Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

ELSEVIER

DRUG DISCOVERY TODAY TECHNOLOGIES

Drug Discovery Today: Technologies

Vol. 32, No. 2019

The good, the bad, and the ugly in chemical and biological data for machine learning

Tiago Rodrigues[1,2]

[1]Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina da Universidade de Lisboa, Av Prof Egaz Moniz, 1649-028 Lisboa, Portugal
[2]Research Institute for Medicines (iMed.ULisboa), Faculdade de Farmácia, Universidade de Lisboa, Av. Prof. Gama Pinto 1649-003, Lisboa, Portugal

Machine learning and artificial intelligence (ML/AI) have become important research tools in molecular medicine and chemistry. Their rise and recent success in drug discovery promises a rapid progression of development pipelines while reshaping how fundamental and clinical research is conducted. By taking advantage of the ever-growing wealth of publicly available and proprietary data, learning algorithms now provide an attractive means to generate statistically motivated research hypotheses. Hitherto unknown data patterns may guide and prioritize experiments, and augment expert intuition. Therefore, data is a key component in the model building workflow. Herein, I aim to discuss types of chemical and biological data according to their quality and reemphasize general recommendations for their use in ML/AI.

Introduction

Drug discovery is cornerstone for improved and sustainable healthcare. Its success is tightly connected to advances in chemistry and biology research that eventually provide innovative chemical matter for unexplored, yet disease-relevant drug targets and signalling pathways [1–3]. Over the years, a significant amount of data has been accumulated – by academia and the pharmaceutical industry – to a point where it is now humanly impossible to process all information enclosed [4]. Indeed, the generalized non-linearity of data correlations coupled to a perceived human inefficiency at integrating information from more than four variables simultaneously warrants the development of improved strategies to extract knowledge and efficiently advance discovery programs [5,6].

Recent advances in machine learning/artificial intelligence (ML/AI) heuristics, computing power and storage capacity now allow for correctly parsing and performing correlation analyses with the ever-growing amount of chemical and biological data, as annotated in both publicly-available (*e.g.* ChEMBL) and proprietary/corporate datasets. Moreover, the diminishing hardware costs and wide advocacy for open source tools democratizes the access to and development of bespoke ML/AI for myriad, real-world applications [7,8]. The use of ML/AI for retrosynthetic planning [9–13], *de novo* design [14–16], reaction product prediction [17–19], and drug target deconvolution [20,21], among others has been prospectively validated and thoroughly reviewed on multiple occasions [22–24]. While the latter contributions commonly focus on the ML/AI model architectures and strategies for formalizing expert knowledge, less attention has been given

E-mail address: (tiago.rodrigues@ff.ulisboa.pt)

 https://doi.org/10.1016/j.ddtec.2020.07.001

to a key component of the model development process – data source quality. One may argue that despite the growing amount of data, its collection is still a bottleneck in ML/AI for drug discovery. This is especially true for new ML/AI applications where labelled data is not necessarily available, or when a deep learning heuristics is employed and responsible for the feature engineering task. Data curation and quality check is also critical, time-consuming and an underappreciated task by the less informed community. However, the use of non-curated datasets can have a significant impact on the harmonization of information and, subsequently, on model quality and utility. Taking into account potential pitfalls, herein, I highlight the good (high quality and complete), the bad (moderate quality and sparse) and the ugly (low quality) in chemical and biological data for knowledge abstraction in ML/AI. To that end, I analyze and discuss the quality of information enclosed in select publicly available and proprietary databases. Based on those examples, I aim to pinpoint their caveats and reemphasize general guidelines for data inspection and curation to the least experienced medicinal chemistry researchers.

The good

ML/AI has seen numerous applications in discovery sciences and engineering through abstraction of expert knowledge. Thus, access to data for learning is key. While working with labelled data can influence the choice for supervised (classification/regression models for labelled data) over unsupervised (clustering/dimensionality reduction models for unlabelled data) methods, it is unquestionable that any ML/AI method requires quality data from where meaningful patterns are identified to describe a given event [5]. The utility of supervised ML/AI methods can be evaluated retrospectively and in intuitive fashion by comparing predicted values and classes with the ground truth, for the examples used in the model-training phase. Indeed, good practice usually suggests the assessment of supervised learning models through cross-validation studies without or with randomized target variables [25]. In the latter case, real data patterns are expected to be disrupted and lead to poorer model performance – a realization that was recently re-emphasized by machine learning practitioners in chemistry and biology [26]. Unsupervised learning can also be evaluated retrospectively, albeit less intuitively. For example, using well-established principles in information theory to compare the difference of two probability distributions, *i.e.* cross-entropy.

