross-validation remains the most appropriate surrogate for the future application of a model. However, it assumes that experimental decisions remain unchanged over time and ignores the impact of the model in decision-making.

While partitioning training data is recommended for model building, an external or evaluation set[68,79] — that is never used in training — is also useful for assessing the utility of the model on unseen data. In some instances, there might be data availability constraints that prohibit holding out a substantial fraction of data for that purpose. It is important to outline which approach was taken, but the main goal here is to discuss how to standardize the evaluation of ML models. We recommend computing an array of metrics, rather than a single, user-preferred one, and scrutinizing them as an ensemble, given their complementary information (TABLE 2).

Metrics in regression ML. In regression models, the mean absolute error (MAE) and mean squared error (MSE) — residuals methods — should be provided to understand the uniformity of prediction errors better. The MSE is related to the MAE but ascribes an exponential penalty to predictions that are increasingly further away from the expected value. Calculating both the MSE and MAE is important because the MAE can assume identical values when resulting from a small number of complete mispredictions or, alternatively, from many near-misses. These disparate cases can be distinguished using the MSE, which is more sensitive to outliers. Further, the MAE should be evaluated with care as it can be misleading (or over-optimistic) if the dynamic range of the data is small. The coefficient of determination (r^2) value should also be provided, despite the misconception that it generally describes the goodness of fit in property modelling[80,81], which is usually nonlinear. Further, model comparison based on r^2 values also requires that training/test splits be kept constant, since those metrics are normalized over the variance of the response being predicted. The r^2 value can be over-optimistic if the dynamic range of the data is large. It can also be arbitrarily near 0 or close to 1 when the model is either correct or incorrect, respectively[21]. Finally, r^2 may not be valid for true nonlinear models. For instance, $r^2 < 0$ was the result of a nonlinear relationship between variables and yield in a reaction optimization problem[21] (FIG. 1b). As an alternative, the cross-validated r^2 (also known as q^2)[81] may be calculated for an external evaluation set[82], provided that it has an identical distribution to the training set. The conditional Kendall's tau[83] — a nonparametic metric — provides information about the directionality and strength of association between random variables, and should also be provided.

Metrics in classification ML. For classifiers with balanced data sets, one should employ the number of true/false positive/negative predictions, and their derived composite metrics, such as precision, specificity, accuracy, recall/sensitivity, F1 score (a harmonic mean of recall and precision),

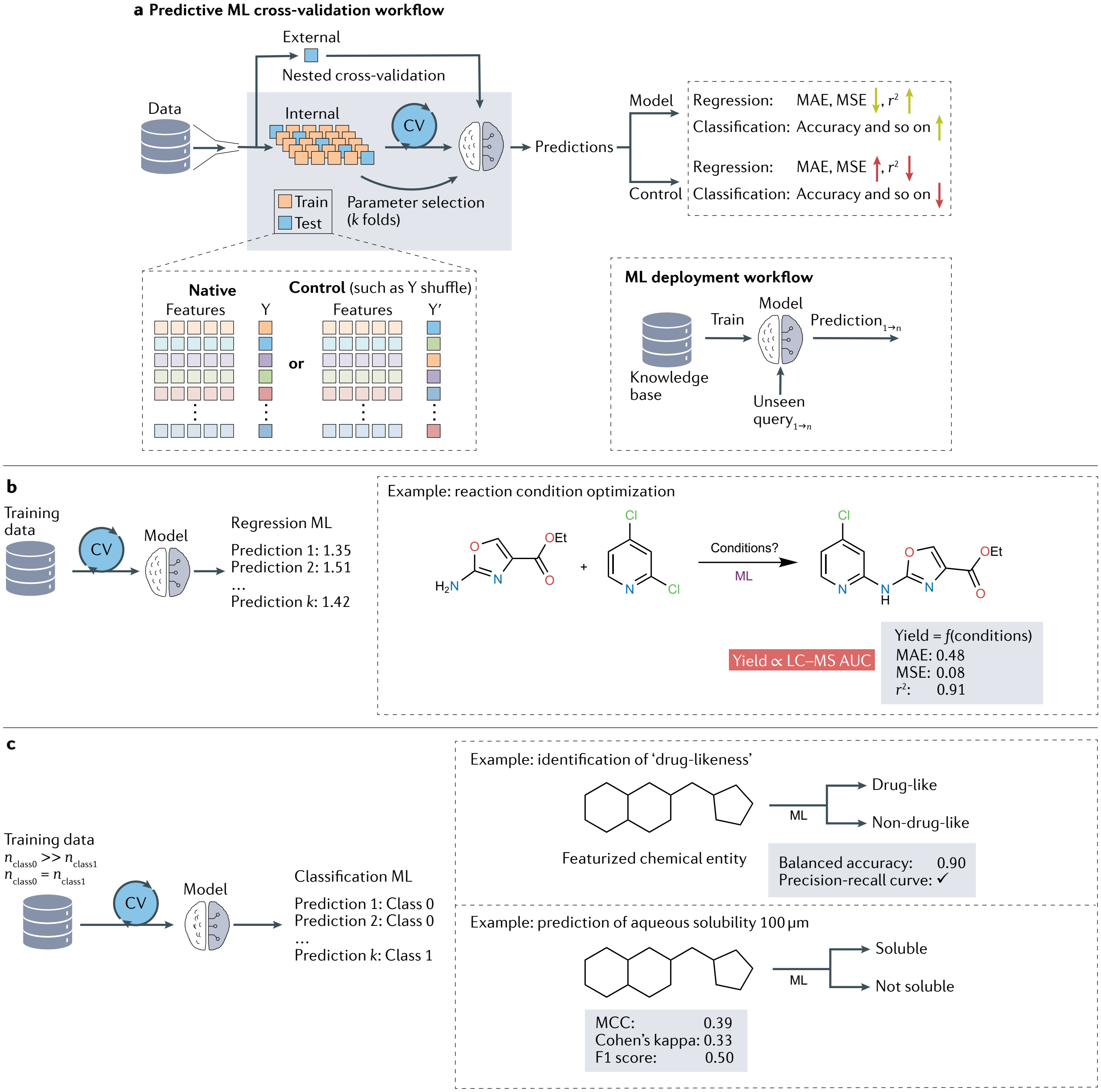


Fig. 1 | **Retrospective evaluation in ML.** **a** | Overall concept for retrospective evaluation and control computations, wherein training data are partitioned. Partitioned data are used to iteratively build machine learning (ML) models whose performance is assessed through a suite of metrics. **b** | Retrospective evaluation in regression ML, exemplified through a reaction condition optimization study[21], where the algorithm predicts reaction yields as a function of conditions. The mean absolute error (MAE) provides an arithmetic mean of absolute errors for the predicted variable (area under the curve in liquid chromatography–mass spectrometry (LC–MS AUC) traces). The mean squared error (MSE) ascribes a penalty by considering the square of the difference between experimental and predicted values. **c** | Retrospective evaluation in classification ML with imbalanced training data sets, that is, the number of training examples for one class is far superior to the number of training examples for another. Two examples are included: the classification of molecules as 'drug-like' or 'non-drug-like'[86]; and the prediction of aqueous solubility of small molecules at a concentration of 100 µM (REF.[87]). See TABLE 1 for metrics. CV, cross-validation; Y, target value; Y′, shuffled target value.

the Matthews correlation coefficient (MCC), receiver operating characteristic curve (ROC) and the corresponding area under the curve (AUC). However, imbalanced training data sets are more likely to map onto real-world scenarios and should not be evaluated by precision, specificity, accuracy and recall alone[84]. For example, a model trained on a majority class comprising 90% of all training instances (imbalanced data) is bound to afford high accuracy irrespective of the ML utility. Metrics that account for imbalanced data should be employed (TABLE 2). However, one should acknowledge that high true positive/negative rates might

