Artificial Intelligence Review (2023) 56:9089–9114
https://doi.org/10.1007/s10462-023-10391-w

Navigating with chemometrics and machine learning in chemistry

Payal B. Joshi[1]

Accepted: 9 January 2023 / Published online: 24 January 2023

Abstract
Chemometrics and machine learning are artificial intelligence-based methods stirring a transformative change in chemistry. Organic synthesis, drug discovery and analytical techniques are incorporating machine learning techniques at an accelerated pace. However, machine-assisted chemistry faces challenges while solving critical problems in chemistry due to complex relationships in data sets. Even with increasing publishing volumes on machine learning, its application in areas of chemistry is not a straightforward endeavour. A particular concern in applying machine learning in chemistry is data availability and reproducibility. The present review article discusses the various chemometric methods, expert systems, and machine learning techniques developed for solving problems of organic synthesis and drug discovery with selected examples. Further, a concise discussion on chemometrics and ML deployed in analytical techniques such as, spectroscopy, microscopy and chromatography are presented. Finally, the review reflects the challenges, opportunities and future perspectives on machine learning and automation in chemistry. The review concludes by pondering on some tough questions on applying machine learning and their possibility of navigation in the different terrains of chemistry.

Keywords Machine learning · Retrosynthesis · Automation · Chemometrics · Expert systems

Abbreviations

AI	Artificial Intelligence
ANN	Artificial neural network
CAMEO	Computer-assisted mechanistic evaluation of organic reactions
CAOSP	Computer-assisted organic synthesis planning
CASP	Computer-assisted synthesis planning
CHIRON	Chiral synthon
CHMTRN	CHeMistry TRaNslator
^{13}C-NMR	Carbon-13 nuclear magnetic resonance
CNN	Convoluted neural networks

✉ Payal B. Joshi
payalchem@gmail.com

[1] Operations and Method Development, Shefali Research Laboratories, Ambernath (East), Thane, Maharashtra 421501, India

COVID-19	Coronavirus disease of 2019
DENDRAL	Dendritic algorithm
DNN	Deep neural networks
EROS	Elaboration of reactions for organic synthesis
FAIR	Findable, Accessible, Interoperable, Reusable
FG	Functional group
FGI	Functional group interconversion
FIEM	Family of isomeric ensembles of molecules
GC–MS	Gas chromatography–Mass spectrometry
HIV	Human immunodeficiency virus
INTERLISP	List processing
IR	Infrared spectroscopy
KNN	k-nearest neighbours
KOSP	Knowledge base—Oriented system for synthesis planning
LC–MS	Liquid chromatography–Mass spectrometry
LDA	Linear discriminant analysis
LHASA	Logic and heuristics applied to synthetic analysis
LR	Linear regression
LSTM	Long short term memory
MALDI-TOF-MS	Matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry
ML	Machine learning
MS	Mass spectrometry
NIST	National institute of standards and technology
NIPALS	Nonlinear iterative partial least squares
NIR	Near infrared spectroscopy
NMR	Nuclear magnetic resonance
NNs	Neural networks
PASCOP	Programme d'Aide a la Synthèse en Chimi Organique et Organo Phosphorèe
PCANet	Principal component analysis networks
PLSR	Partial least square regression
ROBIA	Reaction outcomes by informatics analysis
SARS-CoV	Severe acute respiratory syndrome-associated coronavirus
SARS-CoV-2	Severe acute respiratory syndrome coronavirus-2
SERS	Surface-enhanced Raman scattering
SIMCA	Soft independent modelling by class analogy
SLING	SYNCHEM linear input graphs
SMILES	Simplified molecular-input line-entry system
SMARTS	Simplified molecular-input line-entry system arbitrary target specification
SOPHIA	System for organic reaction prediction by heuristic approach
SPRESI	Storage and retrieval of chemical structure information
SVM	Support vector machine
SWATH-MS	Sequential window acquisition of all theoretical fragment ion spectra-mass spectrometry
SYNCHEM	SYNthetic CHEMistry
SYNCHEM2	SYNthetic CHEMistry2
SYNSUP	Synthetic route design system

TOM	Target organic molecule
UV	Ultraviolet spectroscopy
WBCs	White blood cells
WODCA	Workbench for the organization of data for chemical applications
WLN	Wiswesser line notation

1 Introduction

Chemistry is the central science—organic synthesis, drug discovery and analytical techniques are the major domains that is utilizing artificial intelligent methods such as, chemometrics and machine learning resulting in a major transformation. Artificial intelligence (AI) garnered attention of chemists in early 1950s, yet at that time, computer-based learning was obscure or esoteric for solving chemistry problems. However, this situation did not persist for long. Over centuries, chemists have amassed huge data of chemical structures by performing several experiments. At that time, chemometrics was used to demonstrate computer usage in chemistry that assisted in solving complex problems. Massart et al. (1997) defined chemometrics as a "chemical discipline that uses mathematics, statistics, and formal logic (a) to design or select optimal experimental procedures; (b) to provide maximum relevant chemical information by analysing chemical data; and (c) to obtain knowledge about chemical systems." In 1975, a seminal paper featured 'chemometrics' in the title bringing a novel idea of utilizing computing tools to study complex chemical data (Kowalski 1975). In 1977, Analytical Chimica Acta introduced a section to communicate developing area of chemometrics pertaining to computer-assisted analysis especially for chromatography, UV, IR, ^{13}C-NMR, and mass spectrometric data (Clerc and Ziegler 1977). The section was devoted to pioneering work on NIPALS algorithm for principal component analysis, SIMCA and KNN algorithms for pattern recognition. Hence, chemometrics was primarily applied in pattern recognition that were influenced by two-fold approaches viz.

(a) kernel methods, machine learning, self-organizing maps, and support vector machines
(b) statistical methods such as, discriminant analysis, method validation, Bayesian models.

In a strict sense, chemometrics is typically a mathematical and statistical computer-based modelling utilized for optimizing methods and extracting results from analytical data. It was only from 1988, that the term "machine learning," made its debut in chemical literature titles (Appel et al. 1988; Gelernter et al. 1990; Sternberg et al. 1992; Salin and Winston 1992) and has ever since been used till date. In essence, chemometrics and machine learning has a fine distinction, as the former relies on linear relationships of data, while the latter deals with large and non-linear datasets. Machine learning involves the training of algorithms with chemical data and allows them to learn by examples. A trained machine learning model is deployed to deliver intelligent decisions. This necessitates using good data for machine learning models to navigate in solving chemical problems.

Today, chemists are consistently exploiting ML and chemometrics to solve challenging problems. This upsurge became apparent when Baum et al. (2021) reported the rise in journals and patents featuring AI-based methods in chemistry. Considering this increased interest and the hype of rapid march towards chemical automation led to the genesis of this review. The present review article describes the utilization of chemometrics and ML in

chemistry, particularly in organic synthesis and analytical chemistry. It discusses different expert systems utilized in organic synthesis. It essentially covers the earliest attempts of retrosynthesis and current ML methods applied to organic synthesis with chosen examples. Next, we describe reported literature showcasing efforts undertaken by medicinal chemists for COVID-19 therapy. Further, the progress of ML techniques applied to spectroscopy, microscopy and chromatography is presented. Due to the interdisciplinary nature of this review, discussions between chemists, computer scientists and mathematicians may lead to better investigations and unravelling mysteries of the chemical world. An attempt to address some tough questions on the current scenario of ML-based methods on chemistry is reflected. Rather than focusing on one particular domain, the present review aims to address selected domains of chemistry so as to bring out the divergent role of AI wholly.

2 Pacing organic synthesis with machine learning

Chemical space is a conceptual area that contains all possible chemical entities. It was envisioned by Lipinski and Hopkins (2004) that there are about 10^{180} number of possible molecules and about 10^{60} number of small organic molecules. Organic chemists are delving in this chemical space for exploring novel drug molecules. Due to the large possibility of molecules in chemical space, the search for novel molecules is challenging as a human endeavour bringing machine learning techniques as an attractive technology to the fore.

Chemistry is a new language to be learnt by machines that can efficiently predict organic synthesis routes at a faster pace. Before, we delve into machine learning methods, it is essential to describe earlier attempts made to study and predict organic reaction outcomes. Lederberg (1964) made an earliest attempt of an intelligent system in chemistry called the DENDRAL project that assisted chemists in identifying organic molecules from MS data. DENDRAL has been considered a pioneering expert system that automated problem-solving tasks of synthetic chemists. It was coded in INTERLISP and comprised of heuristic-DENDRAL and meta-DENDRAL modules. The heuristic-DENDRAL expert system worked on 'Plan-Generate-Test' sequence for organic structure elucidation using MS data. The meta-DENDRAL module predicted correct spectral data of novel molecules using chemistry rules. DENDRAL came across as a precursor for upcoming expert systems in chemistry and with pioneering work of developing knowledgebase of organic reactions by Elias J Corey led to retrosynthesis and their computing tools (Corey 1967). We also come across seminal work by Dugundji & Ugi (1973) who conceptualized algebraic matrix model called FIEM for understanding organic synthesis and mechanisms. In the following sections, various retrosynthetic tools developed by far, are discussed to express their growth in synthesis planning.

2.1 Solving maze of organic synthesis using retrosynthesis

The journey of organic synthesis dates to about 200 years ago when Wöhler (1828) prepared urea and oxalic acid. A typical problem in an organic synthesis is the structural description of the molecule to be prepared, called as the target organic molecule (TOM). TOM are compounds with important properties that could be a promising therapeutic agent or an industrially important intermediate.

Routinely, synthetic routes were performed by chemists with innate retrosynthesis—primarily a pen-paper method where chemists hand-draw the pathways based on

chemistry-based general rules and their intuition. Retrosynthesis is a conceptual problem-solving strategy for transforming the TOM to simpler starting materials that allows tracing of the feasible organic synthetic route to original target molecule (Fig. 1).

A synthetic chemist works backwards from the TOM by assuming possible disconnections about the chemical bonds. These disconnections generate synthons that refers to fragments, usually unstable species such as, ion or a radical. All these disconnections are the not real bond-breaking steps, rather is a mental foresight of the chemist based on general rules. Retrosynthesis is conceived in two forms namely, target-oriented or (whole molecule) and functional group interconversion (FGI) provided as an example of gabapentin (TOM), an anticonvulsant drug. It is only axiomatic to predict synthons for gabapentin via whole-molecule retrosynthetic strategy; it is FGI that exhibits wider options to arrive at possible starting reactants (Santos and Heggie 2020). In Fig. 1b, either 1-methylene cyclohexane or cyclohexanone are potential starting materials for synthesizing gabapentin. 1-methylene cyclohexane is problematic and costly whereas, cyclohexanone is toxic and an irritant in nature. Hence, chemists are supposed to make choices with certain trade-offs. It is advisable to select synthetic routes on the basis of availability of reagents, cost, and fewer reaction steps. Hence, in the case of gabapentin, either choose cyclohexanone or, a whole another class of organic reaction that shall generate fewer reaction steps and lesser toxic starting materials.