In real world, drug discovery scenarios, quality data in chemistry and biology can be heterogeneous and scarce. Specifically, data that is related to new discoveries or in new research fields is, by definition, not abundant. Also, the generation of data in different laboratories with different equipment and researchers, the different data acquisition methodologies and the often non-comparable experimental endpoints affect the harmonization of the information in datasets [7]. Taken together, experiments tend to suit a very specific research need, and are not designed for building databases or with a ML/AI application in mind. Moreover, data collection and labelling is very time consuming, expensive and requires domain knowledge [5]. This offers limited operational solutions for the majority of ML/AI researchers but to mine pre-existing datasets. Ideally, collection of information follows a standardized method while monitoring goal-oriented endpoints (*e.g.* reaction yield, IC_{50} value) that can be abstracted by relevant descriptors/features, such as physicochemical properties. Also, the number of training objects must be representative of the search space. This not only impacts on the applicability domain of the implemented ML/AI, but also influences the utility of the models for prospective use. Admittedly, extrapolation from a ML/AI model with narrow domain of applicability confers high uncertainty to its predictions, relative to in-sample predictions.

The abovementioned limitations can be surpassed by high throughput experimentation enabled by fully automated robotic systems. These offer a solution to more rapidly acquire data [27,28]. Among the numerous advantages of miniaturizing and parallelizing experiments, a salient feature of automating chemistry and biology experiments is the reproducibility of procedures that must be thoroughly documented and hardcoded in order to correctly capture information for abstraction. Given the specificity of chemistry experiments and absence of tailored datasets for most needs, generating on-demand information to deploy ML/AI models has been pursued and can be recommended (Fig. 1). As reported by Doyle and co-workers [29], 4608 Buchwald-Hartwig amination reactions were performed and simultaneously evaluated in 1536-well plates to generate the knowledge base required for ML/AI modelling. Furthermore, 640 deoxyfluorination reaction data was generated to train a random forest that could efficiently predict reaction outcomes given a descriptor set encoding educts [30]. Similarly, Cronin and colleagues generated own reactivity training data to leverage an active learning strategy as means of charting hitherto un-

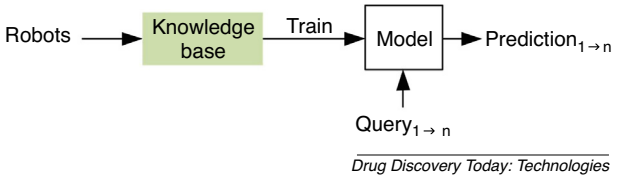

Fig. 1. High-throughput experimentation driven by robotic systems generates reproducible, high quality datasets for machine learning applications. The knowledge base may then be harnessed to generate statistical models that could be employed for predicting outcomes of new experiments.

known reactivity space and ultimately discover new reactions that may find applicability in the chemist's toolbox [31]. Understandably, despite the potential of these approaches, their feasibility may be currently limited to research groups with appropriate automation equipment. However, the absence of full automation to generate training datasets does not preclude the use of ML/AI in chemistry applications by the wider community. One may argue that more important than collecting data from hundreds or thousands of experiments is ensuring that all performed experiments are highly informative [32]. Moreover, the experiments ought to be diverse and devoid of anthropogenic biases [33] that could skew ML/AI models towards decisions in human-preferred search spaces – *i.e.* including a high percentage of positive, low risk outcomes. With that in mind, Reker *et al.* recently demonstrated that bespoke active learning heuristics exploring a small set of randomly performed experiments afford a solution to minimize biases and to identify optimized reaction conditions with minimal synthetic effort and in competitive fashion to human intuition [34]. Despite this latter application, one can foresee that future data-driven research may become increasingly automated through robotic systems, as these technologies are progressively democratized and made accessible to the wider academia and industry communities [27,35–37]. As such, information generated through probability computations in statistical learning coupled to state-of-the-art robotics, *i.e.* with minimal human intervention, will be preferred given the potentially balanced nature of the enclosed information, which will include both positive and negative results.