Table 2 | **Recommended metrics for machine learning model evaluation**

Metric	Range	Desired	Comments and recommendations
Regression			
Mean absolute error (MAE)	[0, +∞)	0	Optimistic for mispredictions in small dynamic ranges All errors count the same
Mean squared error (MSE)	[0, +∞)	0	Together with MAE, allows the type of mispredictions (large or small) to be inferred Penalizes large errors
Coefficient of determination (r^2)	[0, 1]	1	Optimistic in large dynamic ranges Can be arbitrarily lower or higher than 0 and 1 Parametric and not suitable for nonlinear correlations
Kendall's tau	[−1, 1]	1	Informs on the directionality and strength of association between variables Non-parametric and suitable for nonlinear correlations Suitable for small sample sizes, especially if data have a monotonic relationship
Classification			
Accuracy $\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$	[0, 1]	1	Use in balanced data sets May lead to optimistic conclusions in case of imbalanced (ratio of classes) data
Precision $\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$	[0, 1]	1	Use in balanced dat asets Measures fraction of retrieved instances that are relevant to the query
Recall $\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$	[0, 1]	1	Use in balanced data sets Measures fraction of relevant instances retrieved correctly
Specificity $\frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$	[0, 1]	1	Use in balanced data sets Measures fraction of unwanted instances identified as such
F1 score $\frac{2 \times (\mathrm{Precision} \times \mathrm{Recall})}{\mathrm{Precision} + \mathrm{Recall}}$	[0, 1]	1	Use in (im)balanced data sets Harmonic mean of precision and recall with equal weights for both
Matthews correlation coefficient (MCC) $\frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$	[−1, 1]	1	Use in (im)balanced data sets Suitable for binary classification Provides a measure of correlation between observations and predictions A value of 1 represents perfect agreement; 0 and −1 correspond to random and perfect disagreement, respectively
Receiver operating characteristic area under the curve (ROC AUC)	[0, 1]	1	Use in (im)balanced data sets Suitable for binary classification A value of 1 represents a perfect classifier and <0.5 worse than random
Balanced accuracy	[0, 1]	1	Use in (im)balanced data sets Normalizes true positive and negative predictions according to samples in each class Stronger emphasis on minority class and a value of 0.5 is expected for random guesses
Precision-recall AUC	[0, 1]	1	Use in imbalanced data sets Stronger emphasis on minority class and presents the tradeoff between both metrics
Cohen's kappa	[0, 1]	1	Use in imbalanced data sets A value of 1 represents a perfect agreement and <0.5 worse than random
Binary cross-entropy	[0, +∞)	0	Use in (im)balanced data sets Aggregates information on probability distributions Goal is the approximation of a predicted to real probability (zero loss) Uncertainty increases with higher values
Confusion entropy (CEN)	[0, 1]	0	Use in multi-classification tasks Measures the uncertainty of a source of information (prediction) Uncertainty increases with higher CEN values

FN, false negative; FP, false positive; TN, true negative; TP, true positive.

not be a sufficient condition to retrieve impressive metric values, as in the case of MCC[85], and thus caution in interpretation is recommended. In fact, every performance measure that is derived from multiple other performance measures, and that simplifies the problem at first glance, will introduce its particular bias regarding input performance, measures used and data set size. Balanced accuracy and the precision-recall curve (and its AUC) are sensitive to class imbalances and should be reported, together with the F1 score, Cohen's kappa, the MCC and the ROC curve. The precision-recall curve focuses on performance for the minority class, which will give a more realistic assessment of the ML method. This has been demonstrated in distinguishing between 'drug-like' and 'non-drug-like' entities[86] and aqueous solubility[87] (FIG. 1c). It is less clear which metrics should be computed for multi-class classification, but the confusion entropy (CEN) is advisable. Based on Shannon's formalism for information gains, lower values of CEN are desirable. They reflect a higher certainty of the model, as recently demonstrated in the inference of reaction classes with attention-based neural networks[88]. The top-*N* accuracy can also be used in multi-classification problems and gauges the identification of a set of potentially viable solutions. In the case of retrosynthetic analyses, it does not account for the possibility of obtaining the same product via multiple routes. As an alternative, one can consider assessing retrosynthetic tools through the number of successfully applied templates and the overall ability to generate full synthetic routes[62], through coverage, class diversity and Jensen–Shannon divergence[89]. These strategies still require caution because nonsensical templates can provide solvable, yet meaningless routes and so continuous development of evaluation strategies is warranted[90].

Metrics in unsupervised learning. Although this Perspective is mainly focused on supervised learning, unsupervised methods are also valuable. They can be retrospectively evaluated using principles of information theory that enable comparison of two probability distributions, such as cross-entropy or even the Silhouette coefficient in clustering problems[91]. Together with deep-learning-based molecular de novo design, reinforcement learning can be powerful in navigating the chemical space to identify molecules that maximize a reward function[92]. However, defining the reward function must be done with care because reward hacking might occur and unrealistic molecules might be generated[56]. The easiest way to be alerted to potential reward hacking is to compare the log-likelihood of the generated molecules with the log-likelihood of the same molecules based on the neural network before applying reinforcement learning[93]. Reward hacking might have occurred if the log-likelihood has become substantially higher when using reinforcement learning. Another potential issue with reinforcement learning is that it might get stuck in the first identified minimum, which can be avoided by including a memory function[93].

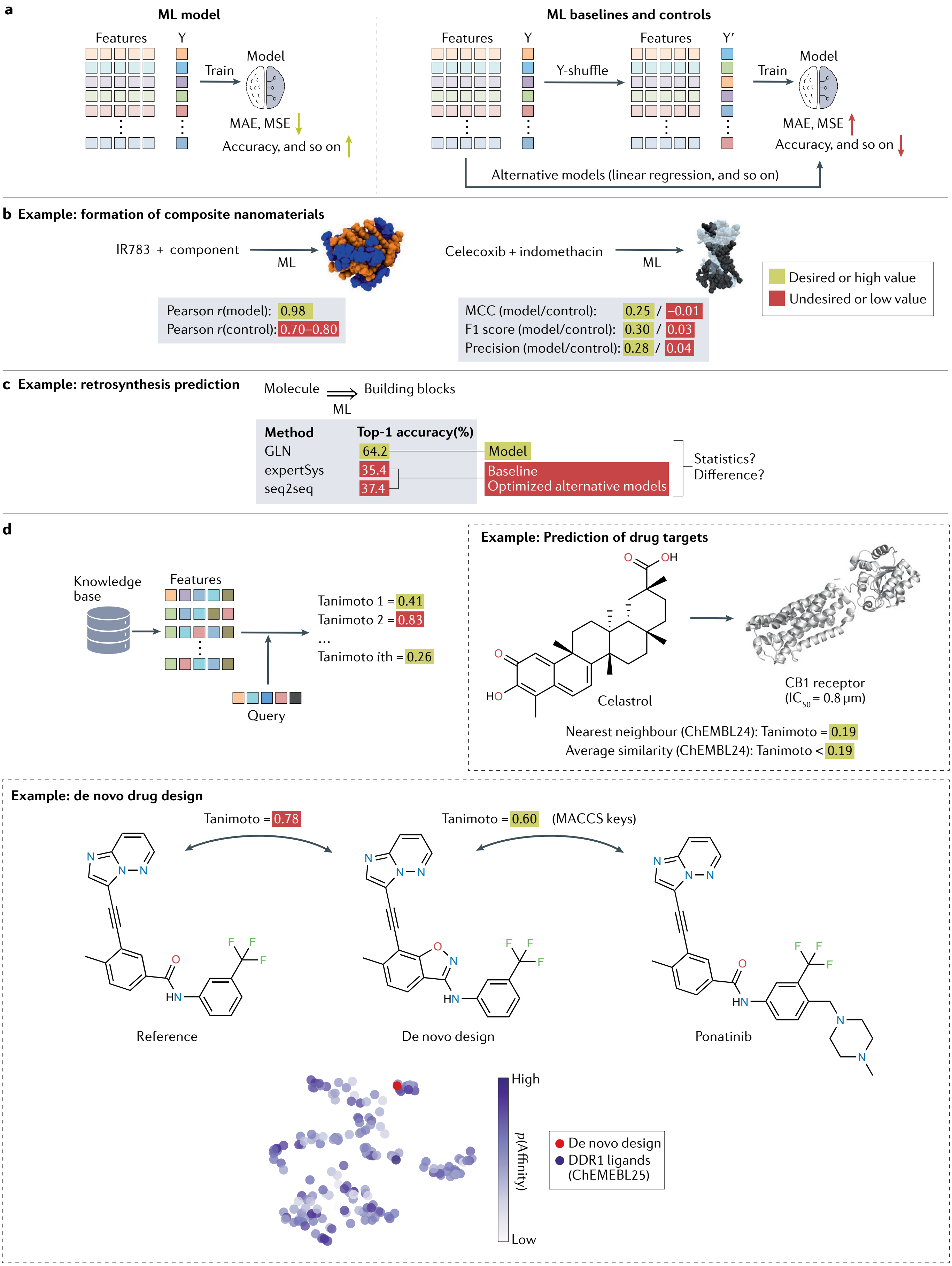
All recommended metrics can be readily computed through dedicated Python libraries, such as scikit-learn[2], SciPy[94] or PyCM[95]. Arguably, retrospective evaluation studies are widely employed in ML applications to chemical synthesis[96–98], materials design[99] and chemical biology[31], or any other research area where predictive modelling might be valuable. Cross-validations and external evaluation are important because they can help to gauge the model-wide uncertainty and whether a model is overfitted or underfitted — all these aspects affect the ML model's utility[64].

Comparison to other tools

A wide adoption of emerging technologies by target audiences is partly contingent on robust performance and the demonstration of competitive advantages over more established methods[100]. To this end, we recommend comparisons to null model(s) and established tools previously developed for the same purpose and in general use by the community. Superior performance claims should be statistically supported with *p*-values corrected for multiple comparisons, since they increase the possibility of one method randomly appearing superior to another. This correction is typically done using the Bonferroni or Holm methods[101]. On the other hand, small effect sizes can become significant for large sample sizes, which does not always translate into practical relevance. Therefore, significance values should always be reported together with effect sizes given that statistical significance does not always equal practical relevance.

