WEBSITE: chem-tox-ecotox.org/ScientificReviews ……5 May 2020

SCIENTIFIC REVIEWS

Synthetic Organic Chemistry and the Emergence of Artificial Intelligence-Driven Technology to Synthesize Target Chemical Compounds

Athanasios Valavanidis
Departmenty of Chemistry, National and Kapodistrian University of Athens, University Campus Zografou, 15784 Athens, Greece
E-mail : Valavanidis@chem.uoa.gr

Abstract. Organic synthesis in the last century was a scientific and experimental branch involved in the practical synthesis of organic chemicals from simple chemical reagents to highly compicated molecular structures for the chemical industry and for research purposes. The synthesis of organic molecules has developed into one of the most important branches of organic chemistry and a cornestone sector of the economy worldwide. Organic chemicals are used for plastics, rubber materials, pharmaceuticals, cosmetics, detergents, coatings, dyestuff, agrichemical chemicals and a variety of materials for the heavy industry and commercial businesses. Also, synthetic organic chemicals are important for biochemistry, biotechnology, medicinal products and health treatments. With the emergence of artificial intelligence-driven technology, the challenge of generalized predictions of organic reaction outcomes and streamlined reaction optimization has gained significant importance. Artificial Intelligence (AI) is considered one of the most important disruptive innovations with the potential to revolutionize research in synthetic organic chemistry. Also computer programmes supported by AI and specialized deep learning techniques can help organic chemists to predict mechanisms, reactivity, yields and final products of given reagents, aiming to accelerate the outcome of organic synthetic reactions. The advent of AI and algorithms to automatize, improve, and generalize organic reaction prediction is gaining importance in parallel with conventional synthetic methodologies. There are now several published studies in this area that reviewed and explored the results of experimental research with connection to advanced methods of AI. In the last decade scientists employed AI techniques and deep reinforcement learning to optimize organic chemical reactions, giving researchers the opportunity to use model iteratively records leading to new experimental conditions to improve organic reaction outcomes. This review collected some interesting studies from the scientific literature of the last decade on the subject of AI and organic reaction synthesis.

Introduction: Organic synthesis a creativity-based scientific discipline

Organic synthesis in chemistry is a fundamental scientific and experimental discipline which has been in existence for more than a century. During this period a great number of synthetic techniques and technologies were developed and are available to synthetic chemists for designing appropriate routes to synthesize target chemical compounds. Synthesis of organic compounds is central to the chemical industry, which produces useful materials and contributes to the economic growth of the most developed countries. Organic materials are used in the rubber industry, for the synthesis of plastics, new pharmaceuticas, cosmetics, detergents, coatings, dyestuff, and agrichemical chemicals. But also, organic synthesis is important for the foundations of biochemistry, biotechnology, and medicine through highly specialised organic compounds playing substantial role in life processes. Many modern, high-tech materials are at least partially composed of organic compounds. Organic chemists spend much of their time creating new compounds and developing better ways of synthesizing previously known compounds. The organic chemical industry plays an important role in modern world economies by converting raw feedstock materials into more than 80,000 different chemical products. It must be emphasised that 75% of the chemical industry's output worldwide is polymers and plastics. Chemicals are used to make a wide variety of consumer goods, as well as thousands of products that are inputs to the agriculture, manufacturing, construction, and all types of service industries.[1-4]

Organic synthesis has a long tradition as a creativity-based and meticulous scientific discipline, and the total synthesis of complex molecules is often referred to as "the art of synthesis".[5] A key limitation of organic synthesis still lies in the tedious reaction optimization for single steps and route adaptations, which are on the one hand important for learning and the advancement of science but on the other hand hamper rapid synthesis. Organic synthetic chemists start typical synthesis design with scientific literature search, retrosynthetic analysis, identification of synthesis conditions,

and finally the practical organic synthesis in a well equipped chemical laboratory and spectroscopic methodology for indentification.[6]

Synthesis in Organic Chemistry involves a variety of experimental methodologies and research involving the fundametal stages in the synthesis of organic compounds: discovery, optimisation, and studies of scope and limitations. All these stages require extensive knowledge of and experience with chemical reactivities of appropriate reagents. Optimisation is a methodological process in which one or two starting compounds are tested in the reaction under a wide variety of conditions of temperature, use of solvent to dissolve the reagents, reaction time, etc., until the optimal conditions for the highest yield of the final product and its purity are found. Finally, the researcher tries to extend the method to a broad range of different starting materials, to find the scope and limitations.[7-10]

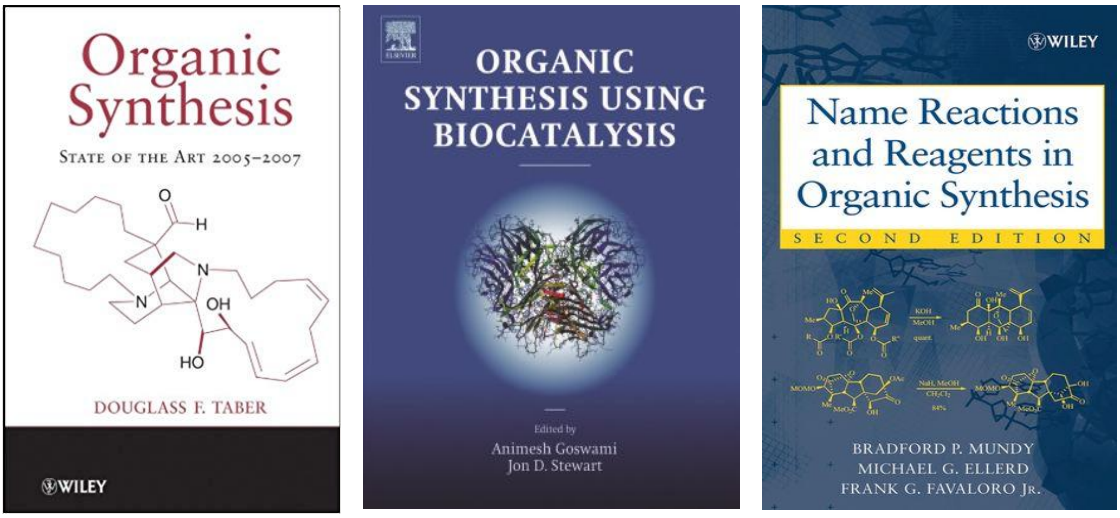


Figure 1. There are numerous textbooks and specialised books on Organic Chemistry and in particular for Organic Syntheis of chemical compounds.