The bad

Undeniably, chemical and biological sciences of today evolve at an unprecedented speed with new chemical reactions, drug targets and signalling pathways being constantly discovered or clarified. While an immense amount of data is generated in this process, it is also true that databases aggregating chemical and biological findings are usually imbalanced [38,39] – *i.e.* focused on certain regions of search spaces or biased towards certain outputs (*e.g.* high yielding reactions) – and redundant in some instances, given that experiments are performed with a clearly defined goal and research question in mind. In this regard, data and information are not interchangeable terms, as a high amount of data may offer limited information for ML/AI. One may also argue that despite the wealth of available data, most database entries are not fully annotated resulting in sparse datasets, which may not be ideal for statistical learning. Therefore, construction of ML/AI models from out-of-the-box chemistry (*e.g.* US patents) and chemogenomic (*e.g.* ChEMBL and BindingDB) data repositories – when access to high-throughput experimentation is limited – might be ill advised. The resulting models will most likely under-perform in previously unseen prospective examples if careful data curation is not made. Common errors in databases include impossible valences in chemical structures, wrongly annotated tautomers and miscalculation or swapping of concentration values and units [40–42]. While a quality check can be manually executed, the need for a standardized data curation procedure has led to the implementation of automated pipelines capable of identifying erroneous structures, scoring the annotated data and removing outliers, to promote meaningful cheminformatics studies [43–45]. Some of these methods are open-source and do not require a specialized knowledge in order to be employed by the medicinal chemistry community [44].

Pharmaceutical patents have been a rich source of reliable information for the implementation of retrosyntheses recommender tools based on deep learning architectures [7]. However, mining of >1.1 million unique patent reactions in the 1976–2015 period revealed selection biases towards certain reaction types, *e.g.* acylation, alkylation, arylation among others [46]. Understandably, the accuracy of ML/AI tools built from this dataset will be poorer when predicting retrosynthetic routes for molecules requiring transformations less commonly present in the knowledge base. Additionally, a recent survey on the same patent data showed that reported yields are frequently inconsistent when comparing text-mined and calculated values for a given reaction in the same method description (Fig. 2). For example, 47,358 reactions (or 10% of all parsable patent data) presented a discrepancy of >10% yield between text-mined and calculated values. Moreover, the distribution of reported yields was highly skewed and in almost 50% of all reaction entries a yield was not reported [7]. These observations do not invalidate the utility of the dataset, but rather should caution the less experienced researcher regarding potential limitations that may invalidate downstream modelling.

Issues can also be found in chemogenomic databases, warranting caution in their use. As mentioned before, more often than desirable, molecular structures are not machine readable, which can invalidate all respective bioactivity annotations and values. While this might not impact the implementation of robust quantitative–structure activity relationship models for highly explored targets, the same is not true when information is scarce, as in the case of new and unexplored targets. It is however worth reemphasizing that sufficient amount of data might be available but a high bias for certain bioactivity values again limits the applicability of ML/AI. One of such examples refers to the global emergence of SARS-CoV2. The inexistence of medications/vaccines with proven clinical efficacy against this coronavirus at a time of pandemics has prompted an enormous worldwide effort to identify therapeutics to mitigate the disease, control the pressure on health systems and limit the death burden. Numerous reports have recently surfaced, detailing not only the epidemics of the disease, but also disclosing the potentially actionable drug targets [47] or potential