If the TOM is a complex entity, there is a greater chance of various distinct synthetic routes to prepare them. One can have over 10^{18} feasible one-step reaction routes to prepare target molecules. This led to Corey and Wipke (1969) to propose logic-oriented computer approach referred as *synthesis tree search*. In this approach, organic reactions are viewed as AND/OR tree, where the tree descends from the TOM i.e., goal node to the terminal nodes

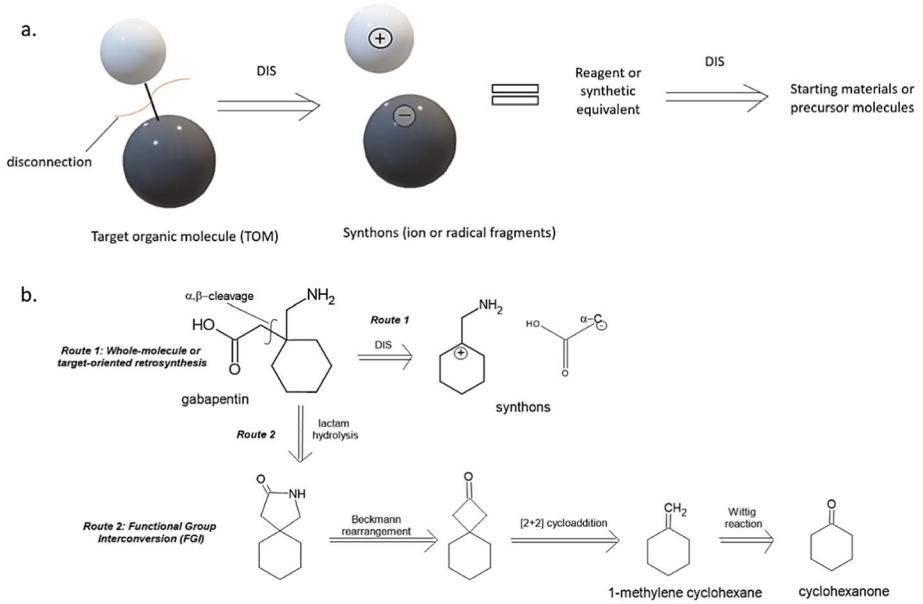

Fig. 1 **a** Schematic representation of retrosynthesis. DIS means disconnection. **b** retrosynthesis of gabapentin via whole-molecule or FGI types are depicted. There could be more than one way of interconverting FG of gabapentin; other than those depicted in the figure

that routes further to the simpler molecular fragments. These molecular fragments are the possible starting molecules. The branches that connect the goal and terminal nodes are the organic reactions. To put it simply, the AND-node and OR node links of the decision tree refers to organic reactions and molecules, respectively. Figure 2 depicts FGI scheme of AND/OR synthesis tree of gabapentin. Corey's seminal work was devoted to developing retrosynthetic tools that shifted focus of chemists from intuition-based strategies to logic-based instructions. (Corey et al. 1985). The earliest known retrosynthetic tool for assisting chemists was the LHASA program (Pensak and Corey 1977). LHASA was based on a sixfold strategy for retrosynthesis viz, transform, mechanistic transform, structure goal, stereochemical, topological and FG strategies. It used a special language CHMTRN constructed to note and search for disconnections.

Following LHASA, almost all retrosynthetic tools worked on these six-fold strategies for synthesis planning. CASP (Salatin and Jorgensen 1980) and CAMEO (Jorgensen et al. 1990) expert systems assisted chemists to find feasible synthetic routes and predict products respectively.

The provision of graphical knowledgebase editor was an innovative step in CASP that allowed communication with the chemist. CAMEO was a forward prediction tool that had selected rules on nucleophilicity, pKa, nature of leaving groups and steric effects which were used to rank the chemical reactions that the target molecule undergoes during substitution reactions. The forward prediction programs SOPHIA (Satoh and Funatsu 1995) and EROS (Gasteiger and Jochum 1978) assisted chemists in identification of active functional groups using reactivity rules and calculations. Ellerman et al. (1997) reported COSYMA

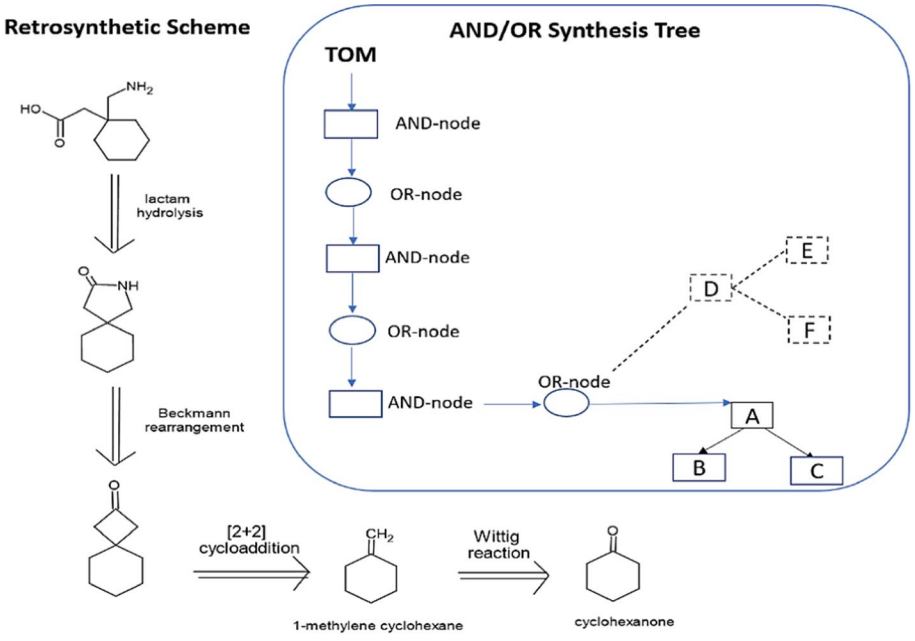

Fig. 2 AND/OR synthesis tree representation of FGI scheme of gabapentin. The AND node is the organic reaction and OR node refers to molecules. There can be more AND/OR nodes than those shown in the figure. The dashed boxes and lines of D, E and F refer to different chemical routes generating different starting materials other than those depicted in retrosynthetic scheme

program that searched for FGI and protecting-deprotecting groups. With advancing computing power, the task moved from number crunching to logic and reasoning leading to SYNCHEM and SYNCHEM2 programs. In SYNCHEM program, the initial stage involved chemists' choice of synthetic strategies to be tried out called 'synthemes.' Each syntheme had its own set of transforms that led to retrosynthetic routes and resultant precursors were assessed and ranked. The higher ranked precursors were processed further, which led to suitable material search in the reaction library in SYNCHEM (Gelernter et al. 1977). Further, Benstock et al. (1988) included stereochemistry that led to development of the SYNCHEM2 program. They entered chemical structures in SYNCHEM using WLN representation, whereas in SYNCHEM2, a linear SLING representation was used. Mehta et al. (1998) reported the SESAM program that utilized a backtracking algorithm to determine suitable starting materials to the target molecule. Hanessian et al. (1990) demonstrated CHIRON database search for synthetic routes to stereochemical compounds to obtain starting materials that showed maximum overlap of carbon skeleton, FGs and stereochemistry.

Many programs helped retrosynthetic planning such as PASCOP (Choplin et al. 1978), RETROSYN (Blurock 1990), WODCA (Gasteiger and Ihlenfeldt 1990), KOSP (Satoh and Funatsu 1999), ROBIA (Socorro et al. 2005), and CAOSP (Bersohn 1972; Tanaka et al. 2010) that were either a retrosynthetic or a forward prediction program. It is evident that synthetic chemists were utilizing computer-aided retrosynthesis. Further, if one incorporates machine learning in organic synthesis, it shall lead to evolutionary change in proposing forward syntheses. It is evident that post-LHASA, many expert systems allowed automation for planning multistep synthesis in chemistry laboratories. However, they could visualize only one step at a time for simpler target organic molecules. Thus, such a program caused an impediment for its application in multistep natural product syntheses.

Table 1 Features of present computer programs that support organic synthesis planning

Name of the program	Features
ChemPlanner 1.0	Analysing known synthetic target molecules and generating one-step reaction pathways from literature (Stark et al. 2016)
Chematica	Algorithms that draw utilizing hand-coded rules available in the database. It allows both retro- and forward- prediction of reactions. It avoids predicting patented synthetic routes (Klucznik et al. 2018)
Spaya AI	Based on deep machine learning using databases such as Mcule, Chemspace, EMolecules (Parrot et al. 2021)
LillyMol	Utilizes machine learning approach, atom mapping and train reaction transformation rules (Watson et al. 2019)
AutoSynRoute	Utilizes Montel-Carlo tree search with heuristic scoring function, transformer-type-seq-2-seq model (Lin et al. 2020)
AiZynthFinder	Monte Carlo tree search by ANN policy that allowed prioritizing reaction templates to generate novel precursors (Genheden et al. 2020)
IBM RXN for Chemistry	Forward prediction, one-step retrosynthetic tool that uses seq-2-seq database, natural language approach and trained on automatic extracted chemical data (IBM 2018)
ICSYNTH	Utilizes machine learning to generate chemical rules from SPRESI database (Bøgevig et al. 2015)
PostEra Manifold	Open-source retrosynthesis tool that allows search for different synthetic routes and generates comparison of raw materials from different vendors (PostEra 2021)

Table 1 enlists the current programs that assist chemists for selecting novel route of organic syntheses.

As the computing capacity kept increasing, algorithms needed improvisations and organic reaction forward prediction programs were visualized in two modalities viz, template-based and template-free methods, both of which had their trials in a chemical prediction task. Template-based methods are rule-based with reaction libraries and scoring functions; those that are discussed earlier in this section. Template-based approach may be a good starting point, but the basic premise of generating and extracting algorithms from set templates may spruce bias in the data as it largely relies on chemists' intuition. Template-free approach solves the bias issue which includes utilizing NNs and seq-2-seq models.

Nam and Kim (2016) pioneered neural machine translation for predicting reactions from patent dataset and Wade's Organic chemistry textbook. They trained their model with patent reactions spanning from 2001 to 2013 US applications and 75 reactions for five different starting molecules given as text problem in Wade's book. Liu et al. (2017) pioneered data-driven model that learnt reaction predictions by seq-2-seq recurrent NNs that was trained with 50,000 experiments from the US patent literature using SMILES text representation.