In the 20th century one of the revolutionary advances were the approach to synthesis design, based on retrosynthetic analysis, for which Prof. Corey won the Nobel Prize for Chemistry in 1990 (Prof. Elias James Corey, Harvard University, MA, USA). The organic synthesis was planned backwards from the product, using standard rules. The retrosynthetic methodology was the most important contribution to synthetic organic chemistry that have made it possible to produce a large variety of biologically

highly active, complicated natural products, thereby making, among other things, certain pharmaceuticals commercially available. [7]

Organic synthetic chemistry and the emergence of artificial intelligenece (AI) methodologies

With the emergence of artificial intelligence-driven technology, the challenge of generalized predictions of organic reaction outcomes and thus streamlined reaction optimization has gained significant interest in organic synthesis. Artificial Intelligence (AI) is considered one of the most important disruptive innovations with the potential to revolutionize everyday life in a similar fashion and scale as the innovation of internet and communications technology.[11]

Artificial Intelligenece (AI) is not a new field of research in organic chemistry, as chemists have been using computers for years for their reaesearch in organic synthetic activities. Preliminary research studies started for more than 50 years ago with the DENDRAL project. Professors Corey and Wipke envisioned that both synthesis and retrosynthesis could be designed by a computer machine using handcrafted rules known as reaction templates.However, a deep chemical expertise was still required, and writing rules remained a time-consuming task.[12-14]

With the advances in the last decade of automated AI-driven reaction methodology, became clear to organic scientists that total synthesis of complex molecules comprising fully automated multistep synthesis can be regarded as the next evolutionary step. It is noteworthy that a subtle differentiation between automated and autonomous synthesis is needed. While the latter describes a self-governing synthesis that can react to surrounding parameters (reaction yield and stereoselectivity) independently without human input, automated synthesis requires human input to define boundaries, thresholds, reaction parameters, and synthesis protocols. For the purpose of automated multistep synthesis, continuous-flow chemistry is probably one of the key technologies to drive multistep synthesis and once optimized only limited human input is required. The combination of AI-driven

reaction optimization, robotics, and a modular continuous-flow platform has been proposed as a practical apporach to organic step-by-step of organic synthesis. In 2015, Whitesides already claimed in his essay *Reinventing Chemistry* …. "*Chemistry is ending an era of extraordinary intellectual growth and commercial contribution to society, powered by an explosion of science and technology, and a parallel and mutually beneficial expansion of academic and industrial chemistry. The close links between the two were mutually deeply beneficial. This prolific era is over, and chemistry is now facing new opportunities, and obligations to society, that are even more interesting, but entirely different. These new opportunities are, however, much broader in scope and greater in complexity than the simpler, previous problems, and require new structures and methods. Chemistry is no longer just about atoms and molecules, but a field with unique capabilities in manipulating molecules and matter, can do to understand, manipulate, and control complex systems composed (in part) of atoms and molecules: its future extends from living cells to megacities, and from harvesting sunlight to improving healthcare. To deal efficiently with these problems, academic chemistry will need to integrate "solving problems" and "generating understanding" better. It should teach students the skills necessary to attack problems that do not even exist as problems when the students are being taught. Industry must either augment its commodity- and service-based model to re-engage with invention, or face the prospect of settling into a corner of an industrial society that is comfortable, but largely irrelevant to the flows of technology that change the world….".*[15]

Scientists use already sophisticated computer programmes (artificial intelligence, AI, deep learning, etc) and robotics that can have a fundamental impact on synthetic organic chemistry. The scientific literature is full of papers and methods in showcase organic synthesis with the use artificial intelligence-driven and automated synthetic methodology. Organic chemistry scientists have faced for a long time the challenge to predict the outcome of complex chemical transformations by computer-assisted technological applications.[16,17]

The development of quantum-chemical approache has already opened some opportunities in this direction, and in many cases, the outcomes of organic synthesis experiments can be efficiently modeled by computer programmes (*in silico*).[18,19]

Approaches of computational prediction of chemical reactivity usually requires expert knowledge. Also, at present there are relatively few computational tools that can be used by a bench chemist to help guide organic synthesis. One example is the RegioSQM method for predicting the regioselectivity of electrophilic aromatic substitution reactions of heteroaromatic systems. RegioSQM has been proved a good predictive tool to guide organic synthesis.[20]

Organic stereoselectivity by computer–assisted methods has been advanced as the processing power of computers has increased substantially and can be used as a predictive tool in the discovery of new asymmetric catalysts. There are recent examples in the scientific literature of how the fundamental principles and application of state-of-the-art computational methods may be used to gain mechanistic insight into organic and organometallic reactions.[21] There are plenty of studies and examples of computational chemistry that has become an established tool for the study of the origins of chemical phenomena and examination of molecular properties. Because of major advances in theory, hardware and software, calculations of molecular processes can predict reactivities for applications in organic sythetic methods.[22]

Organic chemists know from their experimental experience that a major challenge in organic synthesis is the reaction prediction, if it possible among the reactants or energetically unfavorable to proceed under certain conditions. So, for organic synthetic chemist is desirable to develop algorithms that will “learn” (deep learning techniques) from being exposed to examples of the application of the rules of organic chemistry. In a recent paper scientists explored the use of neural networks for predicting reaction types, using a new reaction fingerprinting method. They combined this predictor with SMARTS transformations (SMARTS is a computer language that allows scientists to specify substructures using rules that are straightforward extensions of