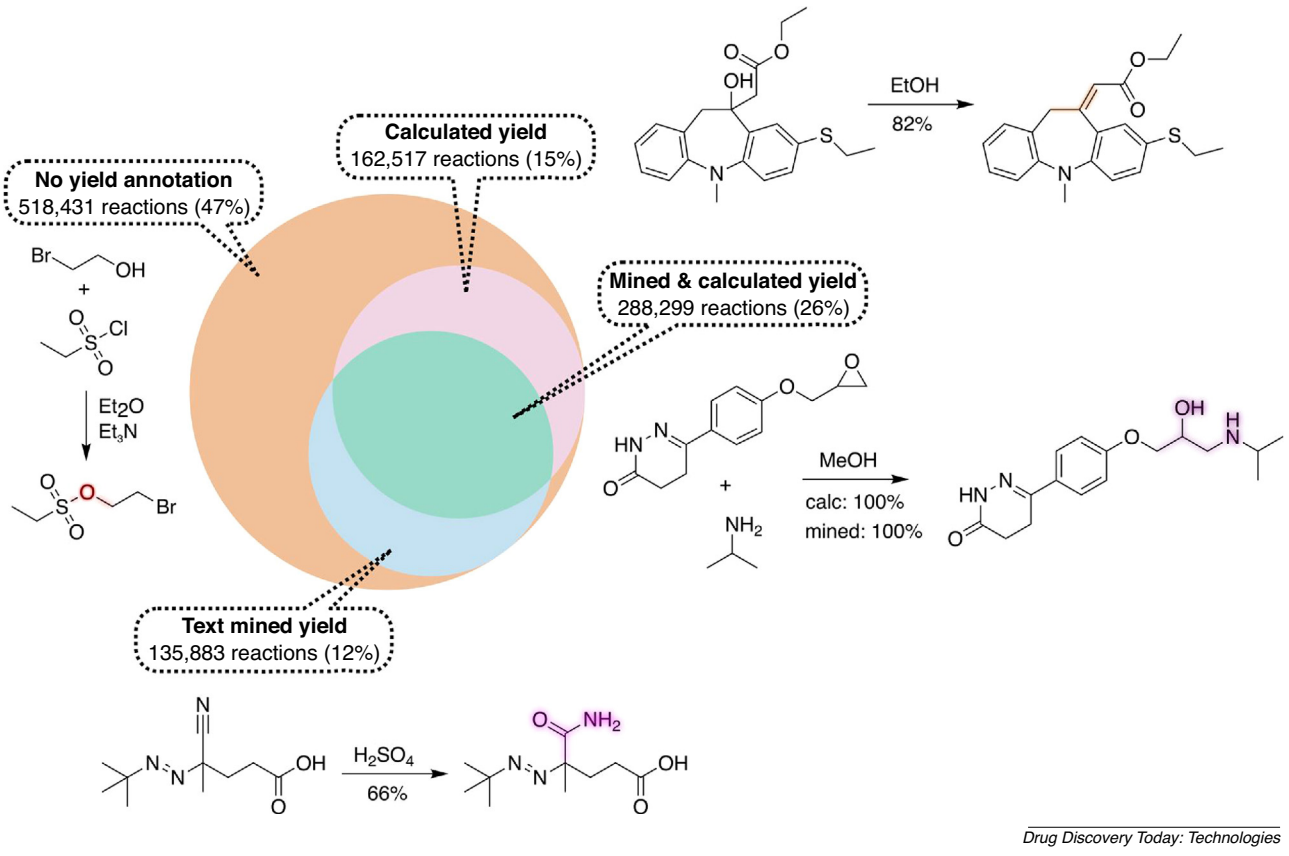

Fig. 2. Chemical reaction data in US patents from 1976 to 2015 period is highly heterogeneous and potentially unsuitable for bespoke machine learning applications. A high proportion of all reactions (47%) do not report a yield and >10% of reactions with both a calculated and text-mined yield show a discrepancy between both values. The reaction centre and functional groups originating from the depicted reaction are highlighted. Data extracted from Ref. [7].

repurposing strategies [48]. Given there is a high sequence identity (>95%) [49] between the main protease (3C-like) of SARS-CoV2 and SARS-CoV it is reasonable to recycle data from previous target-based high-throughput primary screen programs (*e.g.* https://pubchem.ncbi.nlm.nih.gov/bioassay/1706) for ML/AI. The dataset is however highly unbalanced, as only 0.1% of all screening compounds showed any sort of weak/moderate activity at a concentration of 6 μM against the main protease of SARS-CoV. Thus, while classification-based ML/AI algorithms are likely impracticable here, due to the lack of positive (active) examples, regression methods may be more applicable, yet not advisable. In regression, the goal is not to distinguish between a highly (inactive) and lowly (active) represented class, but predicting a real value. The distribution of target values should cover a range as wide as possible, which, admittedly, is not verifiable in this case since >99% of all entities are inactive against the main protease. Due to the narrow spread of bioactivity values, the utility of such a regression model likely is limited. Ultimately, this leads to a focused domain of applicability and high uncertainty for potential out-of-sample predictions. For quantitative structure activity relationship studies a minimum of 3 log unit spread in IC_{50}, $K_D/_i$ values is advisable [50], which, from my experience, is not always the case for all targets in ChEMBL – a publicly available chemogenomic data repository.