Recalling the point to consider chemistry as a language by machines (refer Fig. 3), Schwaller et al. (2018) moved a step ahead by demonstrating computational liguistics to solve chemical predictions. They related organic chemistry to a language and applied template-free seq-2-seq models. Adopting a model reported by Vaswani et al. (2017) and using SMILES representation, Schwaller's team developed Molecular Transformer that demonstrated higher accuracy for predicting reaction outcomes (Schwaller et al. 2019). Further, it could accurately predict selectivity, specificity, regioselectivity and chemoselectivity of the reactions.

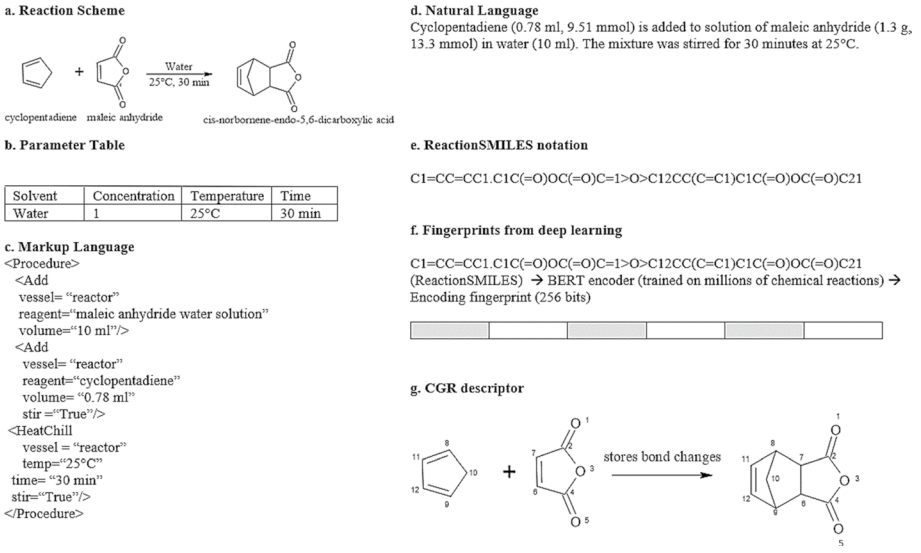

Fig. 3 Representation of typical Diels–Alder reaction between cyclopentadiene and maleic anhydride. **a** Kekulé type reaction graph; **b** parameter table allows optimization data capture for chemical reaction; **c** and **d** uses markup and natural languages respectively, of which the former is of greater significance; **e**, **f** and **g** describes reactions as ReactionSMILES, chemical fingerprints and descriptors that is easier for machines to understand. CGR means Condensed Graph of Reaction

Intriguingly, this model is utilized in IBM RXN (refer Table 1). Other efforts of ML in organic synthesis worth mentioning are automation in chemical sciences (Dragone et al. 2017), ML-based reaction optimization (Gao et al. 2018), and DL-based chemical pattern prediction (Cova and Pais 2019).

It is opined that natural product synthesis, organocatalysis and drug discovery are the three fundamental areas of chemistry that are utilizing state-of-the-art ML techniques. Natural product synthesis and organocatalysis particularly, fall under the category of organic chemistry and have witnessed major transformation in terms of retrosynthesis, and hence covered in the next section. Considering the expanse of drug discovery and repurposing, it is discussed in a separate Sect. 3.

2.1.1 Natural product syntheses

Natural products are complex target molecules with multiple cyclization reactions making the synthetic routes difficult to interpret. Chemists find planning multistep natural product synthesis a challenging endeavour. If one integrates computational methods with AI technique, it shall be of great relevance to understand natural product synthesis. Tantillo (2018) discussed typical questions that can be solved using computational modelling of natural product synthesis. Marth et al. (2015) reported natural product synthesis of weisaconitine D and liljestrandinine by modelling network analysis along with AI- assisted retrosynthesis. Kim et al. (2019) reported total synthesis of Paspaline A and Emindole PB using a computational model integrated with AI assisted retrosynthesis. Chematica team designed machine-tuned natural product syntheses of (−)-dauricine, (R,R,S)-tacamonidine and lamellodysidine A that were reported to be comparable to those designed by skilled chemists (Klucznik et al. 2020).

2.1.2 Organocatalysis

Asymmetric enantioselective organocatalysis is ranked as one of the emerging chemically sustainable technologies (Gomollón-Bel 2019). The effect of isoxazole additives on carbon–nitrogen coupling Buchwald-Hartwig reaction was reported that used machine learning to predict reaction outcomes using random forest (Ahneman et al. 2018). Kondo et al. (2020) demonstrated atom-efficient organocatalyzed enantioselective Rauhut-Currier and [3 + 2] annulation reactions for chiral spirooxindole analogue in a flow system. The authors applied Gaussian regression to multi-parameter reaction screening processes. It is realized that determining transition states of enantioselective reactions is time-consuming and lacks accuracy, bringing ML to the rescue. Gallarati et al. (2021) developed ML model that predicted enantioselectivity of Lewis-catalysed propargylation reactions. Further, the ML model predicted absolute configuration of enantiomeric excess product independently. This work is unique, as enantioselectivity of an organocatalyst is a challenging task to be predicted by ML models. The authors represented propargylation reaction to ML model trained an algorithm for calculating activation energies of competing catalytic pathways. This novel strategy of utilizing activation energy differences of organocatalytic products has paved way of deploying ML algorithms to solve complex enantioselective catalyst systems.

3 Facilitating drug discovery and repurposing

Drug discovery process involves identifying new chemical entities as potential therapeutic agents. By now, it is realized that emerging infectious diseases (EIDs) are a part of the human race, AI-based methods are sought after for their predictive modelling. It is felt that intelligent systems, if in place, shall be able to predict emerging diseases, prior to its occurrence. ML methods are particularly robust, when applied as predictive model in drug discovery and public health. Figure 4 describes supervised, unsupervised or reinforcement learning to represent drug molecules and understand their therapeutic potential. Target validation, biomarker identification and computational pathology are the three key areas of drug discovery that have adapted DL methods particularly, for therapeutics in cancer, and most recently in SARS-CoV-2 disease. Considering the expanse of drug discovery, this section is particularly focused to highlight the recent efforts undertaken for discovering antiviral COVID-19 agents using advanced ML methods. A brief section is devoted to describe the recent progress witnessed in drug repurposing methods for COVID-19. For a more comprehensive review on drug discovery, readers can refer reviews by Dara et al. (2022), Kolluri et al. (2022), Shehab et al. (2022) and Pillai et al. (2022). Though, drug discovery is described separately, the understanding of organic synthesis is symbiotic with this field.

3.1 Drug discovery for COVID-19

Drug discovery, particularly the stages of target drug identification, compound screening and preclinical studies necessitates tremendous scope for applying ML-based methods. If machine learning and deep learning techniques can assist in bringing a causal relationship between target novel molecule and the disease, drug discovery shall become cost and time efficient endeavour for pharmaceutical industries. In this section, a concise discussion is presented on the progress of drug discovery for antiviral agents against COVID-19.

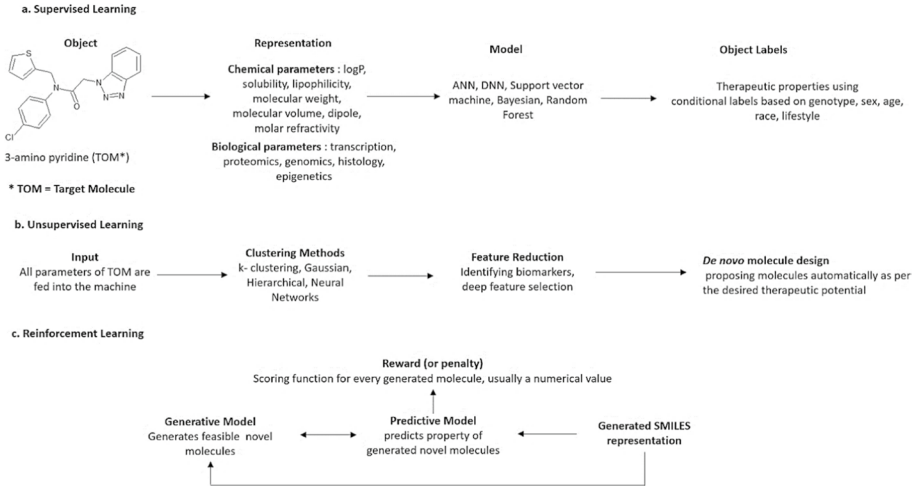

Fig. 4 Supervised, unsupervised and reinforcement learning in drug discovery

Amilpur and Bhukya (2022) reported LSTM model for searching and generating novel molecules that can potentially bind with main 3CLPro protease of coronavirus. They screened about 2.9 million molecules from ChemBL, Moses and RDKit databases and represented by SMILES prior to deploying on generative LSTM model. Using binding affinity scores, 10 potential drug candidates were suggested by their model for treating infections. A state-of-the-art quantum computing ML-based framework was designed as an in silico tool for discovering novel drug candidates against COVID-19 (Mensa et al. 2022). A novel MP-GNN model and featurization was reported to designing COVID-19 drugs (Li et al. 2022). Their model comprised of two unique properties viz, multiscale interactions that utilized more than one type of molecular graph and simplified feature generation. They validated MP-GNN model with datasets from PDBbind. Over 185 complexes of inhibitors for SARS-CoV-2 were evaluated for their binding affinities using their unique model. Drug molecules and chirality have always presented a unique relationship. Exploiting this premise for natural remedies, natural products were screened for finding novel drug candidates. Vasighi et al. (2022) proposed a ML-based technique to classify and discover COVID-19 inhibitors obtained from natural products. They prepared docking protocol with 125 ligands and analyzed protein–ligand interactions and drug-likeness properties of inhibitors using statistical exploratory data analyses. Structural characteristics of SARS-CoV-2 especially the spike proteins were immensely investigated. It was revealed that Cathepsin-L (CSTL) increased the severity of COVID-related infections by activating spike protein of the coronavirus (Zhao et al. 2021). Hence, CSTL became a promising target and the search for their inhibitors were widely investigated using advanced DL-based techniques and statistical models. Yang et al. (2022a) reported DNN alongwith Chemprop for identifying novel molecules and approved drugs that blocked CSTL activity. Five molecules namely, daptomycin, Mg-132, Mg-102, Z-FA-FMK and calpeptin potentially blocked CSTL activity and alleviated severity of secondary COVID infections.

All these reports elicit that most researchers did not solely rely on vaccines, but rather focused on novel molecules as potential drug candidates for alleviating COVID infections. Inspite of public misinformation, vaccines are a safer therapy to combat the disease, albeit it cannot be entirely relied upon essentially due to resistance of mutant SARS-CoV-2 and subsequent breakthrough infections.