SMILES, and SMILES, acronym, Simplified Molecular-Input Line-Entry System, is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings). SMILES is now the most widely used chemical line notation] to build a system which, given a set of reagents and reactants, predicts the likely products. The advent of artificial intelligence (AI) algorithms to automatize, improve, and generalize organic reaction predictions is gaining importance in this field, and several recent studies have been published in this area. For example, in 2016, Aspuru-Guzik and co-workers reported their attempt to apply neural networks to basic reactions of alkenes and alkyl halides, and they were able to identify the correct reaction type for the majority of a set of textbook problems.[23]

Figure 2. Books on artificial intelligence and organic synthesis. Hippe Z. *Artificial Intelligence in Chemistry. Structure Elucidation and Simulation of Organic Reactions*. Elsevier, Amsterdam, 1991. Linsay RK., et al. Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL project (McGraw-Hill advanced computer science series). McGraw Hill, New York, 1980. Brown N (Ed). *Artificial Intelligence in Drug Discovery*. Royal Society of Chemistry publs, Cambridge, UK, 2020,

Researchers from Warsaw University, Poland (Gambin and co-workers) tested Artificial Intelligenece (AI) algorithms to predict a large set (450, 000 cases) of manifold organic reactions, emphasizing that it might be essential to identify new chemoinformatic descriptors for future developments. The paper demonstrates that the applicability of machine learning to the

problems of chemical reactivity over diverse types of chemistries remains limited. Scientists noticed that with the currently available chemical descriptors, fundamental mathematical theorems impose upper bounds on the accuracy with which raction yields and times can be predicted. Improving the performance of machine-learning methods calls for the development of fundamentally new chemical descriptors.[24]

In another example deep reinforcement learning was employed to optimize organic chemical reactions. Researchers use model iteratively records the results of a chemical reaction and chooses new experimental conditions to improve the reaction outcome. This model outperformed a state-of-the-art methodology. This was another important attempt to predict and optimize organic reactions on the basis of Artificial Intelligence (AI), in recent studies by the group of Zare.[25]

Although computer assistance in organic synthesis design has existed for over 40 years, yet retrosynthesis planning software has struggled to achieve widespread adoption. One critical challenge in developing high-quality pathway suggestions is that proposed reaction steps often fail when attempted in the laboratory, despite initially seeming viable. The true measure of success for any synthesis program is whether the predicted outcome matches what is observed experimentally. Researchers report a model framework for anticipating reaction outcomes with noteworthy examples.[26]

Although organic synthetic chemistry has been advanced for over decades into modern synthetic methods that can provide access to molecules of considerable complexity, predicting the outcome of a single chemical reaction remained a major challenge. Although the first computer programmes to design organic syntheses emerged around the 1960s they failed to capture the imagination of organic chemists. Synthesis laboratories have remained sceptical of the ability of computer programs to learn the ‘art’ of organic chemistry. But now the digitization of multistep organic synthesis is fast approaching and the selection of reaction conditions is a key element of the automation of the synthesis planning. Recently, machine-learning-based tools have been developed that provide information on route planning for a target molecule. These algorithms are trained on the chemical literature, learning the

'rules and reasoning' of synthesis, and then predict a suitable synthetic routes.[27,28,29]

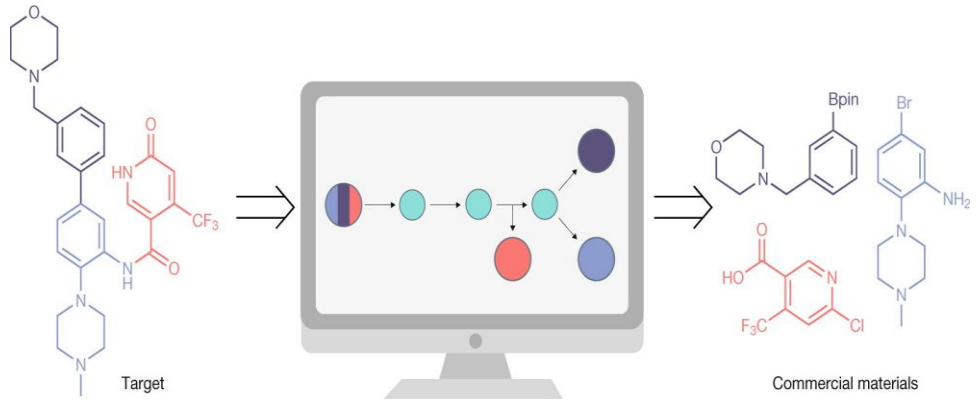


Figure 3. A schematic organic synthetic route is predicted from commercial materials to give the desired target molecule. These critical advances in machine-aided synthesis are still limited in their application to more complex molecules such as natural products, as well as in dealing with the intricacies of medicinal and process chemistry. They rely on the datasets published in journal articles, which represent only a fraction of the raw data collected in a given research project [Davies W.The digitization of organic synthesis. *Nature* 570:175-181, 2019. https://www.nature.com/ articles/s41586-019-1288-y].