The ugly

Several data resources are, *per se*, not ideal for direct use as knowledge bases in ML/AI applications. Rather, they require extensive curation by expert chemists, biologists and data scientists, in order to extract only the relevant and informative data, and eliminate noise that could decisively compromise the utility of the implemented methods. Although I argue that no dataset is in itself useless, inappropriate exploitation of the underlying information may lead to important pitfalls and erroneous conclusions. Thus, bad practice constitutes a disservice to the community and may further generate scepticism relative to the utility of ML/AI tools in drug discovery. For example, in line with the discussed information heterogeneity in pharmaceutical patents, one may also find a high percentage of reaction protocols (9%) reporting yields above 100% and a skewed distribution of yields towards high values [7]. As knowledgeable synthetic chemists may recognize, this is an alternative reality that frequently

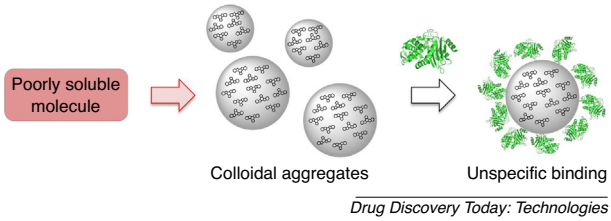

Fig. 3. Colloidal aggregation is the major source of false positive readouts in biological screening assays. Poorly soluble molecules form colloidal structures at micromolar concentrations, which are able to sequester and partly denature proteins adsorbed at their surface. This unspecific interaction leads to apparent drug target inhibition.

arises from misannotation and no reasonable repair can be recommended. Naturally, while the reaction protocols and product structures may still be valid and useful for bespoke applications – *e.g.* prediction of reaction products, reaction condition design or retrosynthetic planning – the same reaction description should not be considered if the goal is the prediction of a reaction yield from a set of educts.

Similarly, ML/AI heuristics have gained traction in recent years as a means of predicting on- and off-target effects for small molecules, either in quantitative or qualitative fashion [51,52]. These computations require a significant amount of biological data that is usually collected at different compound concentrations in order to obtain concentration–response curves. From those, an EC_{50}, K_D or K_i value is calculated and can be used as target value in ML/AI. However, at typical high-throughput screen concentrations, aggregation of small molecules into colloidal particles (SCAMs) is common, which may lead to sequestration and partial protein denaturation [24,53] (Fig. 3). Indeed, this is a widespread and unspecific mode of action for the grand majority of single and double-digit micromolar hits. As colloidal aggregation of hits and leads for drug discovery is still insufficiently controlled for, it is comprehensible that publicly available datasets are silently polluted with nuisances that may diverge attention in development programs to attrition prone and less promising chemical matter. Indeed, it has been estimated that up to 7% of all ChEMBL ligands may aggregate, which ultimately compromises the reliability of the biological data annotated to those entities [24,54]. ML/AI heuristics for affinity prediction harness this data and their utility may be questioned in cases where the number of training examples is small – *i.e.* where false biological patterns encoded in small colloidally aggregating molecules may stand out. With appropriate awareness for the aggregation problematic, chemogenomic datasets can be substantially improved. Dynamic light scattering and target-based screens with and without Triton X-100 should be more routinely performed to (de-)validate the identified ligand–target relationships.

Outlook

Data is a key component for any ML/AI-driven research program. The fast pace at which drug discovery unfolds, coupled to the rise of automation technologies has led to the generation of massive amounts of data. As discussed herein, such level of data wealth can become both a blessing (for experienced practitioners) and a curse (for the less experienced) given their heterogeneity. Data will tend to be appropriate for modelling when collected for a specific need. Conversely, repurposing previously assembled and curated datasets may lead to ill-informed decisions based on ML/AI artefacts. In this concise review I highlight a grand challenge in modern ML/AI research – identifying the good, the bad, and the ugly in chemical and biological data, as a function of their quality and completeness. This has been discussed by providing select examples for each case and with the goal of cautioning the casual ML/AI user for the associated shortcomings. In line with the on-going call for disclosure of ML/AI code and knowledge bases in published studies, it would also be beneficial for the community to standardize data collection and reporting methods. Although recommendations have been made in this direction, their adoption is still insufficient. Such measures can improve the quality of the generated ML/AI models to more effectively advance future drug discovery.