3.2 Drug repurposing for COVID-19

When the world was hit by COVID-19 pandemic, there was an urgent need to handle the spread of coronavirus and its treatment. With no vaccines then, the pandemic forced researchers to innovate and strategize antiviral treatment using AI-based techniques. This urgency also led researchers to find old drugs utilizing AI-based learning methods for treating COVID infection. This process of finding existing approved drugs for treating emerging diseases is called *drug repurposing*. As SARS-CoV and SARS-CoV-2 viruses display similar receptor binding mode (Lan et al. 2020), AI-assisted models utilized their structural data and predicted drug molecules that could alleviate COVID-19 symptoms. Until the vaccines arrived, these old, marketed drugs were repositioned for treating COVID-19 infected patients (Mohanty et al. 2020). The AI-assisted drug repurposing required an open drug database, repurposed drug database as input labels and then various algorithms are applied to them. All these processes generate the drug molecule required for the purpose. The critical issue in drug repurposing is the determination of a unique drug-disease relationship. AI-learning

modelled with molecular descriptors, functional-class fingerprints (FCFPs), chemical fingerprints, and physico-chemical properties like partition coefficients could screen and identify drugs for treating coronavirus patients. It is revealed that drug repurposing for COVID-19 primarily utilized three types of algorithms viz, network-based (Ge et al. 2021), expression-based (Pham et al. 2021) and integrated docking simulations (Ahmed et al. 2022). Sibilio et al. (2021) examined three different network-based algorithms to identify potential drug molecules using transcriptomic data from the WBCs of COVID infected patients. They performed in silico studies that predicted drug-disease association and disease-likeness of COVID with other diseases. Yang et al. (2022b) demonstrated utilization of a novel web-server called D3AI-CoV for target identification and screening of drugs to combat COVID infections. They employed advanced DL-based models with canonical SMILES representation and more than 800 bioactives and 29 targets against nine coronavirus variants. Xie et al. (2022) proposed a compressed sensing algorithm combined with centered kernel alignment that short-listed total 15 drug candidates as therapeutics for COVID-19.

Most of the reported literature focused on network-based, expression-based and docking simulation algorithms for identifying drug-disease relationships, viral gene expressions and host protein target interactions. It is argued that even with these reports, DL-based are limiting in scope while determining repurposed drugs for their potential use as COVID-19 treatment. Most of the DL-based methods require huge patient dataset that is not publicly available hindering the infection and survival predictions of COVID-19 infection. Hence, most of the reported literature utilized smaller data set that cannot be extrapolated for public health studies.

Proceeding with the discussion, the review now shifts focus to analytical chemistry especially, chemometrics and ML techniques on spectroscopy, microscopy and chromatography. A tremendous scope of successful chemometrics and ML-based techniques are witnessed in analytical chemistry. It is envisioned that, on further advances, automated analytical systems will be a reality. Following section describes the current progress of AI-based techniques and automation in spectroscopy, microscopy and chromatography.

4 AI and automation in analytical chemistry

Modern analytical techniques create huge data for heterogenous samples that needs to be interpreted by the chemists. Analytical chemists spend most of their time identifying and quantifying molecules in laboratory samples ranging from food, drug molecules to industrially important molecules. Chromatograms and spectra are generated that undergo chemometric and standard mathematical algorithms to derive useful information, though a huge subset of data remains ignored. Earlier, the Library Search Algorithm was employed to obtain crucial information about molecular structures from spectral data. Today, the situation has matured to a certain extent that utilizes machine learning techniques such as, convoluted neural networks on spectral peaks, microscopic images and chromatograms.

Prior to data interpretation, chemical data retrieved from instrumental techniques are composed of distortions called artefacts. These artefacts are caused due to noise levels from instruments, sample type, solvent effects and physico-chemical factors. Their

presence in spectral and chromatogram data adversely affects crucial data sets leading to loss of chemical information. When these distortions are eliminated or suppressed for data enhancement, it is called pre-processing method. It involves correcting peak shifts, baseline corrections, noise removal, stray light suppression and retrieving missing data values (Chalmers 2006). In the following section, chemometrics and ML methods employed in spectroscopy, microscopy and chromatography are discussed (refer Fig. 5).

4.1 Chemometrics and machine learning in spectroscopy

Chemometrics became successful as a statistical technique for its application in near-infrared spectroscopy. Near-infrared spectra contain deeply convoluted signals that are not separated by baseline, thus making it difficult to quantify crucial information regarding molecules. Vibrational spectroscopy, NMR, MS, and hyphenated techniques generate multidimensional spectral data that contains a plethora of critical information related to molecular structures. These data are optimized and studied using chemometrics and machine learning methods with enhanced precision and accuracy.

4.1.1 Vibrational spectroscopy

NIR, IR and Raman spectroscopy are typical vibrational spectroscopic methods that derives structural information by measuring vibrations of molecules. An open-source python module called "nippy" was employed for NIR spectral data (Torniainen et al. 2020). Roger et al. (2020) reported the utilization of sequential and orthogonal PLS regression for pre-processing NIR spectral data of wheat grains, tablet and meat samples. Martyna et al. (2020) applied genetic algorithm to Raman spectral data. The deployed genetic algorithm assessed the pre-processing technique by calculating variance ratios and validated it by

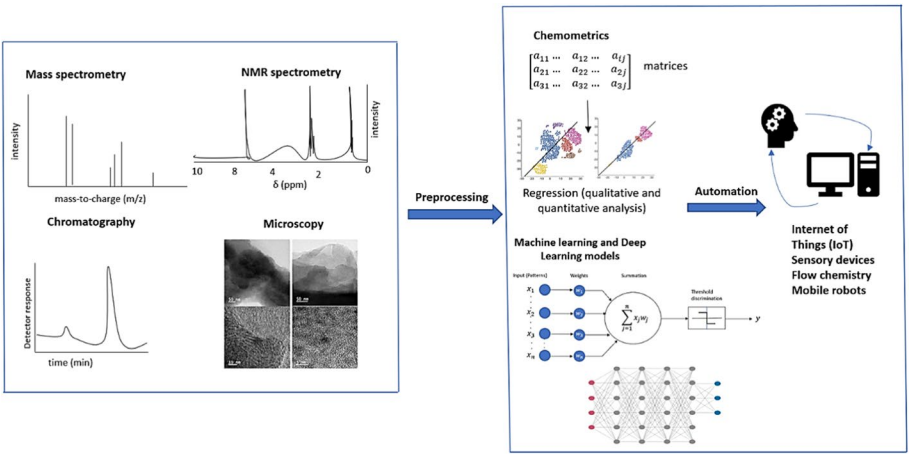

Fig. 5 Overview of chemometrics and ML methods applied to analytical techniques. Spectroscopy, chromatography and microscopy are depicted on the left panel (not drawn to scale, not representative of any data). The right panel depicts chemometrics and ML models applied on analytical data after pre-processing. Finally, it depicts navigation towards automation that utilizes IoT, sensory devices, flow chemistry and mobile robots

applying it on forensic Raman spectral data. Post chemical data enhancement, they applied various chemometric algorithms to obtain critical information from spectroscopic data. Raman and SERS techniques produce complex vibrational spectra of chemical mixtures which are exploited to obtain critical information.

Until now, LR analysis was performed to obtain useful data from Raman and SERS vibration spectra, but deep learning has replaced these statistical models. Weng et al. (2020) modelled a deep learning with CNN and PCANet that identified drugs in human urine with an accuracy above 98.05%. They also measured pirimiphos-methyl in wheat extract that was quantified using fully convoluted neural networks with a determination coefficient of 0.9997. Table 2 lists selected spectroscopic techniques, different chemometric and ML-based techniques with their potential applications.

4.1.2 NMR spectroscopy and mass spectrometry

NMR spectroscopy and mass spectrometry are sophisticated analytical techniques that provide critical information on type of nuclei and m/z of chemical molecules respectively. Particularly, deep neural networks gained importance in NMR spectral interpretation to enable time-efficient data acquisition and lower chemists' training endeavours (Chen et al. 2020). Kong et al. (2020) reported deep learning through CNN coupled with sparse matrix completion to suppress noise and speeding up 2D nanoscale NMR spectroscopy. A momentum of interest was witnessed for DNNs being utilized for reconstructing non-uniformly sampled NMR to enhanced resolution at shorter time (Hansen 2019; Karunanithy and Hansen 2021). One particular concern was to unravel critical structural information from multidimensional NMR spectra obtained during metabolomic studies. Metabolomic study generates large data with crowded NMR spectral peaks and hence peak picking is an old yet a hard problem. Conventional peak picking methods in a routine NMR instrument may be insufficient. The specialized DNNs are providing respite to analytical chemists to decode these utilizing advanced GUI interface (Rahimi et al. 2021) and DNNs (Li et al. 2022). Native MS spectrometry is utilized for unravelling macromolecule structures particularly nucleic acids and proteins. An intriguing study was reported by Allison et al. (2022) on applying native MS for structural elucidation of selected protein complexes that complemented ML methods.

When spectrometric methods are combined with chromatography, they are called *hyphenated techniques*. Hyphenated techniques such as, LC–MS, GC–MS, etc. produce multidimensional data that requires advanced DL techniques for data interpretation. Qiu et al. (2018) reported GC–MS data interpretation without spectral library database query and efficiently prioritized biological candidate molecules by orthogonal datasets of retention indices, mass spectra and other physicochemical parameters of compounds. Recently, a deep learning algorithm called 'peakonly' was developed by Melnikov et al. (2020) that provided precise peak identification and integration in LC–MS data.

4.2 Chemometrics and machine learning in microscopy and chromatography

The advances in chemometrics and ML methods have led to utilizing them in chemical data image processing in electron microscopy, atomic force microscopy and 2D chromatographic techniques. It has allowed insights to crucial information about the molecular structures where chemical images are obtained either as grayscale or hyperspectral images.

Table 2 Selected spectroscopic techniques, different chemometrics and machine learning methods and their applications

Analytical technique	Chemometric software/machine learning method	Features
MS	SWATH-MS	Raw mass spectral (MS) image data of prostate cancer tumors were sampled. Image data were encoded as feature vectors and classified using ML-based models (Cadow et al. 2021)
MS	VOCCluster	Python-based algorithm analyzed MS data of deconvolved human breath with an accuracy of about 96% (Alkhalifah et al. 2020)
MALDI-TOF-MS	CNN algorithm	Accurate identification of pathogenic bacterial species in urine samples utilized DL-based model (Papagiannopoulou et al. 2020)
Raman spectroscopy	SIMCA, ANN and PLSR	Real-time, non-invasive carotenoid analysis with enhanced quantification of *trans*-lycopene in fruits (Akpolat et al. 2020)
Raman spectroscopy	ANN, SVM, LDA, LR, Interval PLS method	Estimated sugar concentration in industrial food products using Raman spectroscopy with feed-forward NNs (Viveros et al. 2021)
Mid-infrared laser spectroscopy	Multivariate analysis, pre-processing by classical least squares, random forest model	Quantified explosives in soil using knowledge base of spectral data, demonstrated regression analysis of 0.996 accuracy (Pacheco-Londoño et al. 2020)
Nuclear magnetic resonance spectroscopy	Orthogonal projection to latent structures discriminant analysis and random forest	Differentiated a variety of American beers (Vasas et al. 2021)

In this section, the recent advancement of ML methods reported in imaging techniques and chromatography is explored.