Although the predictions of organic reactions had some limitations, in general, the AI algorithms showed an encouragingly good performance even for sophisticated organic systems. Already machine-learning methods are becoming integral part to scientific inquiry in synthetiv organic reactions in multidimensional chemical space using data obtainaed via high-throughput experimentation. Researchers created scripts to compute and extract atomic, molecular and vibrational descriptors for the components of a palladium-catalyzed Buchward-Hartwig cross coupling of aryl halides with 4-methylaniline. Using these descriptor as inputs and reaction yield as output, the results showed that a random forest algorithm provides significanty improved predictive performance over linear regression analysis.[30,31]

Organic synthetic chemists predict that applications of artificial intelligenece (AI) methodologies will make in the futture great progress in the acceleration of organic synthetic process of useful chemicals. There is little doubt to many scientists in this field that these AI systems, once fully operational in organic chemistry, will dramatically speed up development of new chemical materials and innovative drugs in the near future.[32]

Web-based databases for chemistry information and organic synthetic reactions

There is a great number of chemical databases specifically designed to store chemical information, such chemical structure, spectroscopic data (NMR, FTIR, mass spectrometry, etc), pathways of chemical reactions, synthetic mechanisms, crystal structures, and thermophysical data.

There is a great variety of chemical structure databases. Some of these chemical databases are free to use. Example, PubChem, ChemExper, ChemSpider, Emolecules etc., are very useful databases to locate compound suppliers or to find biological/physical properties of chemical substances. But only two databases are currently providing synthesis or synthesis references free of charge: OrgSyn and Heterocycles. These websites are very useful for organic synthetic chemists. There is only one disadvantage: the data in these databases is limited to a few journals. [Services for Research Chemistry Databases, https://research.csc.fi/chemistry/databases].

The most important Chemistry databases

There are at present more that 60 free chemistry databases that contain millions of chemical compounds, spectroscopic data, organic reaction mechanisms, protein structures, crystallographic data and scientific literature [Depth-First, Sixty-Four Free Chemistry Databaes, https://depth-first.com/articles/2011/10/12/sixty-four-free-chemistry-databases/].

ChemExper (https://www.chemexper.com/**).** Database that contains currently more than 10 millions chemicals, 16,000 Material Safety Data Sheets (MSDS), 10,000 IR spectra and more than 2,000 chemical suppliers.

PubChem (https://pubchem.ncbi.nlm.nih.gov/**).** PubChem is an open chemistry database at the National Institutes of Health (NIH, USA), U.S. National Library of Medicine, National Center for Biotechnology Information). PubChem is organized as three linked databases: PubChem Substance, PubChem Compound, and PubChem BioAssay

PubMed contains 103 million chemical compounds, 252 million substances, 268 million bioactivities and 31 million literature references.

PubChem is the world's largest collection of freely accessible chemical information. Scientists can search chemicals by name, molecular formula, structure, and other identifiers. It has chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations and more.

ChemSpider (http://www.chemspider.com/, it is owned by the Royal Society of Chemistry, UK). *ChemSpider* is a free chemical structure database providing fast text and structure search access to over 67 million structures from hundreds of data sources, properties and associated information. Structures (2D, 3D), properties, structural searches, etc. ChemSpider builds on the collected sources by adding additional properties, related information, and links back to original data sources. ChemSpider offers text and structure searching to find compounds of interest and provides unique services to improve this data by curation and annotation, and to integrate it with users' applications.

eMolecules [San Diego, Californis, USA, eMolecules Reagents and Building Blocks Search for Reaxys Users , https://www.emolecules.com/info/plus/download-database]
eMolecules empowers researchers to explore uncharted chemical and biological spaces and deliver more efficient drug-discovery programs. It contains 5.9 million commercially available, unique chemical structures from the eMolecules database.

Organic Syntheses is a peer-reviewed scientific journal that was established in 1921. It publishes detailed and checked procedures for the synthesis of organic compounds. A unique feature of the review process is that all of the data and experiments reported in an article must be successfully repeated in the laboratory of a member of the editorial board as a check for reproducibility prior to publication.The journal is published by Organic Syntheses, Inc., a non-profit corporation. An annual print version is published by John Wiley & Sons on behalf of Organic Syntheses, Inc. The journal from 1998 is published online.[http://www.orgsyn.org/].

Heterocycles (International Journal for Reviews and Communications, the oficial journal of the Japan Institute of Heterocycle Chemistry**)**
Since its launch in 1973, ***Heterocycles*** is an authoritative international journal that has provided a platform for the rapid exchange of research in the areas of organic, pharmaceutical, analytical, and medicinal chemistry of heterocyclic compounds. In addition to communications, papers and reviews, a special section of the journal presents newly-discovered natural products whose structure has recently been established. Another section is devoted to the total synthesis of previously documented natural products with heterocyclic ring systems. [https://www.heterocycles.jp/newlibrary/libraries/prepress].
Heterocycles Web Edition. This journal will list the new natural products with a heterocyclic ring system, collected from current chemical literature, whose structure has been established, new natural products with a heterocyclic ring system, collected from current chemical literature, whose structure has been established. [http://www.heterocycles.jp/ synthesis/synthesis_resultNew.php].

CAS (Chemical Abstracts Service) American Chemical Society (Columbus, Ohio, USA, founded 1907) is a global organization of expert scientists, technologists, and business leaders with a successful and extended history of delivering scientific information opportunities.

CAS®
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

CAS REGISTRY® contains more than 160 million unique organic and inorganic chemical substances, such as alloys, coordination compounds, minerals, mixtures, polymers and salts, and more than 68 million biosequences - more than any other database of its kind.
A CAS Registry Number® is universally recognized and used to provide a unique, unmistakable identifier for chemical substances.
Chemical Abstract Systems uses 8,000 journals, technical reports, dissertations, conference proceedings, and new books, available in at least 50 different languages, are monitored yearly, as are patent specifications from 27 countries and two international organizations. *Chemical Abstracts* ceased print publication on January 1, 2010 and is in digital form. **The CAS Registry** contains information on more than 130 million organic and inorganic substances, and more than 64 million protein and nucleic acid sequences.

WebReactions. Synthetic chemists think of a reaction by making and breaking of bonds at the reaction center as the defining nature of the reaction. Researtch chemists consider the effects of surrounding groups on rate, hindrance, or resistance to change under the reaction conditions. The WebReactions web database is a program that approach in similar ways for indexing reaction entries in any database."