Minimum comparison requirements. It is good practice to run positive and negative controls in biological assays to gauge results more directly under the same experimental conditions. The ML community can benefit from routinely adopting similar practices to support the relevance of data patterns exploited by algorithms. We recommend comparisons to mean values and prior class distributions as the minimum test. Repeated control computations provide a means of statistically (in)validating ML heuristics. In this regard, we recommend computing mean predictors (such as a linear model) or majority class classifiers (predicting the majority class for every input) to justify the use of models with higher complexity, based on their better performance. Experimental artefacts and meaningless variables may also affect ML models[102]. The randomization of target variables within the knowledge base — for example, Y-randomization/shuffling — followed by training and cross-validation tests[98] can help to challenge the hypothesis of the confounding patterns. The generated models should have lower performance than their counterparts if true patterns are disrupted in the randomization step[102] (FIG. 2a). This lower performance is of utmost importance because over-optimistic retrospective performance might lead to poor ML generalizability and inflated model expectations, which can subsequently lead to an erroneous perception of the domain of applicability. Recently, this problem was observed in a workflow for reaction yield prediction[103,104]. Substituting quantum chemical descriptors with random numbers or a 'one-hot' vector provided models of identical performance, thereby exposing an underlying pattern in the training data set resulting from the experimental design[104]. Control computations help to clarify the utility of ML tools by ascertaining the relevance of the engineered features, and interrogating whether patterns are meaningful and learned (higher generalizability), rather than artefactual and memorized (lower generalizability). In another example, the meaningfulness of substructural fingerprints and physicochemical descriptors in ML models — for a drug delivery application — was evaluated using Y-shuffling[20,37] (FIG. 2b). The relevance of these examples does not lie in the ML algorithms, but rather in their utility for prioritizing screenings, which is in part supported by these controls. Similarly, ablation studies can serve an identical purpose, as exemplified by a molecular transformer for reaction prediction[105]. In essence, model performance should be affected if a relevant feature is eliminated from the learning process[106].

Simple ML models can provide viable solutions to chemistry problems in some instances[21,37]. To assert the viability of simple and optimized ML baselines,


a
ML model
Features
Y
Train
Model
MAE, MSE ↓
Accuracy, and so on ↑
ML baselines and controls
Features
Y
Y-shuffle
Features
Y′
Train
Model
MAE, MSE ↑
Accuracy, and so on ↓
Alternative models (linear regression, and so on)
b Example: formation of composite nanomaterials
IR783 + component
ML
Pearson r(model): 0.98
Pearson r(control): 0.70–0.80
Celecoxib + indomethacin
ML
MCC (model/control): 0.25 / −0.01
F1 score (model/control): 0.30 / 0.03
Precision (model/control): 0.28 / 0.04
🟨 Desired or high value
🟥 Undesired or low value
c Example: retrosynthesis prediction
Molecule ⟹ Building blocks
ML

Method	Top-1 accuracy(%)
GLN	64.2
expertSys	35.4
seq2seq	37.4

Model
Baseline
Optimized alternative models
Statistics?
Difference?
d
Knowledge base
Features
Query
Tanimoto 1 = 0.41
Tanimoto 2 = 0.83
…
Tanimoto ith = 0.26
Example: Prediction of drug targets
O
OH
O
HO
Celastrol
CB1 receptor
(IC₅₀ = 0.8 μm)
Nearest neighbour (ChEMBL24): Tanimoto = 0.19
Average similarity (ChEMBL24): Tanimoto < 0.19
Example: de novo drug design
Tanimoto = 0.78
Tanimoto = 0.60
(MACCS keys)
N
N–N
F
F
F
O
HN
Reference
O
N
De novo design
N
Ponatinib
High
p(Affinity)
Low
● De novo design
● DDR1 ligands (ChEMBL25)

◀ Fig. 2 | **Comparison of ML utility relative to competing methods. a** | The evaluation of machine learning (ML) with competing methods and controls. Shuffling the target variable (Y-shuffle or randomization) yields an alternative model that can work as control. **b** | Control models help to identify confounding variables and artefacts. Y-randomization/shuffling (control) desirably leads to lower model performance (lower r^2, F1 score and precision values). The examples show those metrics for models predicting the formation of two-component nanomaterials[20,37]. In the computational model, IR783 is depicted in blue and indocyanine is depicted in orange. Similarly, celecoxib is depicted in black and indomethacin in grey. **c** | Comparison of a retrosynthesis prediction tool with optimized models or baselines[115,165]. Comparisons should be supported by statistical analyses whenever possible. The top-1 accuracy (number of correct classes predicted in the top 1% of predictions, sorted by probability) is provided. **d** | Comparison of ML model to non-learning approaches (for example, rule- or similarity-based searches or any meaningful in silico technology) to test the added value of ML. A drug target was predicted for celasterol (left), which presents low Tanimoto similarity (Morgan fingerprints radius 2, bits = 2,048) to ChEMBL molecules used for training[122]. In the second example (right), the designed molecule[100,125] shares structural features with an entity in the training set/literature (Tanimoto = 0.78; MACSS keys) as well as identical 2D pharmacophores[147] relative to active discoidin domain receptor 1 ligands. UMAP[166] and ChEMBL25 (the version in the publication date) were used for projection. Dots correspond to ligands and colour indicates the *p*(Affinity) (–logIC/EC$_{50}$ or $K_{\mathrm{D}}/_{\mathrm{i}}$) of the molecule as reported in ChEMBL25. GLN, graph logic network. In panel **b**, the molecular dynamics simulations map for IR783 with indocyanine (blue and orange) was adapted from REF.[20], Springer Nature Limited; and that for celecoxib and indomethacin (black and grey) was adapted from REF.[37], Springer Nature Limited.

one should compute them from the same training data[107] to contextualize and make usefulness comparisons. Eventually, those comparisons can attest that the implemented software outperforms the computationally cheaper and potentially more interpretable approaches. However, this might not always be the case. Multivariate linear regression is of high value in different settings[108–111] and enforcing 'regularization' — as in lasso, elastic net, Bayesian regression and others — can afford simpler yet equally reasonable statistical models[112,113]. Baselines need to be more complex to afford meaningful benchmarks in other instances, such as retrosynthetic analyses[114,115] (FIG. 2c) or the prediction of reaction recipes[116]. Overall, a side-by-side comparison with subsequent statistical analysis is recommended. These studies aim to reinforce the practical utility of any given approach in close-to-real-world scenarios (for example, the discovery of new materials)[37].

Desirable additional comparisons. Before the recent surge of ML algorithms[117], rule-, distance-, similarity-based[118] and non-ML methods such as molecular docking[119] have shown utility. Moreover, these methods were designed to solve the same chemical problems[120], and their practicality has led to wide adoption and extensive experimental validation. As such, comparing outputs from different learning and non-learning approaches can help to identify competitive advantages, caveats and opportunities for future research (FIG. 2d). Although not all research groups have the means to run exhaustive comparisons to multiple methods, it is good practice to consider similarity searches as a baseline given their easier access, simplicity and relevance. Similarity searches are not outdated and there is continued interest in designing universal molecular fingerprint systems[121]. For example, would it be possible to identify the same macromolecular drug targets (such as proteins) for a given small molecule following a similarity search, rather than using an ML method[122]? The answer depends on many variables, such as the amount of available information and its quality, monitored end points and others. In general, the interested reader should not be left wondering about the relevance of the ML used relative to more established (and complementary) technologies. Despite being slow as a result of enumerating all possible pairwise relationships, similarity searches are amenable to different settings, such as retrosynthetic planning[120], prediction of physicochemical properties[36], and controls in de novo design of drugs and materials[26,123,124]. In addition, without explicitly learning patterns, they can provide a fair and straightforward contextualization relative to what a computer should obtain. Such comparisons based on molecular fingerprints are occasionally overlooked, which may lead to confirmation biases and less motivated claims of novelty[100,125] (FIG. 2d).

In this section, our recommendations are meant to be inclusive and accessible. If a study focuses on developing a new method, then a simple baseline (for example, similarity search) might be sufficient for a preliminary proof-of-concept. Most importantly, this strategy applies to almost any subfield of chemistry in which alternative technologies have been developed to solve numerous challenges. Prospective evaluation is necessary if a ML model is deployed for discovery, as in the recent case of piperlongumine for TRPV2 modulation[32], or de novo design[126]. Such studies may be time-consuming, and the budget required to conduct chemical and biological experiments is prohibitive for many research teams. However, unfounded claims of model utility are more likely to be made[49,50] if said discriminative/generative ML models are not properly evaluated with bench experiments. We caution that the opposite is equally true, such that exaggerated claims should be avoided if only limited experimental evaluation is available.

Prospective evaluation

While retrospective evaluation offers an important means of testing the potential utility of a ML workflow, lower-than-expected performances are observed in prospective, real-world applications. These underwhelming performances can be a result of deploying a model outside its applicability domain. For example, the algorithm may not fully grasp some nuanced patterns owing to heuristics or training data limitations. Prospective studies do not evaluate ML models but the whole process[127], which includes the data preparation, featurization, hyperparameter and experiment selection routines.