4.2.1 Atomic force microscopy (AFM)

AFM is an advanced analytical topographic imaging technique that produces high—resolved images at atomic resolution allowing nanoscale characterization of important materials such as biological and inorganic samples. Previous attempts were made to minimize heuristic probe conditioning while imaging using algorithms (Villarrubia 1997), inverse imaging of probe (Schull et al. 2011; Welker and Giessibl 2012; Chiutu et al. 2012) and probe manipulation in atomic force microscopy (Paul et al. 2014). However, these methods are not suitable for large dataset acquisition. AFM imaging presents challenges such as scan speed, optimization, and artefacts in scanned images. An autonomous atomic force microscopy utilizing an AI framework was reported by Krull et al. (2020) that allowed probe quality assessment, conditioning, and its repair along with large data acquisition. Javazm and Pishkenari (2020) proposed adaptive and multi-layered neural fuzzy inference system NNs for solving the problem of AFM restricted scan speeds. Payam et al. (2021) reported AFM data acquisition and imaging using continuous wavelet transform on photodetector data. Their approach generated data rapidly and provided information of amplitude and phase for AFM probe with variation of sample materials.

4.2.2 Electron microscopy (EM)

In EM, an electron beam illuminates the sample to generate an image that provides critical information of surface characteristics and their detailed morphology. In EM imaging technique, chemists scan selected regions of the sample and assess the quality of the image based on their past experiences. If the chemist considers the EM scan as a poor-quality image, they shall change the conditions of the instrument and rescan another region of the specimen. Thus, most of the endeavour is based on trial- and error that is often time-consuming due to optimizing specimen region scan, probe type, voltage pulse between specimen and the probe for obtaining highly—resolved images.

Ilett et al. (2020) reported a validated automated agglomerate measurement for characterizing dispersion of nanoparticles in biological fluids using machine learning open-source software called ilastik and CellProfiler. Their approach utilized automated STEM imaging to obtain statistically relevant image data coupled with machine learning analysis. Further, the approach was extrapolated to confirm FeO nanoparticles agglomerate in cell culture medium that was deficient of surface-stabilising serum proteins. Yu et al. (2020) applied semantic image segmentation technique to analyze pore spaces of sandstone and its relationship with permeability characteristics. Their work demonstrated deep learning using neural networks precisely recognizing SEM images that led to improved identification of pores in sandstone samples. Wang et al. (2021) developed an unsupervised ML algorithm for automated transmission electron microscopic image analysis of metal nanoparticles. They explored the automated algorithm on palladium nanocubes and CdSe/CdS quantum dots that showed quantitative results.

4.2.3 Chromatography

Chromatography is a separation technique that involves partitioning of individual compounds of complex mixtures between mobile and stationary phases. This method of separation faces a problem of peak overlaps and analysing one type of data over time. With need to separate multiple samples with complex matrices led to development of 2D chromatography. 2D chromatography uses two chromatographic columns with different phases. During separation run, the sequential aliquots collected from the first chromatographic column are reinjected onto the second chromatographic column (Jones 2020). Thus, the components that could not be separated in the first column, get separated in the second column. The resulting data after separation is plotted in 2D or 3D space leading to complex data generation that is essentially solved using algorithms (Huygens et al. 2020).

Pérez-Cova et al. (2021) developed ROIMCR (Region of Interest Multivariate Curve Resolution) method for 2D liquid chromatographic separation method. Retention index (RI) is a critical parameter of chromatography that depends on the chemical structure and type of stationary phase employed during chromatographic separation. Several efforts are taken to determine retention indices that enhances the identification of analyte moelcules. Matyushin and Buryak (2020) utilized four machine learning models viz, 1D and 2D CNN, deep residual multilayer perceptron, and gradient boosting. They described molecules for input labels as strings notation, 2D representation, molecular fingerprints and descriptors in all the four machine learning models. The model was deployed and tested on flavoring agents, essential oils and metabolomic compounds of interest and exhibited error of about 0.8–2.2% only. Further, they utilized a free software, thereby demonstrating their models as being easily transferrable on a lab bench towards automation. Vrzal et al. (2021) proposed DeepReI model based on deep learning for accurate retention index prediction. They used SMILES notation as input labels and a predictive model of 2D CNN layers that had percentage error of $< 0.81\%$. Qu et al. (2021) described the training of graph neural networks to predict retention indices for NIST listed compounds and compared the results with earlier published work. They demonstrated that RI predictive, systematic and data-driven approach of deep learning outperforms previous machine learning models.

5 Challenges, opportunities and future perspectives

Organic synthesis, drug discovery and analytical techniques are no longer a sole human activity that requires numerous experiment protocols and reaction optimization. Even with a significant uptick of ML methods in chemistry, we are facing failures in applying them. As uncomfortable as it may sound, there are some serious problems which are presented as questions below and their subsequent reflections:

(1) How mature is the status of machine learning and chemometrics in chemistry?
(2) Are we training and deploying ML models in chemistry in the right manner? and;
(3) Can we completely automate our chemical laboratory bench?

It is already known that utility and application of ML models in chemistry rely heavily on quality and quantity of data. In most chemical experiments, protocols are based on previously optimized reaction conditions that lack reproducibility (Bergman and Danheiser 2016). Over the years, chemical data reproducibility issues are being addressed that

includes, reaction optimization with minimal information (Reker et al. 2020; Shields et al. 2021). Efforts are initiated in this direction by employing FAIR guidelines (Wilkinson et al. 2016) for chemical data. These guidelines are now transpiring as a research consortium amongst chemists for data sharing practices and fostering digital chemistry culture (Herres-Pawlis et al. 2019). Next, chemical process optimization that remained in dormancy is gradually showing progress as flow chemistry methods (Cherkasov et al. 2018). Mateos et al. (2019) reported continuous flow self-optimization platforms that included intelligent algorithms and monitoring techniques for a chemical reaction. Inspired by FAIR guidelines, two novel open-source machine learning benchmarking frameworks Summit (Felton et al. 2021) and Olympus (Häse et al. 2021) were reported for rapid optimization of reaction conditions. In same breath, it is reiterated that the emergence of ML in organic synthesis must not take away the elegance of discussing synthetic routes amongst chemists.

Though ML methods are transforming chemistry, yet these methods must not be exaggerated as we navigate on the Gartner hype cycle of AI (not a cycle, but a curve) (Gartner 2022). When Beker et al. (2022) investigated application of ML model on Suzuki–Miyaura reaction optimization, it was quite evident that data acquisition is a problem. Most of the data fed to machines are extracted from published journal papers and patents that are skewed towards high yielding reactions. Hence, the bias creeps in the data thereby, causing the ML deployment in organic synthesis planning dicey. As we advocate the success of AI in chemistry, we need to obtain reproducible data of high yielding reactions and standardize low yielding reactions. Utilizing and augmenting both the data sets is a better proposition, rather than merely feeding huge datasets of popular organic reactions. The same scenario holds true for drug discovery where, medicinal chemists are searching for drug molecules in infinity chemical space. Recalling Lipinski's idea of chemical space, medicinal chemists are utilizing rule of five (or Ro5) to search drug molecules (Lipinski 2016). DL methods are robust techniques when applied to drug discovery and repurposing. These holds promise in prediction modelling studies of emerging diseases for potential target identification. Medicinal chemists have plethora of choices to represent molecules. Apart from SMILES and SMARTS notations, Coulomb matrices, bag-of-bonds, fingerprinting and deep tensor networks are successfully implemented to find druggable molecules.

Another concern of experimental chemists is the failure to generate large datasets for "data crazy" ML models. It necessitated the application of transfer learning in chemistry that allowed algorithms to extract knowledge from the pre-trained model. Apart from a standard dataset, the pre-trained model with a similar application task as the target set is fed to the machine model to enhance performance. Few reports were published that trialled for applying transfer learning in chemical science (Tran et al. 2017; Wen et al. 2022). However, one cannot be fool-proof with transfer learning if the chemist chooses a pre-trained model dataset that has lower similarity index with the target set.

The final question is based on the premise that chemists are data generators whereas, computer scientists are programming experts. We are convinced that machines are good with images; hence their application on spectral, chromatogram and microscopic data is less problematic. The images are broken down to pixels and affixed a numeric value which is fairly easy yet, images generated are with artefacts. Artefacts are resolved easily with chemometric pre-processing methods prior to deploying ML models to extract crucial information. As AI-based models are good at deriving critical information from large high-quality data sets, it is possible to deploy them in atomic force microscopy, chromatography, and spectroscopy as discussed in Sect. 4. These sophisticated analytical methods have large data sets available for training and easily available to chemists. Machine learning models

are easily navigating in the different areas of analytical techniques, although cannot be fully automated, as the analytical instrument hardware are designed to be operated by humans.

Just as we wonder if AI is a dream for chemists, some path-breaking reports on mobile robots (Peplow 2014; Burger et al 2020; Fakhruldeen et al 2022) on a chemical lab bench brings our hype back. A chemical reaction robotic system controlled by machine-learning algorithm explored over 6000 organic reactions faster than those carried out by synthetic chemist's laboratory processes (Granda et al. 2018). All efforts discussed by far, are signalling towards digitization of chemical laboratories. However, automation in chemistry is not new, in fact the earliest attempt on chemical automation was demonstrated by Merrifield (1965) called solid phase peptide synthesis. Till date, solid phase peptide synthesis finds its applications in biochemical laboratories. We are not far from automated lab bench with sensory devices, IoT, digital twin and robust hardware in place. Yet, the scaling-up of the robotic work-flow from lab to industrial bench needs practical augmentation. We are in a triad of hope, disillusion and productivity when it comes to reflecting AI in drug discovery, organic synthesis and analytical methods, respectively (Fig. 6).