Some chemical databases on the web (for free)

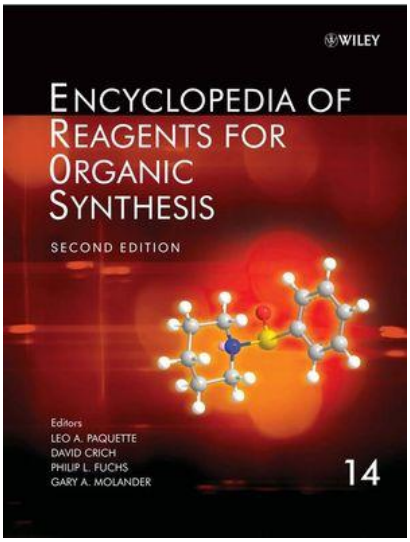
NIST (https://webbook.nist.gov/chemistry/ National Institute of Standards and Technology, U.S. Dpt of Commerce) Chemistry, cxhemical substances, webBook- Structures, spectra (NMR, IR, etc), properties.

PDB Protein Database (https://www.rcsb.org/pdb/home/sitemap.do) . This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease. The data, typically obtained by X-ray difraction crystallography, NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations.
The PDB protein database (4.7.2019) contains 142,433 proteins, 3,360 nucleic acids, and 153,601 protein/nucleic acid complecxes,

Research Collaboratory for Structurural Bioinformatics (RCSB)

ZINC Database (https://zinc.docking.org/ **).** Department of Pharmaceutical Chemistry, Universityu of California, San Francisco**.** A free database for virtual screening - Structures, properties, structural searches, etc. (renamed ZINC15 in 2015). A free database of commercially available chemical compounds, 230 million purchasable compounds and over 750 million purchasable compounds that can be searched for analogs in minutes.

e-EROS (Wiley, publisher) (https://library.medicine.yale.edu/find/title/eeros-encyclopedia-reagents-organic-synthesis). The online version of the critically acclaimed **Encyclopedia of Reagents for Organic Synthesis (EROS)** has rapidly become the definitive reference source for reagents used in organic synthesis. e-EROS allows chemical structure, substructure and reaction searching straight from a chemist's desktop. It contains a database of around 70,000 reactions and more than 4,500 of the most frequently consulted chemical reagents. It is searchable by structure and substructure, reagent, reaction type, experimental conditions, and allows for sophisticated full text searches.



Spectral Database for Organic Compounds SDBS

(https://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi)- NMR, IR, Raman, etc. spectra from a Japanese database.

SDBSWeb : https://sdbs.db.aist.go.jp (National Institute of Advanced Industrial Science and Technology, JAPAN).

SDBS, **National Institute of Materials and Chemical Research** in Japan, contains full spectra for 34,600 compounds, including about 25,000 mass spectra, 14,200 ^{13}C NMR, 15,900 proton ^{1}NMR, 54,100 IR, 3,500 Raman and 2,000 Electron Spin Resonance (ESR) spectra. The database is searchable by compound name, CAS Registry Number, molecular formula and NMR, IR or MS peaks.

MassBank- Mass Spectral database

(https://massbank.eu/MassBank/)

MassBank, the first public repository of mass spectral database of small chemical compounds (<3000 Da) for life sciences. Research groups contributing to the repository make their mass spectral data available to the public as supporting experimental data for other researchers. MassBank

accepts mass spectral data analyzed on chemical compounds using optimized, up-to-date analytical methods.

Crystallography Open Database (http://www.crystallography.net **)**. It is a collection of organic, inorganic and complex crystal structures. It is an open-access database with registered users. In 2016, the database had more than 360,000 entries.

Chemsynthesis (http://www.chemsynthesis.com/)
ChemSynthesis is a freely accessible database of chemicals. This website contains substances with their synthesis references and physical properties such as melting point, boiling point and density. There are currently more than 40,000 compounds and more than 45,000 synthesis references in the database.Database of chemicals with synthesis references that are not limited to a few journals. Along with these synthesis references our database also contains physical properties for the listed substances.

ChEBI. Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds."

DrugBank. [https://www.drugbank.ca/**]** A unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target information. The database contains 6,707 drug entries including 1,436 FDA-approved small molecule drugs, 134 FDA-approved biotech (protein/peptide) drugs, 83 nutraceuticals and 5,086 experimental drugs.

Electronic Encyclopedia of Reagents for Organic Synthesis. (Wiley online Library). This online edition is more than a Major Reference Work, as it combines the complete text of the Encyclopedia with the sophistication of a Database including all the chemical reactions and structures [https://onlinelibrary.wiley.com/doi/book/10.1002/047084289X]. "

REAXYS, a web-based tool for chemistry information

Reaxys is a web-based tool for the retrieval of chemistry information and data from published literature, including journals and patents. The information includes chemical compounds, chemical reactions, chemical properties, related bibliographic data, substance data with synthesis planning information, as well as experimental procedures from selected journals and patents. Reaxys is licensed by the well known scientific publisher Elsevier. Reaxys was launched in 2009 as the successor to the CrossFire databases.

Figure 4. Reaxy's covers basically 450 core chemistry journals and textbooks. Reaxys core literature is a carefully curated collection of essential titles in organic, inorganic and physical chemistry as well as material science, petro-chemistry, pharmacology and medicinal and computational chemistry. Reaxy's covers 16,000 chemistry-related periodicals, including conference abstracts to ensure comprenhensive literature searches.

Reaxys database gives access to over 1,000 periodicals containing 500 million .abstracts, and substanc**e** data from publicly available and proprietary databases. Reaxys is a web-based tool for the retrieval of chemistry information and data from published literature, including journals and patents. The information includes chemical compounds, chemical reactions, chemical properties, related bibliographic data, substance data with synthesis planning information, as well as experimental procedures from selected journals and patents. It is licensed by Elsevier. Reaxys was launched in 2009 as the successor to the CrossFire databases. It was developed to provide research chemists with access to current and historical, relevant,

organic, inorganic and organometallic chemistry information, from reliable sources via an easy-to-use interface.