As ML tools strive to speed up the pace for making novel discoveries, we advocate that an expanded set of recommendations should be implemented to engage wet-lab researchers in this digital chemistry era. We recognize that engaging wet-lab researchers over long evaluation campaigns is not a trivial task. However, ML may become a more integral part of the discovery sciences if ML evaluation protocols are better harmonized and their associated expectations are managed. Automation with robots and the possibility of human-in-loop configurations can have important roles here. These automation and human–machine strategies will promote the acquisition of high-quality data and enable a certain degree of prospective deployment and trust. However, experimental design is in the hands of humans (even if only the 'design of the experimental design'), so expert input and a suitable man–machine interaction will continue to be key. We argue that prospective examples must be diverse, such as covering disparate application scenarios and regions of search space. They should also challenge and probe different aspects of the ML tool while maintaining current scientific interest. Above all, prospective examples should be carefully designed to provide a meaningful assessment of the process and extract conclusions on the capabilities and limitations that may have passed unnoticed in retrospective evaluations, but are still pertinent for productive deployment. Prospective evaluation in drug discovery usually requires multiple replicates, including the use of orthogonal assay technologies. For example, single concentration primary screens might be appropriate as a crude evaluation of bioactivity. These must then be scaled to full dose–response curves using different ligand-binding technologies (for example, surface plasmon resonance, radioligand displacement assay and saturation transfer difference nuclear magnetic resonance (NMR)) and functional assays (for example, evaluating cellular activity).

Rediscovery test. To follow a 'divide-and-conquer' evaluation strategy and attain reporting completeness, we argue that examples with tiered levels of processing difficulty (but that are all equally relevant) are warranted. One of the examples should be performed in semi-controlled settings to assure the research team that a positive outcome is achievable by competing means — either in silico or expert chemical intuition. With such a prospective test, the goal is to limit unknown variables that may unpredictably contribute to the targeted outcome and have the ML tool forecast outcomes for previously unseen yet feasible examples. This example can be configured as a rediscovery or pseudo-prospective test, in which an outcome has been reported elsewhere but is unknown to the ML tool. For example, the synthesis of a previously reported imidazopyridine was optimized using a ML routine[21]. Likewise, a curiosity algorithm was used to explore a reactivity space and re-discover optimized methods for Suzuki–Miyaura cross-couplings[17] (FIG. 3a). In these two examples, the ML method should contribute the preliminary proof-of-concept by assessing whether the overall process is appropriate for the problems it is meant to solve.

Discovery test. Subsequently, at least one test should tackle a genuine research question — assisting in prioritizing experiments and contributing to augmenting knowledge in a specific domain — as in the case of discovering new chemical reactions[17] (FIG. 3b), clarifying phase-transition states in complex materials[128] or discovering coordination rules in inorganic chemistry[129]. Although completion of a challenging test with a positive outcome is desirable, the occasional inability to do so does not statistically invalidate the modelling process, especially if the proof-of-concept showcases how potentially transferable the model is. More importantly, failures can hint at knowledge gaps, augment intuition and enrich the community with high-quality, negative data; all of which are relatively underrated[130,131].

Negative example test. One might also argue that a final test is required, structured as a 'failure-by-design' in which experiments are designed to disprove a research hypothesis rather than confirm it — the classical scientific method. Negative data are largely underexploited, considering the difficulty in extracting said examples from the literature[64], and little is known about false-negative rates in the reported ML workflows. Focusing on accurately identifying positive outcomes is irrevocably the priority, but only informs on a model's applicability on one end. This focus does not provide understanding of whether the ML tool is biased towards certain (and desirable) outcomes or is truly generalizable and equally capable of confidently recognizing experiments that will probably fail. In that regard, the design and execution of negative tests mimics using negative controls in wet-lab experiments, and might provide an integrated vantage point on the capabilities of the ML algorithm. A couple of relevant examples are the design of inactive chemical matter[132] (FIG. 3c) and the prioritization of non-performant reaction conditions[21].

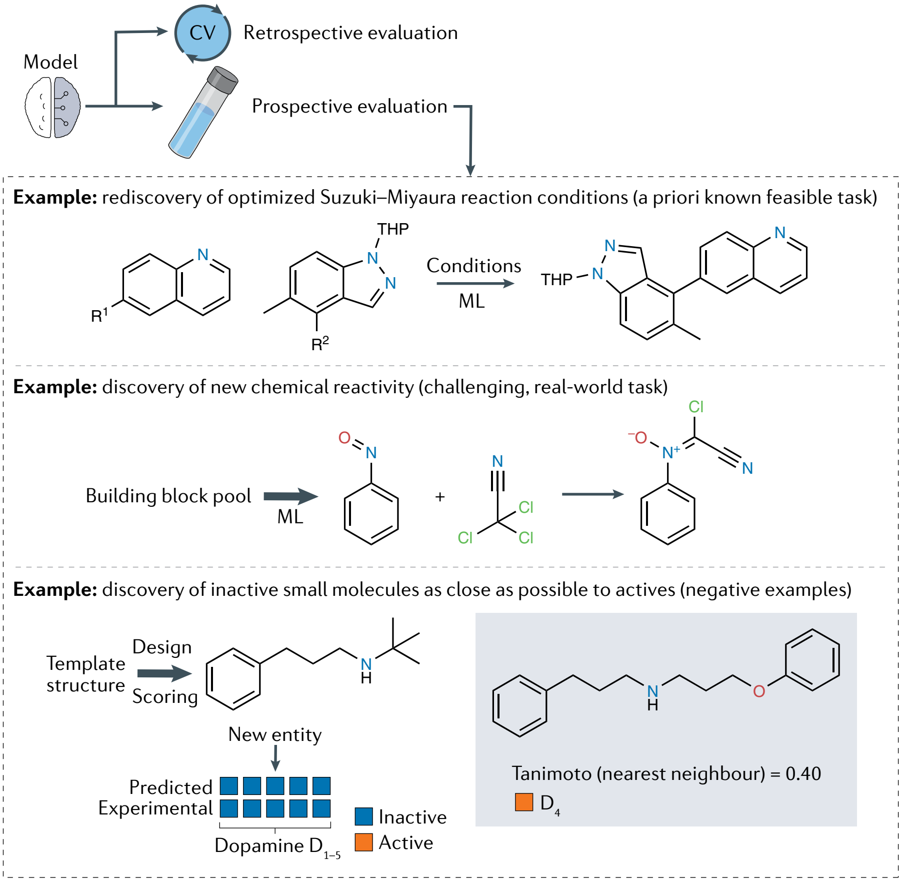

Model
CV
Retrospective evaluation
Prospective evaluation
Example: rediscovery of optimized Suzuki–Miyaura reaction conditions (a priori known feasible task)
N
R1
THP
N
N
R2
Conditions
ML
THP
N
N
N
Example: discovery of new chemical reactivity (challenging, real-world task)
Building block pool
ML
O
N
+
N
Cl
Cl
Cl
Cl
−O
N+
N
Example: discovery of inactive small molecules as close as possible to actives (negative examples)
Template structure
Design
Scoring
N
H
New entity
Predicted
Experimental
Dopamine D1–5
Inactive
Active
N
H
O
Tanimoto (nearest neighbour) = 0.40
D4

Fig. 3 | **Prospective evaluation of ML models.** A minimum of three different experiments should be conducted to clarify the ML model's applicability and hint at future development directions. The bespoke tool should be employed in a semi-controlled setting, to solve a problem that is yet unknown to the machine, but whose solution is attainable by competing methods (first example). Such example will enquire about the potential utility of the machine learning (ML) workflow in real-world scenarios. Next, a challenging test is warranted to solve a previously untapped research question and ideally to generate new and interpretable knowledge on a topic of current interest (second example). Finally, the computational tool ought to be employed in the prioritization of a negative experiment, akin to the execution of negative controls (third example). Dopamine $D_{1–5}$, dopamine receptors 1–5; THP, tetrahydropyranyl.