References

[1] Campbell IB, Macdonald SJF, Procopiou PA. Medicinal chemistry in drug discovery in big pharma: past, present and future. Drug Discov Today 2018;23(2):219–34.
[2] Bostrom J, Brown DG, Young RJ, Keseru GM. Expanding the medicinal chemistry synthetic toolbox. Nat Rev Drug Discov 2018;17(10):709–27.
[3] Lombardino JG, Lowe 3rd JA. The role of the medicinal chemist in drug discovery – then and now. Nat Rev Drug Discov 2004;3(10):853–62.
[4] Brown N, Cambruzzi J, Cox PJ, Davies M, Dunbar J, Plumbley D, et al. Big data in drug discovery. Prog Med Chem 2018;57(1):277–356.
[5] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 2019;18(6):463–77.
[6] Halford GS, Baker R, McCredden JE, Bain JD. How many variables can humans process? Psychol Sci 2005;16(1):70–6.
[7] de Almeida AF, Moreira R, Rodrigues T. Synthetic organic chemistry driven by artificial intelligence. Nat Rev Chem 2019;3:589–604.
[8] Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci 2018;9(24):5441–51.
[9] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018;555(7698):604–10.
[10] Schreck JS, Coley CW, Bishop KJM. Learning retrosynthetic planning through simulated experience. ACS Cent Sci 2019;5(6):970–81.
[11] Schwaller P, Petraglia R, Zullo V, Nair VH, Haeuselmann RA, Pisoni R, et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chem Sci 2020;11:3316–25.
[12] Coley CW, Thomas 3rd DA, Lummiss JAM, Jaworski JN, Breen CP, Schultz V, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science 2019;365(6453). eaax1566.
[13] Klucznik T, Mikulak-Klucznik B, McCormack MP, Lima H, Szymku S, Bhowmick M, et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. Chem 2018;4:522–32.