Reaxys database provides research chemists with access to experimentally measured data: reactions, physical, chemical or pharmacological in one universal platform. Content covers organic, medicinal, synthetic, agro, fine, catalyst, inorganic and process chemistry and provides information on structures, reactions, and citations. Of 200 years of chemistry, abstracted from thousands of journal titles, books and patents.

Reaxys. [https://www.reaxys.com/#/search/quick**]** includes combined data of the Beilstein Database, the Gmelin Database, and the Patent Chemistry Database (Content updated 20.2.2020)

118	Million	Chemical Substances (structure, physicochemical properties)
49	Million	Chemical Reactions (mechanisms, reagents, etc)
59	Million……	Scientific Documents (on chemical reactions and synthesis)
37	Million…	Bioactivities of chemical compounds

CASTREACT database for chemical synthesis information

The **Chemical Abstracts Reaction Search Service** (**CASREACT**) is a chemical reaction database that contains chemical synthesis information derived from documents (1840 to the present). CASREACT is a structure-searchable, document-based database. The CA Abstract Number is the file accession number. In September 2016, the database contained over 91 million single-step and multistep reactions and synthetic preparations. The CASTREACT database is updated daily. The collection of information on chemical reactions has been accumulated with painstakiung and professional work of rthousands of sciwntists in the largest CASReact database [http://www.cas.org/content/reactions].and

[https://www.cas.org/support/documentation/reactions]

CASREACT®, is produced by CAS of the American Chemical Society, contains: (in 2020) approximatelly ~123 million single- and multi-step chemical reactions that are available with SciFinder^{n} (**SciFinder^{n}** is a research database of discovery application that provides integrated access to the world's most comprehensive and authoritative source of references) and on STN (**STN International**, the online service for sci-tech research and

patent information, offers a wide array of databases). Additional synthetic preparations available with SciFinder.

SYNARCHIVE. The Organic Synthesis Archive **(**https://www.synarchive.com/) SynArchive is a free web based application that allows you to browse a growing database of organic syntheses with the sequence of reactions in clear, precise and unambiguous form.

Chemical Synthesis **(**https://www.chemsynthesis.com/ **)** ChemSynthesis is a freely accessible database of chemicals. This website contains substances with their synthesis references and physical properties such as melting point, boiling point and density. There are currently more than 40,000 compounds and more than 45,000 synthesis references in the database.

Synthetic Reaction Update. (Royal Society of Chemistry, UK) (http://pubs.rsc.org/lus/methods-in-organic-synthesis)
Methods in Organic Synthesis was discontinued in 2014 and replaced by Synthetic Reaction Updates. The historical content of Methods in Organic Synthesis is available to Synthetic Reaction Updates subscribers. Each monthly issue of Methods in Organic Synthesis contained around 200 graphical abstracts selected from key journals in the field, covering all areas of synthetic chemistry including new reactions and reagents, asymmetric synthesis, and enzymatic transformations. The online database is searchable and indexed by reaction type, reactant, product, reagent, journal and author.

SciFinder Scholar. [https://www.cas.org/products/scifinder].
Database that explores reactions from the chemical literature with the interface to the following databases: Chemical Abstracts, CASReact, Registry, CHEMCATS, and Medline. Users search by using a chemical structure, chemical reaction, research topic, and author information.

Science of Synthesis. (Georg Thieme Chemistry) [https://science-of-synthesis.thieme.com/ **].** Full-text resource for methods in synthetic organic chemistry. Provides a critical review of the synthetic

methodology developed from the early 1800s to-date for the entire field of organic and organometallic chemistry. The new release of Science of Synthesis is now online – offering new content of two volumes.

Are artificial intelligence methods able to predict organic reactions?

Over the last 100 years that organic chemists experimented with organic reactions and the synthesis of organic chemical substances with complicated structures, there were many attempts to formulate methodologies to predict reactivity, mechanisms and yields. The advent of powerful computers, large databases of organic compounds and machine-learning techniques increased the capabilites of these methods. Although now there are attractive alternatives to predict reactivity of organic substances and mechanisms, such computer-aided reaction designs, are still in their infancy and prediction of reactions based on high-level quantum chemical methods is complex, even for simple molecules.[33,34,35]

Although methods of artificial intelligenece (AI) such as machine learning are powerful for data analysis (organic chemical structures, mechanisms, reactivity) its applications in organic synthetic chemistry are still being developed.[36,37,]

Reasearchers used information on 'dark' reactions--failed or unsuccessful hydrothermal syntheses--collected from archived laboratory notebooks from their laboratory, and added physicochemical property descriptions to the raw notebook information using cheminformatics techniques. In turn they used resulting data to train a machine-learning model to predict reaction success. When carrying out hydrothermal synthesis experiments using previously untested, commercially available organic building blocks, the machine-learning model outperformed traditional human strategies, and successfully predicted conditions for new organically templated inorganic product formation with a success rate of 89%. Inverting the machine-learning model reveals new hypotheses regarding the conditions for successful product formation.[38]

Advances in the last decade in AI have already formulated reaction systems which when controlled by machine learning algorithms may be able to explore the space of chemical reactions in very short time.[39,40,41]