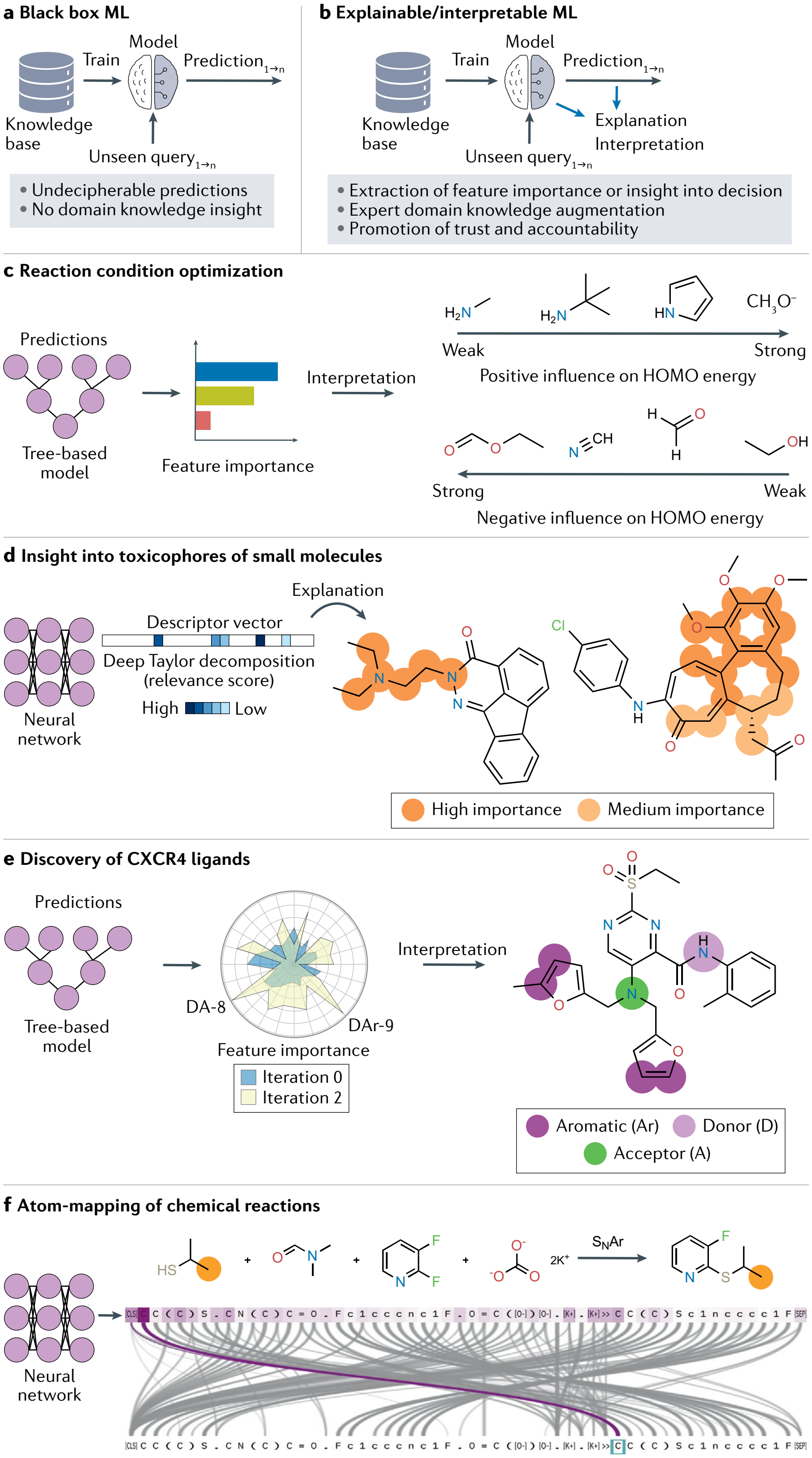

a Black box ML
Model
Train
Prediction$_{1\rightarrow n}$
Knowledge base
Unseen query$_{1\rightarrow n}$
• Undecipherable predictions
• No domain knowledge insight
b Explainable/interpretable ML
Model
Train
Prediction$_{1\rightarrow n}$
Knowledge base
Explanation
Interpretation
Unseen query$_{1\rightarrow n}$
• Extraction of feature importance or insight into decision
• Expert domain knowledge augmentation
• Promotion of trust and accountability
c Reaction condition optimization
Predictions
Tree-based model
Feature importance
Interpretation
H_2N
H_2N
HN
CH_3O^-
Weak
Strong
Positive influence on HOMO energy
O
O
N
CH
H
O
H
OH
Strong
Weak
Negative influence on HOMO energy
d Insight into toxicophores of small molecules
Explanation
Descriptor vector
Deep Taylor decomposition (relevance score)
High
Low
Neural network
N
N
N
N
O
Cl
N
H
O
O
O
O
O
● High importance
● Medium importance
e Discovery of CXCR4 ligands
Predictions
Tree-based model
DA-8
DAr-9
Feature importance
Iteration 0
Iteration 2
Interpretation
O
O
S
N
N
H
N
O
N
O
O
● Aromatic (Ar)
● Donor (D)
● Acceptor (A)
f Atom-mapping of chemical reactions
HS
+
O
N
+
F
N
F
+
O$^-$
$^-$O
O
$2K^+$
S_NAr
F
N
S
Neural network
[CLS] C C (C) S . C N (C) C = O . F c 1 c c c n c 1 F . O = C ([O-]) [O-] . [K+] . [K+] >> C C (C) S c 1 n c c c c 1 F [SEP]
[CLS] C C (C) S . C N (C) C = O . F c 1 c c c n c 1 F . O = C ([O-]) [O-] . [K+] . [K+] >> C C (C) S c 1 n c c c c 1 F [SEP]

Fig. 4 | **ML for knowledge augmentation.** **a** | Black box algorithms provide predictions without yielding domain insight. **b** | Explainable and interpretable algorithms provide insight into domain knowledge and promote trust in end-users, but interpretations are linked to the employed algorithms and descriptors. **c** | Example of interpretable machine learning (ML) wherein motifs are positively or negatively correlated with energy levels of small molecules. **d** | Identification of potential toxicophores in small molecules according to a deep Taylor decomposition routine. **e** | Identification of small molecules active against the CXC-chemokine receptor 4 (CXCR4), as a function of topological descriptor correlations. DA-8 is the donor–acceptor correlation with a distance of 8 bonds, DAr-9 is the donor–aromatic correlation with a distance of 9 bonds. **f** | Annotation of atom rearrangements in chemical reactions as a function of the grammar in SMILES strings. HOMO, highest occupied molecular orbital. The atom mapping in panel **f** was generated using RXNMapper and adapted with permission of AAAS from REF.[149]. © The Authors, some rights reserved; exclusive licensee AAAS. Distributed under a CC BY-NC 4.0 licence.

In summary, it is unrealistic to expect that any given ML tool — or computational technology — will be useful in all possible cases that it was initially designed for. Prospective assessment aims to narrow down an initially ill-perceived domain of applicability and thus bestows a necessary step towards ML evaluation and improvement. Most importantly, we advise that real-world utility claims should not be solely linked to success on a limited number of prospective tests as, arguably, all original research manuscripts will only be able to present a few examples. Those tests should generally be treated as preliminary proof-of-concept rather than robust validation for which statistical support is advised. To this end, we must strive to provide an integrated view of the strengths and deficiencies of the predictive process by including a range of tractable controls, statistical analyses and uncertainty estimation.

Model reasoning

If an ML tool presents substantial advantages over competing methods, we encourage explaining and/or interpreting these to confirm or extract new insights that could push the boundaries of chemistry knowledge (FIG. 4a,b). Explainability (a means of understanding the ML model) and interpretability (an approximation to intelligible information) are not new concepts in the realm of ML, as shown by numerous applications[133]. However, both have remained underappreciated in the chemical sciences until recently. The realization that ML may find immediate applicability as a decision-making assistant, rather than a fully-fledged substitute for brainpower, was the catalyst needed to start embracing the concept. We recognize that explaining and/or interpreting models[134], such as tracking decision paths, understanding the importance of individual features and local perturbations for the probabilistic output, employing data visualization techniques and others, can be effective in the chemical sciences.

Interpretation challenge. Given that experimental chemistry represents an important fraction of the ML audience, the wide adoption and continued development of algorithms that extract knowledge from statistical models is important. Distilling information from the ML model and correlating it with physical phenomena is not straightforward, as it depends heavily on the molecular representations, algorithms and data used[135]. For meaningful interpretations, descriptors must appropriately represent the physical phenomena, which we acknowledge is a challenge in itself. Linking correlation to causality is the prime goal here and the identification of a relevant feature or chemical motif might generally mean one of three things. First, the correlation may be spurious owing to data set/descriptor limitations. For example, particular data set biases, such as the presence of many analogues of a particular chemical series, can lead to spurious model correlations and hence feature interpretations. Second, the correlation may be true but it could result from an indirect observation, such as predicting a lipophilic motif versus lipophilicity. Or third, the correlation is indeed meaningful. This latter case is desirable, but knowing which one is in hand can inform ML development. To that end, benchmarks with clearly defined correlations between structural motifs and end points might find utility in the evaluation of workflows, as shown for the interpretation of a QSAR model[136].

Methods and approaches. It is possible to employ 'off-the-shelf' methods (for example, Shapley additive explanations (SHAP)[137] and local interpretable model-agnostic explanations (LIME)[138]) to interpret ML models on a global (model-wide) or individual (a specific prediction) scale. These methods build surrogate linear models to extract important variables and inform on their extent and direction (in favour of or against the desired outcome). This ability to infer directionality is a distinguishing factor that is not found in ensemble methods, such as random forests. Still, they present an important caveat as model explanation packages approximate the original ML model decisions. Thus, the outputs are in reference to the ML model approximations and are not model itself. Alternatively, tailored data visualization algorithms can also afford an intuitive way to highlight certain aspects of the ML decision process[21,139–141], irrespective of their correctness. One might thus argue that future ML research will increasingly focus on algorithms that provide insights aligned with established chemistry intuition or (ideally) capable of augmenting it. In one instance, the in-built feature importance detection in tree-based ML allowed the extraction of motifs associated with improved control over water–octanol partitioning coefficients, and moieties correlated with the energy levels of organic molecules[142] (FIG. 4c). In other examples, artificial neural networks were deconvoluted by a deep Taylor decomposition method to reveal toxicophores in small molecules[143] (FIG. 4d), decision trees were employed to understand enantiomeric excess values in an asymmetric hydrogenation[144], and graph representations were dissected to reveal reaction centres[145]. One should take care not to fall prey to features spuriously associated with an end point of interest, such as data set biases that include many analogues of a scaffold.