[14] Button A, Merk D, Hiss JA, Schneider G. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. Nat Mach Intell 2019;1:307–15.
[15] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Sci Adv 2018;4(7). eaap7885.
[16] Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol 2019;37(9):1038–40.
[17] Coley CW, Jin W, Rogers L, Jamison TF, Jaakkola TS, Green WH, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem Sci 2019;10(2):370–7.
[18] Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 2019;5(9):1572–83.
[19] Lee AA, Yang Q, Sresht V, Bolgar P, Hou X, Klug-McLeod JL, et al. Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. Chem Commun 2019;55(81):12152–55.
[20] Rodrigues T, Reker D, Welin M, Caldera M, Brunner C, Gabernet G, et al. De novo fragment design for drug discovery and chemical biology. Angew Chem Int Ed 2015;54(50):15079–83.
[21] Rodrigues T, Werner M, Roth J, da Cruz EHG, Marques MC, Akkapeddi P, et al. Machine intelligence decrypts beta-lapachone as an allosteric 5-lipoxygenase inhibitor. Chem Sci 2018;9(34):6899–903.
[22] Gao H, Struble TJ, Coley CW, Wang Y, Green WH, Jensen KF. Using machine learning to predict suitable conditions for organic reactions. ACS Cent Sci 2018;4(11):1465–76.
[23] Jensen KF, Coley CW, Eyke NS. Autonomous discovery in the chemical sciences. Part I: Progress. Angew Chem Int Ed 2019. http://dx.doi.org/10.1002/anie.201909987.
[24] Reker D, Bernardes GJL, Rodrigues T. Computational advances in combating colloidal aggregation in drug discovery. Nat Chem 2019;11(5):402–18.
[25] Mathai N, Chen Y, Kirchmair J. Validation strategies for target prediction methods. Brief Bioinform 2019;21(3):791–802.
[26] Chuang KV, Keiser MJ. Comment on 'Predicting reaction performance in C-N cross-coupling using machine learning'. Science 2018;362(6416).
[27] Roch LM, Häse F, Kreisbeck C, Tamayo-Mendoza T, Yunker LPE, Hein JE, et al. ChemOS: orchestrating autonomous experimentation. Sci Robotics 2018;3(19). eaat5559.
[28] Henson AB, Gromski PS, Cronin L. Designing algorithms to aid discovery by chemical robots. ACS Cent Sci 2018;4(7):793–804.
[29] Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. Predicting reaction performance in C-N cross-coupling using machine learning. Science 2018;360(6385):186–90.
[30] Nielsen MK, Ahneman DT, Riera O, Doyle AG. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. J Am Chem Soc 2018;140(15):5004–8.
[31] Granda JM, Donina L, Dragone V, Long DL, Cronin L. Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature 2018;559(7714):377–81.
[32] Reker D, Schneider G. Active-learning strategies in computer-assisted drug discovery. Drug Discov Today 2015;20(4):458–65.
[33] Jia X, Lynch A, Huang Y, Danielson M, Lang'at I, Milder A, et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. Nature 2019;573(7773):251–5.
[34] Reker D, Bernardes GJL, Rodrigues T. Evolving and nano data enabled machine intelligence for chemical reaction optimization. ChemRxiv 2018. http://dx.doi.org/10.26434/chemrxiv.7291205.v1.
[35] Daponte JA, Guo Y, Ruck RT, Hein JE. Using an automated monitoring platform for investigations of biphasic reactions. ACS Catal 2019;9(12):11484–91.
[36] MacLeod BP, Parlane FGL, Morrissey TD, Häse F, Roch LM, Dettelbach KE, et al. Self-driving laboratory for accelerated discovery of thin-film materials. arXiv 2020. arXiv:1906.05398v2.
[37] Steiner S, Wolf J, Glatzel S, Andreou A, Granda JM, Keenan G, et al. Organic synthesis in a modular robotic system driven by a chemical programming language. Science 2019;363(6423):eaav2211.
[38] Eitrich T, Kless A, Druska C, Meyer W, Grotendorst J. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. J Chem Inf Model 2007;47(1):92–103.
[39] Jain S, Kotsampasakou E, Ecker GF. Comparing the performance of meta-classifiers – a case study on selected imbalanced data sets relevant for prediction of liver toxicity. J Comput Aided Mol Des 2018;32(5):583–90.
[40] Waldman M, Fraczkiewicz R, Clark RD. Tales from the war on error: the art and science of curating QSAR data. J Comput Aided Mol Des 2015;29(9):897–910.
[41] Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 2010;50(7):1189–204.
[42] Kramer C, Lewis R. QSARs, data and error in the modern age of drug discovery. Curr Top Med Chem 2012;12(17):1896–902.
[43] Ambure P, Gajewicz-Skretna A, Cordeiro M, Roy K. New workflow for QSAR model development from small data sets: small dataset curator and small dataset modeler. integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques. J Chem Inf Model 2019;59(10):4070–6.
[44] Gadaleta D, Lombardo A, Toma C, Benfenati E. A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. J Cheminf 2018;10(1):60.
[45] Ruusmann V, Maran U. From data point timelines to a well curated data set, data mining of experimental data and chemical structure data from scientific articles, problems and possible solutions. J Comput Aided Mol Des 2013;27(7):583–603.
[46] Schneider N, Lowe DM, Sayle RA, Tarselli MA, Landrum GA. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. J Med Chem 2016;59(9):4385–402.
[47] Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, O'Meara MJ, et al. A SARS-CoV-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing. bioRxiv 2020. 2020.03.22.002386.
[48] Hoffmann M, Kleine-Weber H, Schroeder S, Kruger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020;181(2). 271–280.e8.
[49] Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. Science 2020;238(6489):409–12.
[50] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inf 2010;29(6–7):476–88.
[51] Reker D, Rodrigues T, Schneider P, Schneider G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. Proc Natl Acad Sci USA 2014;111(11):4067–72.
[52] Reutlinger M, Rodrigues T, Schneider P, Schneider G. Multi-objective molecular de novo design by adaptive fragment prioritization. Angew Chem Int Ed 2014;53(16):4244–8.
[53] Ganesh AN, Donders EN, Shoichet BK, Shoichet MS. Colloidal aggregation: from screening nuisance to formulation nuance. Nano Today 2018;19:188–200.
[54] Feng BY, Shelat A, Doman TN, Guy RK, Shoichet BK. High-throughput assays for promiscuous inhibitors. Nat Chem Biol 2005;1(3):146–8.