Digging further, most of the published literature lacks author diversity that go beyond gender. Few laboratories are working in silos on AI applications in chemistry, which is plausible, considering data privacy issues and funding constraints. Chemists, engineers, mathematicians and data scientists need to have a dialogue and solve the challenging problems of chemistry collaboratively. Automated robots in chemical laboratories are daunting task for scientists, especially those coming from middle-income nations, where grant funding is a problem. It is argued that, AI-based applications must go beyond borders and fruitfully contribute to the research community. One example is DREAM Challenges, a competition solving challenging problems in biomedicine that elicits the need for more such platforms. Such competitions, if explored for chemistry, shall stir up discussions leading to solving complex problems through diverse collaborations. Another

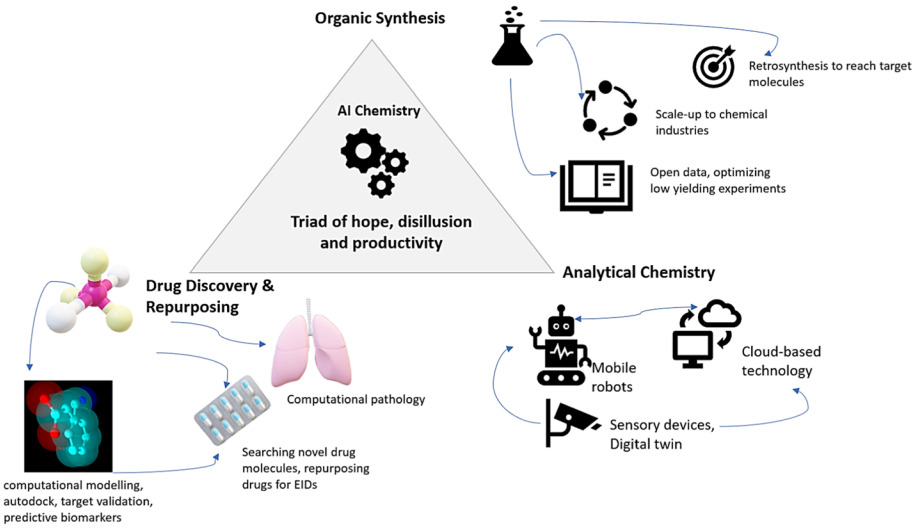

Fig. 6 Triad of hope, disillusion and productivity in drug discovery, organic synthesis and analytical chemistry, respectively

perspective is to introduce ML in chemistry curriculum that focuses only about solving chemistry. There are separate programs for machine learning and artificial intelligence, yet, these courses are curated for engineers rather than chemists. This effort of ML in chemistry curricula shall inspire young chemists to design their own machine algorithms to solve chemistry problems, without taking away the collaborative spirit of interdisciplinary AI research.

With a renaissance of Industry 4.0, chemometrics and machine learning have yet to explore and provide solutions to chemical problems. However, we are not far from reaching advanced ML-based solutions for the challenge. It is well understood that AI essentially derives power by learning from data; in this case, chemical data. If the flaws of data acquisition get resolved for chemical patterns, ML methods shall function more than an auxiliary-checkbox and navigate to explore the intricate chemical world.

Acknowledgements The author acknowledges library resources of NMIMS University, Mumbai and the anonymous reviewers for their insightful suggestions that helped to improve the paper significantly.

Funding Author did not receive any funding for the work.

References

Ahmed F, Lee JW, Samantasinghar A, Su Kim Y et al (2022) SperoPredictor: an integrated machine learning and molecular docking-based drug repurposing framework with use case of COVID-19. Front Public Health 10:902123. https://doi.org/10.3389/fpubh.2022.902123

Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG (2018) Predicting reaction performance in C-N cross-coupling using machine learning. Science 360(6385):186–190. https://doi.org/10.1126/science.aar5169

Akpolat H, Barineau M, Jackson KA, Akpolat MZ, Francis DM, Chen Y-J, Saona L (2020) High-throughput phenotyping approach for screening major carotenoids of tomato by handheld Raman spectroscopy using chemometric methods. Sensors 20(13):3723. https://doi.org/10.3390/s20133723

Alkhalifah Y, Phillips I, Soltoggio A, Darnley K, Nailon WH, McLaren D et al (2020) VOCCluster: untargeted metabolomics feature clustering approach for clinical breath gas chromatography/mass spectrometry data. Anal Chem 92(4):2937–2945. https://doi.org/10.1021/acs.analchem.9b03084

Allison MT, Degiacomi MT, Markland EG, Luca J, Elofsson A, Benesche JLP, Landreh M (2022) Complementing machine learning-based structure predictions with native mass spectrometry. Protein Sci 31(6):4333. https://doi.org/10.1002/pro.4333

Amilpur S, Bhukya R (2022) Predicting novel drug candidates against Covid-19 using generative deep neural networks. J Mol Graph Modell 110:108045. https://doi.org/10.1016/j.jmgm.2021.108045

Appel R, Hochstrasser D, Roch C, Funk M, Muller A, Pellegrini C (1988) Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. Electrophoresis 9(3):136–142. https://doi.org/10.1002/elps.1150090307

Baum ZJ, Yu X, Ayala PY, Zhao Y, Watkins SP, Zhou Q (2021) Artificial intelligence in chemistry: current trends and future directions. J Chem Inf Model 61(7):3197–3212. https://doi.org/10.1021/acs.jcim.1c00619

Beker W, Roszak R, Wołos A, Angello N et al (2022) Machine learning may sometimes simply capture literature: a case study of heterocyclic Suzuki−Miyaura. J Am Chem Soc 144(11):4819–4827. https://doi.org/10.1021/jacs.1c12005

Benstock JD, Berndt DJ, Agarwal KK (1988) Graph embedding in SYNCHEM2, an expert system for organic synthesis discovery. Discret Appl Math 19(1–3):45–63. https://doi.org/10.1016/0166-218X(88)90005-4

Bergman RG, Danheiser RL (2016) Reproducibility in chemical research. Angewante Chemie 55(41):12548–12549. https://doi.org/10.1002/anie.201606591

Bersohn M (1972) Automatic problem solving applied to synthetic chemistry. Bull Chem Soc Jpn 45(6):1897–1903. https://doi.org/10.1246/bcsj.45.1897

Blurock ES (1990) Computer-aided synthesis design at RISC-Linz: automatic extraction and use of reaction classes. J Chem Inf Comput Sci 30(4):505–510. https://doi.org/10.1021/ci00068a024
Bøgevig A, Federsel HJ, Huerta F, Hutchings MG, Kraut H et al (2015) Route design in the 21st century: the ICSYNTH software tool as an idea generator for synthesis prediction. Org Process Res Dev 19:357–368. https://doi.org/10.1021/op500373e
Burger B, Maffettone PM, Gusev VV et al (2020) A mobile robotic chemist. Nature 583:237–241. https://doi.org/10.1038/s41586-020-2442-2
Cadow J, Manica M, Mathis R, Guo T, Aebersold R, Martínez MR (2021) On the feasibility of deep learning applications using raw mass spectrometry data. Bioinformatics 37(1):i245–i253. https://doi.org/10.1093/bioinformatics/btab311
Chalmers JM (2006) Mid-infrared spectroscopy: anomalies, artifacts and common errors. In: Griffiths P, Chalmers JM (eds) Handbook of vibrational spectroscopy. Wiley, Chichester, p 4000. https://doi.org/10.1002/0470027320.s3101
Chen D, Wang Z, Guo D, Orekhov V, Qu X (2020) Review and prospect: deep learning in nuclear magnetic resonance spectroscopy. Chem A Euro J 26(46):10391–10401. https://doi.org/10.1002/chem.202000246
Cherkasov N, Bai Y, Expósito AJ, Rebrov EV (2018) OpenFlowChem—a platform for quick, robust and flexible automation and self-optimisation of flow chemistry. React Chem Eng 3:769–780. https://doi.org/10.1039/C8RE00046H
Chiutu C, Sweetman AM, Lakin AK, Stannard A, Jarvis S, Kantorovich L et al (2012) Precise orientation of a single C60 molecule on the tip of a scanning probe microscope. Phys Rev Lett 108(26):268302. https://doi.org/10.1103/PhysRevLett.108.268302
Choplin F, Laurenco C, Marc R, Kaufmann G, Wipke W (1978) Synthese assistee par ordinateur en chimie des composes organophosphores. N J Chem 2:285–293
Clerc JT, Ziegler E (1977) Computer techniques and optimization. In Analytica Chimica Acta. Elsevier, Amsterdam. https://doi.org/10.1016/S0003-2670(00)84991-0
Corey EJ (1967) General methods for the construction of complex molecules. Pure Appl Chem 14(1):19–38. https://doi.org/10.1351/pac196714010019
Corey EJ, Long AK, Rubenstein SD (1985) Computer-assisted analysis in organic synthesis. Science 228(4698):408–418. https://doi.org/10.1126/science.3838594
Corey EJ, Wipke WT (1969) Computer-assisted design of complex organic syntheses. Science 166(3902):178–192. https://doi.org/10.1126/science.166.3902.178
Cova TF, Pais AA (2019) Deep learning for deep chemistry: optimizing the prediction of chemical patterns. Front Chem 7(809):1–22. https://doi.org/10.3389/fchem.2019.00809
Dara S, Dhamercherla S, Jadav SS, Babu MC, Ahsan MJ (2022) Machine learning in drug discovery: a review. Artif Intell Rev 55(3):1947–1999. https://doi.org/10.1007/s10462-021-10058-4
Dragone V, Sans V, Henson BA, Granda JM, Cronin L (2017) An autonomous organic reaction search engine for chemical reactivity. Nat Commun 15733:1–8. https://doi.org/10.1038/ncomms15733
Dugundji J, Ugi I (1973) An algebraic model of constitutional chemistry as a basis for chemical computer programs. In: Houk K, Hunter C, Krische M, Lehn J, Ley S, Olivucci M et al (eds) Computers in chemistry, vol 39. Springer, Cham, pp 19–64
Ellermann L, Jauffret P, Ostermann C, Kaufmann G (1997) Evolution of the concept of synthesis strategy in the COSYMA system: introduction of the synthesis invariant. Liebigs Ann 1997(7):1401–1406. https://doi.org/10.1002/jlac.199719970717
Fakhruldeen, H., Pizzuto, G., Glowacki, J., & Copper, A. (2022). ARChemist: Autonomous Robotic Chemistry System Architecture. *IEEE International Conference on Robotics and Automation*, (p. 7). https://arxiv.org/abs/2204.13571
Felton KC, Rittig JG, Lapkin A (2021) Summit: benchmarking machine learning methods for reaction optimisation. Chem Methods 1:116–122. https://doi.org/10.1002/cmtd.202000051
Gallarati S, Fabregat R, Laplaza R, Bhattacharjee S, Wodrich MD, Corminboeuf C (2021) Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. Chem Sci 12:6879–6889. https://doi.org/10.1039/D1SC00482D
Gao H, Struble TJ, Coley CW, Wang Y, Green WH, Jensen KF (2018) Using machine learning to predict suitable conditions for organic reactions. ACS Cent Sci 4(11):1465–1476. https://doi.org/10.1021/acscentsci.8b00357
Gartner (2022) What's New in the 2022 Gartner Hype cycle for emerging technologies. https://www.gartner.com/en/articles/what-s-new-in-the-2022-gartner-hype-cycle-for-emerging-technologies. Accessed Aug 2022

Gasteiger J, Ihlenfeldt WD (1990) The WODCA system: an integrating environment for the chemist. In: Gasteiger J (ed) Software development in chemistry 4. Springer, Hochfilzen, pp 57–65. https://doi.org/10.1007/978-3-642-75430-2_7