Although typical ML models often provide a static vantage point on a specific problem, one might also consider iteratively adding information to the training set, as it becomes available. This concept, known as active learning, can improve models if the new data dynamically update feature importance values[146] (FIG. 4e). Those updates should afford better approximations between predictions and the ground truth, thereby compressing the search space into a more feasible set of experiments. By using this strategy and a visualization workflow, superior predictive accuracy and decreased uncertainty were obtained in the search for CXC-chemokine receptor 4 modulators. The importance values of topological pharmacophore descriptors[147], physicochemical properties and Morgan fingerprint bits (RDKit) evolved. As a result, single-digit micromolar inhibitors were discovered in just two iterations[148].

Table 3 | **Summary of required evaluation studies according to the manuscript type**

Manuscript content	Data set availability	Code availability	Retro-spective evaluation	Compar-ison to baseline	Prospective evaluation	Model inter-pretation
Novel concept	+	+	++	++	NR	+
Benchmark	++	++	++	++	NR	+
Discovery	++[a]	++[a]	NR[b]	NR[b]	++	+

+, very beneficial; ++, essential; NR, not required. [a]Omission must be properly justified. [b]If reported in previous study.

We expect to witness cutting-edge research pushing the frontier of explainable ML in the near future, as exemplified in the recent mapping of atoms in chemical reactions[149] (FIG. 4f). These initiatives will ultimately promote trust in end-users, augment intuition and advance the digitalization of chemistry. Although explaining ML models represents a step forward, current limitations in the toolkit must be addressed to ensure that explanation is a standardized procedure and that interpretability becomes a common reality. Until more robust pipelines emerge, we recommend performing repeated analyses, given the stochasticity of the available algorithms, and interpreting outputs as research hypotheses worthy of experimental investigation rather than ground truths.

Outlook

We expect ML to speed up progress in the chemical sciences; either fully independent from human intervention[150] or as an assistant to expert reasoning. As the number of methods developed to tackle a particular challenge increases, harmonized reporting procedures will be necessary in the short term to support recent efforts towards the democratization of said technologies[151]. Streamlining and regulating evaluations in different types of ML reports (TABLE 3) with a minimum set of supporting experiments will benefit both developers and end-users. Developers will gain a means of appropriately benchmarking their software, whereas end-users will have access to transparent information to make informed decisions about the best tools for their needs.

In our internal discussions, there were some divergent opinions, which probably reflect our individual backgrounds, experiences and priorities. This variability is not only expected, but also healthy and suggests that engagement by the community is warranted. In particular, how to standardize and objectively evaluate the outcomes of Turing tests[152] or machine-versus-human competitions is currently unclear to us, given that the number of individuals involved and their cultural and scientific backgrounds can affect comparisons. Nevertheless, such tests can provide interesting benchmarks and have been used[21,22,153–156]. They may promote trust and become a valuable way to manage expectations among ML users, who might unrealistically expect 'super-human' performance by computational tools on a routine basis. We agree that ML research must be based on consistency and accountability towards which interpretable/explainable methods might be of additional value. So far, no firm recommendations can be provided and much research remains to be done in this regard to improve toolkits and workflows. However, we do expect, given the continuous developments in explainable methods, that these will improve the establishment of relevant causality and empower chemists to make verifiable machine-assisted decisions.

While writing this Perspective, other researchers independently published a best-practice checklist for ML development[46]. The recommendations we provide here resonate with that checklist. Similarly, we must conclude that retrospective evaluation, code and data set availability are important, yet still poorly reported. We go one step further and suggest that each manuscript type requires different evaluation studies. For example, a thorough prospective evaluation might be required in some cases but not in others.

We argue that separating evaluation from true ML prospective validation is key but might be confounded. We reached a consensus that proper validation is virtually impossible in a short time frame and can only be achieved over multiple iterations, across numerous studies and years. With that in mind, here we have structured and discussed a list of recommended retro-/prospective evaluation studies for ML in the chemical sciences (TABLE 3 and Supplementary Tables 1,2). We reiterate that there are no universal recipes and that each ML implementation may require specific investigations and controls, which imposes a degree of flexibility in the proposed studies. Taken together, we believe this is a much-needed step for the chemical sciences community and we urge the adoption of ML evaluation guidelines by all intervening stakeholders. As the field matures, we also expect engaging discussions and a continuous update to guidelines, which might spur future investigations and more rigorous ML reports.

Andreas Bender[1], Nadine Schneider[2], Marwin Segler[3], W. Patrick Walters[4], Ola Engkvist[5,6] and Tiago Rodrigues[7]✉

[1]Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK.

[2]Novartis Institutes for BioMedical Research, Novartis Pharma, Novartis Campus, Basel, Switzerland.

[3]Microsoft Research Cambridge, Cambridge, UK.

[4]Relay Therapeutics, Cambridge, MA, USA.

[5]Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden.

[6]Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden.

[7]Research Institute for Medicines (iMed), Faculdade de Farmácia, Universidade de Lisboa, Lisbon, Portugal.

✉e-mail: tiago.rodrigues@ff.ulisboa.pt

https://doi.org/10.1038/s41570-022-00391-9

Published online 24 May 2022

1. Gawehn, E., Hiss, J. A., Brown, J. B. & Schneider, G. Advancing drug discovery via GPU-based deep learning. *Expert Opin. Drug Discov.* **13**, 579–582 (2018).
2. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
3. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8024–8035 (2019).
4. Abadi, M. et al. in *Proc. 12th USENIX Conf. Operating Syst. Design Implement.* 265–283 (USENIX Association, 2016).
5. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
6. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
7. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
8. Myszczynska, M. A. et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat. Rev. Neurol.* **16**, 440–456 (2020).
9. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun.* **11**, 3923 (2020).
10. Yi, P. H., Malone, P., Lin, C. T. & Filice, R. W. Deep learning algorithms for interpretation of upper extremity radiographs: laterality and technologist initial labels as confounding factors. *Am. J. Roentgenol.* **218**, 714–715 (2021).
11. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digital Health* **1**, e271–e297 (2019).
12. Tschandl, P. et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
13. de Almeida, A. F., Moreira, R. & Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **3**, 589–604 (2019).
14. Gromski, P. S., Henson, A. B., Granda, J. M. & Cronin, L. How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.* **3**, 119–128 (2019).
15. Schneider, P. et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
16. Strieth-Kalthoff, F., Sandfort, F., Segler, M. H. S. & Glorius, F. Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chem. Soc. Rev.* **49**, 6154–6168 (2020).
17. Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
18. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
19. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
20. Shamay, Y. et al. Quantitative self-assembly prediction yields targeted nanomedicines. *Nat. Mater.* **17**, 361–368 (2018).
21. Reker, D., Hoyt, E. A., Bernardes, G. J. L. & Rodrigues, T. Adaptive optimization of chemical reactions with minimal experimental information. *Cell Rep. Phys. Sci.* **1**, 100247 (2020).
22. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
23. Schreck, J. S., Coley, C. W. & Bishop, K. J. M. Learning retrosynthetic planning through simulated experience. *ACS Cent. Sci.* **5**, 970–981 (2019).
24. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
25. Tu, K. H. et al. Machine learning predictions of block copolymer self-assembly. *Adv. Mater.* **32**, 2005713 (2020).
26. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2**, 171–180 (2020).
27. Yao, Z. et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* **3**, 76–86 (2021).
28. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
29. Gao, T. & Lu, W. Machine learning toward advanced energy storage devices and systems. *iScience* **24**, 101936 (2021).
30. Severson, K. A. et al. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **4**, 383–391 (2019).
31. Rodrigues, T. et al. Machine intelligence decrypts β-lapachone as an allosteric 5-lipoxygenase inhibitor. *Chem. Sci.* **9**, 6899–6903 (2018).
32. Conde, J. et al. Allosteric antagonist modulation of TRPV2 by piperlongumine impairs glioblastoma progression. *ACS Cent. Sci.* **7**, 868–881 (2021).
33. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
34. Wang, T. et al. Improved fragment sampling for ab initio protein structure prediction using deep neural networks. *Nat. Mach. Intell.* **1**, 347–355 (2019).
35. Tian, Y. et al. Determining multi-component phase diagrams with desired characteristics using active learning. *Adv. Sci.* **8**, 2003165 (2020).
36. Reker, D., Bernardes, G. J. L. & Rodrigues, T. Computational advances in combating colloidal aggregation in drug discovery. *Nat. Chem.* **11**, 402–418 (2019).
37. Reker, D. et al. Computationally guided high-throughput design of self-assembling drug nanoparticles. *Nat. Nanotech.* **16**, 725–733 (2021).
38. Timmreck, R. et al. Characterization of tandem organic solar cells. *Nat. Photon.* **9**, 478–479 (2015).
39. Jones, D. T. Setting the standards for machine learning in biology. *Nat. Rev. Mol. Cell Biol.* **20**, 659–660 (2019).
40. Walsh, I. et al. DOME: recommendations for supervised machine learning validation in biology. *Nat. Mater.* **18**, 1122–1127 (2021).
41. Horstmeyer, R., Heintzmann, R., Popescu, G., Waller, L. & Yang, C. Standardizing the resolution claims for coherent microscopy. *Nat. Photon.* **10**, 68–71 (2016).
42. Faria, M. et al. Minimum information reporting in bio–nano experimental literature. *Nat. Nanotech.* **13**, 777–785 (2018).
43. Miernicki, M., Hofmann, T., Eisenberger, I., Kammer, F. V. D. & Praetorius, A. Legal and practical challenges in classifying nanomaterials according to regulatory definitions. *Nat. Nanotech.* **14**, 208–216 (2019).
44. Aldrich, C. et al. The ecstasy and agony of assay interference compounds. *ACS Cent. Sci.* **3**, 143–147 (2017).
45. Jain, A. N. & Nicholls, A. Recommendations for evaluation of computational methods. *J. Computer Aided Mol. Des.* **22**, 133–139 (2008).
46. Artrith, N. et al. Best practices in machine learning for chemistry. *Nat. Chem.* **13**, 505–508 (2021).
47. Alves, V. M. et al. SCAM detective: accurate predictor of small, colloidally aggregating molecules. *J. Chem. Inf. Model.* **60**, 4056–4063 (2020).
48. Lee, K. et al. Combating small-molecule aggregation with machine learning. *Cell Rep. Phys. Sci.* **2**, 100573 (2021).