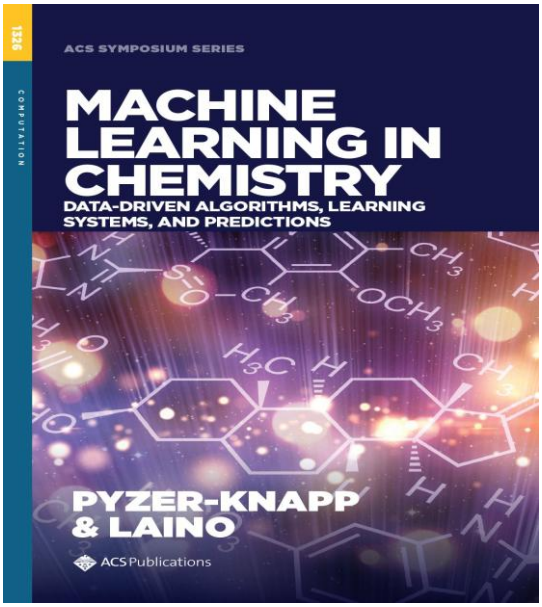
A recent paper (2018) proposed an organic synthesis robot that can perform chemical reactions and analysis at a faster rate, as well as predict the reactivity of possible reagent combinations after conducting a small number of experiments. By using machine learning for decision making, enabled by binary encoding of the chemical inputs, the reactions can be assessed in real time using nuclear magnetic resonance and infrared spectroscopy. The machine learning system was able to predict the reactivity of about 1,000 reaction combinations with accuracy greater than 80%.[42]

Organic synthetic chemists know from experience that to synthesize complex organic compounds is essential to know how to prepare functional compounds, including small-molecules which can be used for drug synthesis. Identification and development of synthetic routes remain until now a manual process and experimental synthesis platforms must be manually configured to suit the type of chemistry to be executed. The ideal automated synthesis platform must be capable to plan its own synthetic routes and execute them under conditions that facilitate scale-up to production goals. Chemists in the last decade have integrated computer-aided synthetic planning and robotically executed chemical organic synthesis. In a recent paper Colley et al., described the progress in automatically synthesis of organic chemicals with algorithmic prediction of viable routes to a target compound. The other aim of the report was to implement a known raction sequence on a computer platform (artificial intelligence) that needs little human intervention by intergration of two protocols. Researchers paired a retrosynthesis prediction algorithm with a robotically reconfigurable flow apparatus (solvent choice and precise stoichiometry supplied by researchers).[43]

Retrosynthetic analysis (working backwards experimental steps of organic synthesis) is the standard common technique used by chemists to plan the synthesis of small organic molecules. Given that transformations are formally reversed chemical reactions, the plan can be then carried out in the laboratory in the forward direction to synthesize the target compound. Sengler

et al., in a recent paper (2018) planned the syntheses of small organic molecules with the help of artificial intelligence (AI) techniques. It is known that computer-aided retrosynthesis would be a valuable tool but at present it is slow and provides results of unsatisfactory quality. Researchers used Monte Carlo tree search and symbolic artificial intelligence (AI) to discover retrosynthetic routes. They combined Monte Carlo tree search with an expansion policy network that guides the search, and a filter network to pre-select the most promising retrosynthetic steps. These deep neural networks were trained on essentially all reactions ever published in organic chemistry. The system of Segler et al., solved for almost twice as many molecules, thirty times faster than the traditional computer-aided search method, which is based on extracted rules and hand-designed heuristics. In a double-blind AB test, chemists on average considered these computer-generated routes to be equivalent to reported literature routes.[44]

The American Chemical Society (ACS) in 2019 organised a symposium on Artificial Intelligence and machine-learning techniques in predicting synthetic organic reactions. ACS considers that machine-learning in organic synthesis is considered a critical challenge in efficient organic synthesis routes. With the current rise of artificial intelligence (AI) algorithms, access to cheap computing power, and the wide availability of chemical data, it became possible by scientists to develop entirely data-driven mathematical models able to predict chemical reactivity. Similar to how a human chemist would learn chemical reactions, those learn by repeatedly looking at examples, the underlying patterns in the data. Researchers compared in a recent paper the state-of-the-art data-driven learning systems for forward chemical reaction prediction, analyzing the reaction representations, the data, and the model architectures. In their paper they discussed the advantages and limitations of the different AI model strategies and made comparisons on standard open-source benchmark datasets. The intention was to provide a critical assessment of the different data-driven approaches recently developed not only for the cheminformatics community, but also for the AI models end-users, the organic chemists, and for early adoption of such technologies.[45]



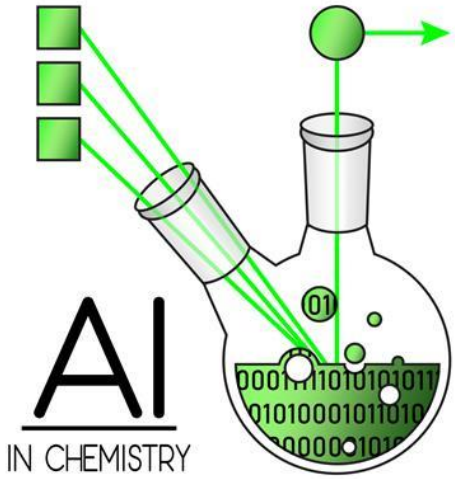


Figure 5. Pyzer-Knapp EO, Laino T (Eds)*. Machine Learning in Chemistry. Data Driven Algorithms Learning Systems and Predictions*. ACS Symposium Series Vol. 1326,Chapter 4:61-79, 2019. [DOI**:**10.1021/bk-2019-1326.ch004].