Gasteiger J, Jochum C (1978) EROS A computer program for generating sequences of reactions. Top Curr Chem 74:93–126. https://doi.org/10.1007/BFb0050147

Ge Y, Tian T, Huang S, Wan F, Li J et al (2021) An integrative drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. Sig Transduct Target Ther 6:165. https://doi.org/10.1038/s41392-021-00568-6

Gelernter H, Rose R, Chen C (1990) Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning. J Chem Inf Comput Sci 30(4):492–504. https://doi.org/10.1021/ci00068a023

Gelernter HL, Sanders AF, Larsen DL, Agarwal KK, Boivie RH, Spritzer GA, Searleman JE (1977) Empirical explorations of SYNCHEM. Science 197(4308):1041–1049. https://doi.org/10.1126/science.197.4308.1041

Genheden S, Thakkar A, Chadimová V, Reymond JL, Engkvist O, Bjerrum E (2020) AizynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. J Cheminform 12:70. https://doi.org/10.1186/s13321-020-00472-1

Gomollón-Bel F (2019) Ten chemical innovations that will change our world: IUPAC identifies emerging technologies in chemistry with potential to make our planet more sustainable. Chem Int 41(2):12–17. https://doi.org/10.1515/ci-2019-0203

Granda JM, Donina L, Dragone V, Long D-L, Cronin L (2018) Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature 559:377–381. https://doi.org/10.1038/s41586-018-0307-8

Hanessian S, Franco J, Larouche B (1990) The psychobiological basis of heuristic synthesis planning—man, machine and the chiron approach. Pure Appl Chem 62(10):1887–1910. https://doi.org/10.1351/pac199062101887

Hansen DF (2019) Using deep neural networks to reconstruct non-uniformly sampled NMR spectra. J Biomol NMR 73:577–585. https://doi.org/10.1007/s10858-019-00265-1

Häse F, Aldeghi M, Hickman RJ, Roch LM, Christensen M, Liles E, Hein JE, Aspuru-Guzik A (2021) Olympus: a benchmarking framework for noisy optimization and experiment planning. Mach Learn Sci Technol 2:035021. https://doi.org/10.1088/2632-2153/abedc8

Herres-Pawlis S, Koepler O, Steinbeck C (2019) NFDI4Chem: shaping a digital and cultural change in chemistry. Angew Chem 58(32):10766–10768. https://doi.org/10.1002/anie.201907260

Huygens B, Efthymiadis K, Nowé A, Desmet G (2020) Application of evolutionary algorithms to optimise one- and two-dimensional gradient chromatographic separations. J Chromatogr A 1628:461435. https://doi.org/10.1016/j.chroma.2020.461435

IBM RXN for Chemistry (2018) https://rxn.res.ibm.com/ Accessed 15 April 2022

Ilett M, Wills J, Rees P, Sharma S, Micklethwaite S, Brown A et al (2020) Application of automated electron microscopy imaging and machine learning to characterise and quantify nanoparticle dispersion in aqueous media. J Microsc 279(3):177–184. https://doi.org/10.1111/jmi.12853

Javazm MR, Pishkenari HN (2020) Observer design for topography estimation in atomic force microscopy using neural and fuzzy networks. Ultramicroscopy 214:113008. https://doi.org/10.1016/j.ultramic.2020.113008

Jones O (2020) Two-dimensional liquid chromatography: principles and practical applications. Springer, Cham

Jorgensen WL, Laird ER, Gushurst AJ, Fleischer JM, Gothe SA, Helson HE et al (1990) CAMEO: a program for the logical prediction of the products of organic reactions. Pure Appl Chem 62(10):1921–1932. https://doi.org/10.1351/pac199062101921

Karunanithy G, Hansen DF (2021) FID-Net: a versatile deep neural network architecture for NMR spectral reconstruction and virtual decoupling. J Biomol NMR 75:179–191. https://doi.org/10.1007/s10858-021-00366-w

Kim DE, Zweig JE, Newhouse TR (2019) Total synthesis of paspaline A and emindole PB enabled by computational augmentation of a transform-guided retrosynthetic strategy. J Am Chem Soc 141(4):1479–1483. https://doi.org/10.1021/jacs.8b13127

Klucznik MB, Gołębiowska P, Bayly AA et al (2020) Computational planning of the synthesis of complex natural products. Nature 588:83–88. https://doi.org/10.1038/s41586-020-2855-y

Klucznik T, Klucznik B, McCormack MP, Lima H, Szymkuć S, Bhowmick M et al (2018) Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. Chem 4(3):522–532. https://doi.org/10.1016/j.chempr.2018.02.002

Kolluri S, Lin J, Liu R, Zhang Y, Zhang W (2022) Machine learning and artificial intelligence in pharmaceutical research and development: a review. AAPS J 24(1):19. https://doi.org/10.1208/s12248-021-00644-3

Kondo M, Wathsala H, Sako M, Hanatani Y, Ishikawa K, Hara S et al (2020) Exploration of flow reaction conditions using machine-learning for enantioselective organocatalyzed Rauhut-Currier and [3+2] annulation sequence. Chem Commun 56(8):1259–1262. https://doi.org/10.1039/C9CC08526B

Kong X, Zhou L, Li Z, Yang Z, Qiu B, Wu X et al (2020) Artificial intelligence enhanced two-dimensional nanoscale nuclear magnetic resonance spectroscopy. NPJ Quantum Inf. https://doi.org/10.1038/s41534-020-00311-z

Kowalski BR (1975) Chemometrics: views and propositions. J Chem Inf Comput Sci 15(4):201–203. https://doi.org/10.1021/ci60004a002

Krull A, Hirsch P, Rother C, Schiffrin A, Krull C (2020) Artificial-intelligence-driven scanning probe. Commun Phys 3:54. https://doi.org/10.1038/s42005-020-0317-3

Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S et al (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 581:215–220. https://doi.org/10.1038/s41586-020-2180-5

Lederberg, J. (1964). DENDRAL-64: a system for computer construction, enumeration and notation of organic molecules as tree structures and cyclic graphs. Part I. Notational algorithm for tree structures. pp. 1–33. http://resource.nlm.nih.gov/101584906X879

Li D-W, Leggett A, Bruschweiler-Li L, Brüschweiler R (2022) COLMARq: a web server for 2D NMR peak picking and quantitative comparative analysis of cohorts of metabolomics samples. Anal Chem 94(24):8674–8682. https://doi.org/10.1021/acs.analchem.2c00891

Li Shuang X, Liu X, Lu L, Sheng Hua X, Chi Y, Xia K (2022) Multiphysical graph neural network (MP-GNN) for COVID-19 drug design. Brief Bioinformatics 23(4):231. https://doi.org/10.1093/bib/bbac231

Lin K, Xu Y, Lai PJ, L, (2020) Automatic retrosynthetic route planning using template-free models. Chem Sci 11(12):3355–3364. https://doi.org/10.1039/C9SC03666K

Lipinski C (2016) Rule of five in 2015 and beyond: target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. Adv Drug Deliv Rev 101:34–41. https://doi.org/10.1016/j.addr.2016.04.029

Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. Nature 432:855–861. https://doi.org/10.1038/nature03193

Liu B, Ramsundar B, Kawthekar P, Shi J, Gomes J et al (2017) Retrosynthetic reaction prediction using neural sequence-to sequence models. ACS Cent Sci 3(10):1103–1113. https://doi.org/10.1021/acscentsci.7b00303

Marth CJ, Gallego GM, Lee JC, Lebold TP, Kulyk S, Kou K et al (2015) Network-analysis-guided synthesis of weisaconitine D and liljestrandinine. Nature 528:493–498. https://doi.org/10.1038/nature16440

Martyna A, Menżyk A, Damin A, Michalska A, Martra G, Alladio E, Zadora G (2020) Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components. Chemom Intell Lab Syst 202:104029. https://doi.org/10.1016/j.chemolab.2020.104029

Massart DL, Vandeginste B, Buydens L, De JS, Lewi PJ, Smeyers-Verbeke J (1997) Handbook of chemometrics and qualimetrics, Part A. Elsevier, Amsterdam

Mateos C, Nieves-Remacha M, Rincón JA (2019) Automated platforms for reaction self-optimization in flow. React Chem Eng 4:1536–1544. https://doi.org/10.1039/C9RE00116F

Matyushin DD, Buryak AK (2020) Gas chromatographic retention index prediction using multimodal machine learning. IEEE Access 8:223140–223155. https://doi.org/10.1109/ACCESS.2020.3045047

Mehta G, Barone R, Chanon M (1998) Computer-aided organic synthesis—SESAM: a simple program to unravel "Hidden" restructured starting materials Skeleta in complex targets. Euro J Organ Chem 1998(7):1409–1412. https://doi.org/10.1002/(SICI)1099-0690(199807)1998:7%3c1409::AID-EJOC1409%3e3.0.CO;2-H

Melnikov AD, Tsentalovich YP, Yanshole VV (2020) Deep learning for the precise peak detection in high-resolution LC–MS data. Anal Chem 92(1):588–592. https://doi.org/10.1021/acs.analchem.9b04811

Mensa S, Sahin E, Tacchino F, Barkoutsos PK, Tavernelli I (2022) Quantum machine learning framework for virtual screening in drug discovery: a prospective quantum advantage. arXiV Preprint. https://doi.org/10.48550/arXiv.2204.04017

Merrifield RB (1965) Automated synthesis of peptides. Science 150(3693):178–185. https://doi.org/10.1126/science.150.3693.178

Mohanty S, Rashid M, Mridul M, Mohanty C, Swayamsiddha S (2020) Application of artificial intelligence in COVID-19 drug repurposing. Diabetes Metab Syndr 14(5):1027–1031. https://doi.org/10.1016/j.dsx.2020.06.068

Nam J, Kim J (2016) Linking the Neural machine translation and the prediction of organic chemistry reactions. *arXiV* , pp. 1–19. https://arxiv.org/pdf/1612.09529.pdf

Pacheco-Londoño LC, Warren E, Galán-Freyle N, Villarreal-González R, Aparicio-Bolaño JA, Ospina-Castro ML et al (2020) Mid-infrared laser spectroscopy detection and quantification of explosives in soils using multivariate analysis and artificial intelligence. Appl Sci 10(12):4178. https://doi.org/10.3390/app10124178

Papagiannopoulou C, Parchen R, Rubbens P, Waegeman W (2020) Fast pathogen identification using single-cell matrix-assisted laser desorption/ionization-aerosol time-of-flight mass spectrometry data and deep learning methods. Anal Chem 92(11):7523–7531. https://doi.org/10.1021/acs.analchem.9b05806

Parrot M, Tajmouati H, Barros Ribeiro da Silva V et al (2021) Integrating synthetic accessibility with AI-based generative drug design. ChemRxiv. https://doi.org/10.26434/chemrxiv-2021-jkhzw-v2