49. 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* **26**, 511–524 (2021).
50. 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* **26**, 1040–1052 (2021).
51. Brown, S. P., Muchmore, S. W. & Hajduk, P. J. Healthy skepticism: assessing realistic model performance. *Drug Discov. Today* **14**, 420–427 (2009).
52. Robinson, M. C., Glen, R. C. & Lee, A. A. Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *J. Computer Aided Mol. Des.* **34**, 717–730 (2020).
53. Cichońska, A. et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nat. Commun.* **12**, 3307 (2021).
54. Brown, N., Fiscato, M., Segler, M. H. S. & Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
55. Raji, I. D., Bender, E. M., Paullada, A., Denton, E. & Hanna, A. AI and the everything in the whole wide world benchmark. Preprint at *arXiv* https://arxiv.org/abs/2111.15366 (2021).
56. Renz, P., Rompaey, D. V., Wegner, J. K., Hochreiter, S. & Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discov. Today Technol.* **32–33**, 55–63 (2019).
57. Chen, L. et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS ONE* **14**, e0220113 (2019).
58. Wallach, I. & Heifets, A. Most ligand-based classification benchmarks reward memorization rather than generalization. *J. Chem. Inf. Model.* **58**, 916–932 (2018).
59. Sieg, J., Flachsenberg, F. & Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* **59**, 947–961 (2019).
60. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
61. Stanley, M. et al. in *35th Conf. Neural Inform. Process. Syst. Datasets Benchmarks Track* (NeurIPS, 2021).
62. Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O. & Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **11**, 154–168 (2020).
63. Chen, G. et al. Alchemy: a quantum chemistry dataset for benchmarking AI models. Preprint at *arXiv* https://arxiv.org/abs/1906.09427 (2019).
64. Rodrigues, T. The good, the bad, and the ugly in chemical and biological data for machine learning. *Drug Discov. Today Technol.* **32–33**, 3–8 (2019).
65. Heil, B. J. et al. Reproducibility standards for machine learning in the life sciences. *Nat. Mater.* **18**, 1132–1135 (2021).
66. McCloskey, K. et al. Machine learning on DNA-encoded libraries: a new paradigm for hit finding. *J. Med. Chem.* **63**, 8857–8866 (2020).
67. Giblin, K. A., Hughes, S. J., Boyd, H., Hansson, P. & Bender, A. Prospectively validated proteochemometric models for the prediction of small-molecule binding to bromodomain proteins. *J. Chem. Inf. Model.* **58**, 1870–1888 (2018).
68. Mathai, N., Chen, Y. & Kirchmair, J. Validation strategies for target prediction methods. *Brief. Bioinform.* **21**, 791–802 (2020).
69. Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 468–481 (2014).
70. Vishwakarma, G., Sonpal, A. & Hachmann, J. Metrics for benchmarking and uncertainty quantification: quality, applicability, and a path to best practices for machine learning in chemistry. Preprint at *arXiv* https://arxiv.org/abs/2010.00110 (2020).
71. Rosario, Z. D., Rupp, M., Kim, Y., Antono, E. & Ling, J. Assessing the frontier: active learning, model accuracy, and multi-objective candidate discovery and optimization. *J. Chem. Phys.* **153**, 024112 (2020).
72. Schwaller, P., Gaudin, T., Lányi, D., Bekas, C. & Laino, T. “Found in translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).
73. Yu, T. & Zhu, H. Hyper-parameter optimization: a review of algorithms and applications. Preprint at *arXiv* https://arxiv.org/abs/2003.05689 (2020).
74. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn. Sci. Technol.* **2**, 0115016 (2021).
75. Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **6**, 1379–1390 (2020).
76. Scikit-learn Developers. Cross-validation: evaluating estimator performance. *Scikit* https://scikit-learn.org/stable/modules/cross_validation.html (2021).
77. Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **53**, 783–790 (2013).
78. Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **11**, 4874 (2020).
79. Ho, S. Y., Phua, K., Wong, L. & Goh, W. W. B. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns* **1**, 100129 (2020).
80. Alexander, D. L. J., Tropsha, A. & Winkler, D. A. Beware of R^2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J. Chem. Inf. Model.* **55**, 1316–1322 (2015).
81. Golbraikh, A. & Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **20**, 269–276 (2002).
82. Consonni, V., Davide, B. & Todeschini, R. Comments on the definition of the Q^2 parameter for QSAR validation. *J. Chem. Inf. Model.* **49**, 1669–1678 (2009).
83. Derumigny, A. & Fermanian, J.-D. A classification point-of-view about conditional Kendall’s tau. Preprint at *arXiv* https://arxiv.org/abs/1806.09048 (2018).
84. Raeder, T., Forman, G. & Chawla, N. V. in *Data Mining: Foundations and Intelligent Paradigms* (eds Holmes, D. E. & Jain, L. C.) 315–331 (Springer, 2012).
85. Brown, J. B. Classifiers and their metrics quantified. *Mol. Inf.* **37**, 1700127 (2018).
86. Beker, W., Wołos, A., Szymkuć, S. & Grzybowski, B. A. Minimal-uncertainty prediction of general drug-likeness based on Bayesian neural networks. *Nat. Mach. Intell.* **2**, 457–465 (2020).
87. Perryman, A. L., Inoyama, D., Patel, J. S., Ekins, S. & Freundlich, J. S. Pruned machine learning models to predict aqueous solubility. *ACS Omega* **5**, 16562–16567 (2020).
88. Schwaller, P. et al. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144–152 (2021).
89. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
90. Mo, Y. et al. Evaluating and clustering retrosynthesis pathways with learned strategy. *Chem. Sci.* **12**, 1469–1478 (2021).
91. Talebian, S. et al. Facts and figures on materials science and nanotechnology progress and investment. *ACS Nano* **15**, 15940–15952 (2021).
92. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminf.* **9**, 48 (2017).
93. Blaschke, T., Engkvist, O., Bajorath, J. & Chen, H. Memory-assisted reinforcement learning for diverse molecular de novo design. *J. Cheminf.* **12**, 68 (2020).
94. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
95. Haghighi, S., Jasemi, M., Hessabi, S. & Zolanvari, A. PyCM: multiclass confusion matrix library in Python. *J. Open Source Softw.* **3**, 729 (2018).
96. Beker, W., Gajewska, E. P., Badowski, T. & Grzybowski, B. A. Prediction of major regio-, site-, and diastereoisomers in Diels–Alder reactions by using machine-learning: the importance of physically meaningful descriptors. *Angew. Chem. Int. Ed.* **58**, 4515–4519 (2019).
97. Haše, F., Roch, Lc. M., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: a Bayesian optimizer for chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).
98. Nielsen, M. K., Ahneman, D. T., Riera, O. & Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *J. Am. Chem. Soc.* **140**, 5004–5008 (2018).
99. MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).
100. Walters, W. P. & Murcko, M. Assessing the impact of generative AI on medicinal chemistry. *Nat. Biotechnol.* **38**, 143–145 (2020).
101. Aickin, M. & Gensler, H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am. J. Public Health* **86**, 726–728 (1996).
102. Chuang, K. V. & Keiser, M. J. Adversarial controls for scientific machine learning. *ACS Chem. Biol.* **13**, 2819–2831 (2018).
103. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
104. Chuang, K. V. & Keiser, M. J. Comment on “Predicting reaction performance in C–N cross-coupling using machine learning”. *Science* **362**, eaat8603 (2018).
105. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
106. Maragakis, P., Nisonoff, H., Cole, B. & Shaw, D. E. A deep-learning view of chemical space designed to facilitate drug discovery. *J. Chem. Inf. Model.* **60**, 4487–4496 (2020).
107. Zahrt, A. F. et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).
108. Reid, J. P., Proctor, R. S. J., Sigman, M. S. & Phipps, R. J. Predictive multivariate linear regression analysis guides successful catalytic enantioselective Minisci reactions of diazines. *J. Am. Chem. Soc.* **141**, 19178–19185 (2019).
109. Brix, K. V., DeForest, D. K., Tear, L., Grose, M. & Adam, W. J. Use of multiple linear regression models for setting water quality criteria for copper: a complementary approach to the biotic ligand model. *Environ. Sci. Technol.* **51**, 5182–5192 (2017).
110. Toste, F. D., Sigman, M. S. & Miller, S. J. Pursuit of noncovalent interactions for strategic site-selective catalysis. *Acc. Chem. Res.* **50**, 609–615 (2017).
111. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).
112. Zahrt, A. F., Athavale, S. V. & Denmark, S. E. Quantitative structure–selectivity relationships in enantioselective catalysis: past, present, and future. *Chem. Rev.* **120**, 1620–1689 (2020).
113. Rodrigues, T. Deriving intuition in catalyst design with machine learning. *Chem* **8**, 15–17 (2022).
114. Liu, B. et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
115. Dai, H., Li, C., Coley, C. W., Dai, B. & Song, L. Retrosynthesis prediction with conditional graph logic network. Preprint at *arXiv* https://arxiv.org/abs/2001.01408 (2020).
116. Vaucher, A. C. et al. Inferring experimental procedures from text-based representations of chemical reactions. *Nat. Commun.* **12**, 2573 (2021).
117. Gillet, V. J., Willett, P. & Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **38**, 165–179 (1998).
118. Edgar, S. J., Holliday, J. D. & Willett, P. Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *J. Mol. Graph. Model.* **18**, 343–357 (2000).
119. Schneider, G. & Böhm, H.-J. Virtual screening and fast automated docking methods. *Drug Discov. Today* **7**, 64–70 (2002).
120. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **3**, 1237–1245 (2017).
121. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminf.* **12**, 43 (2020).
122. Rodrigues, T., Almeida, B. P. D., Barbosa-Morais, N. L. & Bernardes, G. J. L. Dissecting celastrol with machine learning to unveil dark pharmacology. *Chem. Commun.* **55**, 6369–6372 (2019).
123. Rodrigues, T. et al. De novo fragment design for drug discovery and chemical biology. *Angew. Chem. Int. Ed.* **54**, 15079–15083 (2015).
124. Häse, F., Roch, L. M., Friederich, P. & Aspuru-Guzik, A. Designing and understanding light-harvesting devices with machine learning. *Nat. Commun.* **11**, 4587 (2020).
125. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
126. Moret, M., Helmstädter, M., Grisoni, F., Schneider, G. & Merk, D. Beam search for automated design and scoring of novel ROR ligands with machine intelligence. *Angew. Chem. Int. Ed.* **60**, 19477–19482 (2021).