Paul W, Oliver D, Miyahara Y, Grütter P (2014) FIM tips in SPM: apex orientation and temperature considerations on atom transfer and diffusion. Appl Surf Sci. https://doi.org/10.1016/j.apsusc.2014.03.002

Payam AF, Biglarbeigi P, Morelli A, Lemoine P, McLaughlin J, Finlay D (2021) Data acquisition and imaging using wavelet transform: a new path for high speed transient force microscopy. Nanoscale Adv 3(2):383–398. https://doi.org/10.1039/D0NA00531B

Pensak DA, Corey EJ (1977) LHASA—Logic and Heuristics Applied to Synthetic Analysis. In: Wipke WT, Howe J (eds) Computer-assisted organic synthesis; ACS Symposium Series, vol 61. American Chemical Society, Washington, pp 1–32. https://doi.org/10.1021/bk-1977-0061.ch001

Peplow M (2014) Organic synthesis: the robo-chemist. Nature 512:20–22. https://doi.org/10.1038/512020a

Pérez-Cova M, Jaumot J, Tauler R (2021) Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches. TrAC Trends Anal Chem 137:116207. https://doi.org/10.1016/j.trac.2021.116207

Pham T-H, Qui Y, Zeng J, Xie L, Zhang P (2021) A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. Nat Mach Intell 3:247–257. https://doi.org/10.1038/s42256-020-00285-9

Pillai N, Dasgupta A, Sudsakorn S, Fretland J, Mavroudis PD (2022) Machine learning guided early drug discovery of small molecules. Drug Discov Today 27(8):2209–2215. https://doi.org/10.1016/j.drudis.2022.03.017

PostEra Manifold (2021) https://app.postera.ai/. Accessed 8 Sept 2021

Qiu F, Lei Z, Sumner LW (2018) MetExpert: an expert system to enhance gas chromatography–mass spectrometry-based metabolite identifications. Anal Chim Acta 1037:316–326. https://doi.org/10.1016/j.aca.2018.03.052

Qu C, Schneider BI, Kearsley AJ, Keyrouz W, Allison TC (2021) Predicting Kováts retention indices using graph neural networks. J Chromatogr A 1646:462100. https://doi.org/10.1016/j.chroma.2021.462100

Rahimi M, Lee Y, Markley JL, Lee W (2021) iPick: multiprocessing software for integrated NMR signal detection and validation. J Magn Reson 328:106995. https://doi.org/10.1016/j.jmr.2021.106995

Reker D, Hoyt EA, Bernardes GJ (2020) Adaptive optimization of chemical reactions with minimal experimental information. Cell Rep Phys Sci 1(11):100247. https://doi.org/10.1016/j.xcrp.2020.100247

Roger JM, Biancolillo A, Marini F (2020) Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. Chemom Intell Lab Syst 199:103975. https://doi.org/10.1016/j.chemolab.2020.103975

Salatin TD, Jorgensen WL (1980) Computer-assisted mechanistic evaluation of organic reactions. 1. Overview. J Organ Chem 45(11):2043–2051. https://doi.org/10.1021/jo01299a001

Salin ED, Winston PH (1992) Machine learning and artificial intelligence an introduction. Anal Chem 64(1):49A-60A. https://doi.org/10.1021/ac00025a742

Santos PP, Heggie W (2020) Gabapentin. In: Santos PP, Heggie W (eds) Retrosynthesis in the manufacture of generic drugs: selected case studies, 1st edn. Wiley, New Jersey, pp 7–9. https://doi.org/10.1002/9781119155560.ch2

Satoh H, Funatsu K (1995) SOPHIA, a knowledge base-guided reaction prediction system—utilization of a knowledge base derived from a reaction database. J Chem Inf Model 35(1):34–44. https://doi.org/10.1021/ci00023a005

Satoh K, Funatsu K (1999) A novel approach to retrosynthetic analysis using knowledge bases derived from reaction databases. J Chem Inf Comput Sci 39(2):316–325. https://doi.org/10.1021/ci980147y

Schull G, Frederiksen T, Arnau A, Sánchez-Portal D, Berndt R (2011) Atomic-scale engineering of electrodes for single-molecule contacts. Nat Nanotechnol 6:23–27. https://doi.org/10.1038/nnano.2010.215

Schwaller P, Gaudin T, Lányi D, Bekas C, Laino T (2018) "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chem Sci 9:6091–6098. https://doi.org/10.1039/C8SC02339E

Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, Lee AA (2019) Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 5(9):1572–1583. https://doi.org/10.1021/acscentsci.9b00576

Shehab M, Abualigah L, Shambour Q, Abu-Hashem M, Shambour M, Alsalibi AI, Gandomi AH (2022) Machine learning in medical applications: a review of state-of-the-art methods. Comput Biol Med 145:105458. https://doi.org/10.1016/j.compbiomed.2022.105458

Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JI et al (2021) Bayesian reaction optimization as a tool for chemical synthesis. Nature 590:89–96. https://doi.org/10.1038/s41586-021-03213-y

Sibilio P, Bini S, Fiscon G, Sponziello M, Conte F, Pecce V et al (2021) In silico drug repurposing in COVID-19: A network-based analysis. Biomed Pharmacother 142:111954. https://doi.org/10.1016/j.biopha.2021.111954

Socorro IM, Taylor K, Goodman JM (2005) ROBIA: a reaction prediction program. Organic Lett 7(16):3541–3544. https://doi.org/10.1021/ol0512738

Stark SA, Neudert R, Threlfall R (2016) ChemPlanner predicts experimentally verified synthesis routes in medicinal chemistry CHEManager, Wiley. https://www.chemanager-online.com/en/whitepaper/wiley-chemplanner-predicts-experimentally-verified-synthesis-routes-medicinal-chemistry. Accessed 27 Jan 2022

Sternberg MJ, Lewis R, King R, Muggleton S (1992) Modelling the structure and function of enzymes by machine learning. Faraday Discuss 93:269–280. https://doi.org/10.1039/FD9929300269

Tanaka A, Okamoto H, Bersohn M (2010) Construction of functional group reactivity database under various reaction conditions automatically extracted from reaction database in a synthesis design system. J Chem Inf Model 50:327–338. https://doi.org/10.1021/ci9004332

Tantillo DJ (2018) Questions in natural products synthesis research that can (and cannot) be answered using computational chemistry. Chem Soc Rev 47(21):7845–7850. https://doi.org/10.1039/c8cs00298c

Torniainen J, Afara IO, Prakash M, Sarin JK, Stenroth L, Töyräs J (2020) Open-source python module for automated preprocessing of near infrared spectroscopic data. Anal Chim Acta 1108:1–9. https://doi.org/10.1016/j.aca.2020.02.030

Tran HA, Ramsundar B, Pappu AS, Pande V (2017) Low data drug discovery with one-shot learning. ACS Cent Sci 3(4):283–293. https://doi.org/10.1021/acscentsci.6b00367

Vasas M, Tang F, Hatzakis E (2021) Application of NMR and chemometrics for the profiling and classification of ale and lager American craft beer. Foods 10(4):807. https://doi.org/10.3390/foods10040807

Vasighi M, Romanova J, Nedyalkova M (2022) A multilevel approach for screening natural compounds as an antiviral agent for COVID-19. Comput Biol Chem 98:107694. https://doi.org/10.1016/j.compbiolchem.2022.107694

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. (2017) Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 6000–6010). Curran Associates Inc. Retrieved from https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Villarrubia JS (1997) Algorithms for scanned probe microscope image simulation, surface reconstruction, and tip estimation. J Res Nat Inst Stand Technol 102(4):425–454. https://doi.org/10.6028/jres.102.030

Viveros N, Gil P, Ramos J, Núñez H (2021) On the estimation of sugars concentrations using Raman spectroscopy and artificial neural networks. Food Chem 352:129375. https://doi.org/10.1016/j.foodchem.2021.129375

Vrzal T, Malečková M, Olšovská J (2021) DeepReI: deep learning-based gas chromatographic retention index predictor. Anal Chim Acta 1147:64–71. https://doi.org/10.1016/j.aca.2020.12.043

Wang X, Li J, Ha HD, Dahl JC, Ondry JC, Moreno-Hernandez I et al (2021) AutoDetect-mNP: an unsupervised machine learning algorithm for automated analysis of transmission electron microscope images of metal nanoparticles. JACS Au 1(3):316–327. https://doi.org/10.1021/jacsau.0c00030

Watson IA, Wang J, Nicolaou CA (2019) A retrosynthetic analysis algorithm implementation. J Cheminform 11:1. https://doi.org/10.1186/s13321-018-0323-6

Welker J, Giessibl FJ (2012) Revealing the angular symmetry of chemical bonds by atomic force microscopy. Science 336(6080):444–449. https://doi.org/10.1126/science.1219850

Wen M, Samuel MB, Xie X, Dwarkanath S, Persson KA (2022) Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. Chem Sci 13:1446–1458. https://doi.org/10.1039/D1SC06515G

Weng S, Yuan H, Zhang X, Li P, Zheng L, Zhao J, Huang L (2020) Deep learning networks for the recognition and quantitation of surface-enhanced Raman spectroscopy. Analyst 145:4827–4835. https://doi.org/10.1039/D0AN00492H

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Wöhler F (1828) Ueber künstliche Bildung des Harnstoffs. Ann Phys 88(2):253–256. https://doi.org/10.1002/andp.18280880206

Xie G, Xu H, Li J, Gu G, Sun Y et al (2022) DRPADC: a novel drug repositioning algorithm predicting adaptive drugs for COVID-19. Comput Chem Eng. https://doi.org/10.1016/j.compchemeng.2022.107947

Yang W-L, Li Q, Sun J, Tan S, Tang Y-H, Zhao MM et al (2022a) Potential drug discovery for COVID-19 treatment targeting Cathepsin L using a deep learning-based strategy. Comput Struct Biotechnol J 20:2442–2454. https://doi.org/10.1016/j.csbj.2022.05.023

Yang Y, Zhou D, Zhang X, Shi Y, Han J et al (2022b) D3AI-CoV: a deep learning platform for predicting drug targets and for virtual screening against COVID-19. Brief Bioinform 23(3):bbac147. https://doi.org/10.1093/bib/bbac147

Yu Q, Xiong Z, Du C, Dai Z, Soltanian MR, Soltanian M et al (2020) Identification of rock pore structures and permeabilities using electron microscopy experiments and deep learning interpretations. Fuel 268:117416. https://doi.org/10.1016/j.fuel.2020.117416

Zhao MM, Yang WL, Yang F-Y, Zhang L, Huang W, Hou W et al (2021) Cathepsin L plays a key role in SARS-CoV-2 infection in humans and humanized mice and is a promising target for new drug development. Signal Transduct Target Ther 6(1):134. https://doi.org/10.1038/s41392-021-00558-8