127. Kearnes, S. Pursuing a prospective perspective. *Trends Chem.* **3**, 77–79 (2021).
128. Deringer, V. L. et al. Origins of structural and electronic transitions in disordered silicon. *Nature* **589**, 59–64 (2021).
129. Porwol, L. et al. An autonomous chemical robot discovers the rules of inorganic coordination chemistry without prior knowledge. *Angew. Chem. Int. Ed.* **59**, 11256–11261 (2020).
130. Kurczab, R., Smusz, S. & Bojarski, A. J. The influence of negative training set size on machine learning-based virtual screening. *J. Cheminf.* **6**, 32 (2014).
131. Lewis, R. A., Ertl, P., Schneider, N. & Stiefl, N. Reducing the concepts of data science and machine learning to tools for the bench chemist. *Chimia* **73**, 1001–1005 (2019).
132. Reutlinger, M., Rodrigues, T., Schneider, P. & Schneider, G. Multi-objective molecular de novo design by adaptive fragment prioritization. *Angew. Chem. Int. Ed.* **53**, 4244–4248 (2014).
133. Anders, C. J., Montavon, G., Samek, W. & Müller, K.-R. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R.) 297–309 (Springer, 2019).
134. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573–584 (2020).
135. Sheridan, R. P. Interpretation of QSAR models by coloring atoms according to changes in predicted activity: how robust is it? *J. Chem. Inf. Model.* **59**, 1324–1337 (2019).
136. Matveieva, M. & Polishchuk, P. Benchmarks for interpretation of QSAR models. *J. Cheminf.* **13**, 41 (2021).
137. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
138. Ribeiro, M. T., Singh, S. & Guestrin, C. “Why should I trust you?”: explaining the predictions of any classifier. Preprint at *arXiv* https://arxiv.org/abs/1602.04938 (2016).
139. Gao, H. et al. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).
140. Zhong, M. et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. *Nature* **581**, 178–184 (2020).
141. Riniker, S. & Landrum, G. A. Similarity maps — a visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminf.* **5**, 43 (2013).
142. Friederich, P., Krenn, M., Tamblyn, I. & Aspuru-Guzik, A. Scientific intuition inspired by machine learning generated hypotheses. *Mach. Learn. Sci. Technol.* **2**, 025027 (2021).
143. Webel, H. E. et al. Revealing cytotoxic substructures in molecules using deep learning. *J. Computer Aided Mol. Des.* **34**, 731–746 (2020).
144. Singh, S. et al. A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation. *Proc. Natl Acad. Sci. USA* **117**, 1339–1345 (2020).
145. Coley, C. W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
146. Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discov. Today* **20**, 458–465 (2015).
147. Reutlinger, M. et al. Chemically Advanced Template Search (CATS) for scaffold-hopping and prospective target prediction for ‘orphan’ molecules. *Mol. Inf.* **32**, 133–138 (2013).
148. Reker, D., Schneider, P. & Schneider, G. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chem. Sci.* **7**, 3919–3927 (2016).
149. Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H. & Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **7**, eabe4166 (2021).
150. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
151. Gromski, P. S., Granda, J. M. & Cronin, L. Universal chemical synthesis and discovery with ‘The Chemputer’. *Trends Chem.* **2**, 4–12 (2020).
152. Turing, A. M. Computing machinery and intelligence. *Mind* **56**, 433–560 (1950).
153. Mikulak-Klucznik, B. et al. Computational planning of the synthesis of complex natural products. *Nature* **588**, 83–88 (2020).
154. Duros, V. et al. Human versus robots in the discovery and crystallization of gigantic polyoxometalates. *Angew. Chem. Int. Ed.* **56**, 10815–10820 (2017).
155. Klucznik, T. et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522–532 (2018).
156. Shields, B. J. et al. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).
157. Polykovskiy, D. et al. Molecular Sets (MOSES): a benchmarking platform for molecular generation models. *Front. Pharmacol.* **11**, 1931 (2020).
158. Arús-Pous, J. et al. Exploring the GDB-13 chemical space using deep generative models. *J. Cheminf.* **11**, 20 (2019).
159. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
160. Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*. Thesis, Univ. Cambridge (2012).
161. Axelrod, S. & Gómez-Bombarelli, R. GEOM: energy-annotated molecular conformations for property prediction and molecular generation. Preprint at *arXiv* https://arxiv.org/abs/2006.05531 (2020).
162. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **48**, 4111–4119 (2005).
163. García-Ortegón, M. et al. DOCKSTRING: easy molecular docking yields better benchmarks for ligand design. Preprint at *arXiv* https://arxiv.org/abs/2110.15486 (2021).
164. Sun, J. et al. ExCAPE-DB: an integrated large scale dataset facilitating big data analysis in chemogenomics. *J. Cheminf.* **9**, 17 (2017).
165. Segler, M. H. S. & Waller, P. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. Eur. J.* **23**, 5966–5971 (2017).
166. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* https://arxiv.org/abs/1802.03426 (2020).

Acknowledgements

T.R. acknowledges FCT Portugal for funding (CEECIND/00684/2018). T.R. thanks colleagues for discussions on the topic presented here over the years. D. Reker is acknowledged for providing access to original data discussed in the manuscript. T.R., O.E. and A.B. acknowledge that not all suggested evaluation studies might simultaneously be found in their own original research manuscripts. We thank M. Thomas and M. Garcia-Ortegon for help with Table 1.

Author contributions

All authors contributed to the discussion and writing of the manuscript.

Competing interests

T.R. is a co-founder and shareholder of TargTex and has acted as consultant to the pharmaceutical industry. A.B. is a co-founder and shareholder of Healx, PharmEnable and Terra Lumina and acts as a consultant to various pharmaceutical companies.

Peer review information

Nature Reviews Chemistry thanks F. Grisoni and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher’s note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1038/s41570-022-00391-9.

RELATED LINKS

DOCKSTRING: https://github.com/dockstring/dockstring
DUD-E: http://dude.docking.org/
ExCAPE: https://solr.ideaconsult.net/search/excape/
FS-Mol: https://github.com/microsoft/FS-Mol
GDB-13: https://gdb.unibe.ch/downloads/
GEOM: https://github.com/learningmatter-mit/geom
GuacaMol: https://github.com/BenevolentAI/guacamol
Kaggle competitions: http://www.kaggle.com/
MoleculeNet: https://moleculenet.org/
MOSES: https://github.com/molecularsets/moses
PDBbind: http://www.pdbbind.org.cn/
RXNMapper: http://rxnmapper.ai/
SAMPL blind challenges: http://www.samplchallenges.org/
USPTO: https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873