Synthetic organic chemistry by machine learning has been applied to wide variety of difficult challenges of organic synthetic planning. Artificial Intelligence (AI) and Deep learning is a type of machine learning that uses a hierarchical recombination of features to extract pertinent information and then learn the patterns represented in the data and these abilities have increasingly been applied to a wide variety of chemical challenges. A recent review (2019) collected a series of papers that explain the concepts of deep learning to chemists from any background and follows this with an overview of the diverse applications demonstrated in the literature, including organic synthesis.[46] A series of recent papers also explored the use of deep-learning for resolving challenging problems in synthetic organic chemistry.[47-50]

Conclusions

For over a century synthetic organic chemistry proviuded the basic scientific tools for the heavy chemical industry, drug discovery, chemical biology, materials science and chemical engineering. But the execution of complex chemical syntheses required expert knowledge of chemical reactivity of feedstock reagents, solvents, reaction conditions, catalysts, achieving high yields and separation techniques for purification. All this expertise was acquired over many years of study and hands-on laboratory practice by thousands of synthetic chemists working in industry, scientific institutions and university research laboratories. During this prolonged period synthetic organic chemists were interested to introduce computer-supported technologies with potential to resolve fundamental reactivity problems, streamline and automate chemical synthesis and achieve targeted prodution of complex structural chemicals. In the last decades renewed interest in artificial intelligence (AI), machine-learning and deep-learning techniques were driven by improved computing power, and data availability through the use of extended databases of chemicals, their properties, spectroscopic properties, reactivities and other useful characteristics. Organic reaction prediction remains one of the major challenges for organic chemistry and is a prerequisite for efficient synthetic planning of organic chemicals. In the last decade scientists developed AI neural networks and algorithms for predicting organic reactions. The scientific literature is full of studies with novel algorithms for predicting the products of organic chemistry reactions. Oirganic chemists are now using artificial intelligence (AI) techniques, in particular machine learning to identify the reaction type. They trained deep convolutional neural networks to predict the outcome of reactions based on example reactions and then they are applied on well known organic reactions. Considering the recent phenomenal progresses that have been made in computer programmes, artificial intelligenece methodologies, machine learning and deep learning methodologies, there is little doubt that these systems, once fully operational in organic chemistry, will dramatically speed up development of new complex organic chemicals, new materials and drugs.

References

1. Norman ROC, Coxon JM. *Principles of Organic Synthesis*. CRC Press, Taylor & Francis Group, Boca Raton, Fl, 1968, 1978, 1993 (3rd ed),
2. Zweifel GS, Nantz MH, Somfai P. *Modern Orghanic Synthesis*. Wiley and Sons, Hoboken, NJ, 2017.
3. Starkey LS. *Introduction to Strategies for Organic Synthesis*. Wiley and Sons, Hoboken, NJ, 2012.
4. Pirung M. *Handbook of Synthetic Organic Chemistry*, Elsevier, Amsterdam, 2017.
5. Nicolaou KC, Vourloumis D, Winssinger N, Baran PS. The art and science of total synthesis at the dawn of the twenty-first century. *Angewandte Chem Int Ed.* 39: 44–122, 2000.
6. Nicolaou KC, Sorensen EJ. *Classics in Total Synthesis: Targets, Strategies, Methods*. Wiley-VCH, New York, 1996.
7. Corey EJ, Cheng X-M. *The Logic of Chemical Synthesis*, Wiley, New York, 1995,
8. Fuhrhop J-H, Li G, Corey EJ. *Organic Synthesis: Concepts and Methods,* 3rd, Completely Revised and Enlarged Edition, Wiley-VCH, New Yorkm 2003.
9. Ahluwalia VK. *Strategies for Green Organic Synthesis*. CRC Press, Boca Raton, FL, 2012,
10. Taber DF. *Organic Synthesis: State of the Art, 2015-2017*. Wiley, Sheridan Books, New York, 2018.
11. Peiretti F, Brunel JM. Artificial Intelligence: The future for organic chemistry? *ACS Omega* 3(10): 13263-13266, 2018.
12. Gray NAB. Artificial intelligence in chemistry. *Analytica Chimica Acta* 210(1): 9-32, 1988.
13. Lindsay RK.; Buchanan BG, Feigenbaum EA.;Lederberg J. DENDRAL: A case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence 61*: 209–261, 1993,
14. Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses. *Science* 166:178192, 1969.
15. Whitesides GM. Reinventing Chemistry. *Angewandte Chemie Int Edition* 54(1):3196-3209, 2015.
16. Empel C, Koenigs RM, Artificial-Intelligence-Driven Organic Synthesis—En Route towards Autonomous Synthesis? *Angewandte Chemie Intern Edition* 58(48):17114-17116, 2019.
17. Maryasin B, Marquetand P, Maulide N. Machine learning for organic synthesis: are robots replacing chemists? *Angewandte Chemie Intern Edition 57(24): 6978-6980, 2018.*
18. Hie L, Fine NF, Nathel TK, Shah EL, et al. Conversion of amides to esters by the nickel-catalysed activation of amide C–N bonds, *Nature* 524:79–83, 2015.
19. Margrey KA, Mcmanus JB, Bonazzi S, et al. Predictive model for site-selective aryl and heteroaryl C–H functionalization via organic photoredox catalysis. J *Am Chem Soc.* 13 (32):11288–11299, 2017.

20. Kromann JC, Jensen JH, Kruszyk M, et al. Fast and accurate prediction of the regioselectivity of electrophilic aromatic substitution reactions. *Chem Sci* 9: 660–665, 2018.
21. Peng Q, Duarte F, Paton RS., Computing organic stereoselectivity–from concepts to quantitative calculations and predictions. *Chem Soc. Rev* 45: 6093–6107, 2016.
22. Sperger T, Sanhueza IA, Schoenebeck F. Computation and experiment: a powerful combination to understand and predict reactivities. *Acc. Chem. Res* 49:1311–1319, 2016.
23. Wei JN, Duvenaud D, Aspuru-Guzik A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* 2: 725–732, 2016.
24. Skoraczynski G, Dittwald P, Miasojedow B,[….] Gambin A. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?. *Scientific Reports* **7**, Article number 3582, 2017.
25. Zhou Z, Li X, Zare N. Optimizing chemical reactions with deep reinforcement learning. *ACS Cent. Sci* 3:1337–1344, 2017.
26. Coley CW, Barzilay R, Jaakkola RTS, Green WH, Jensen KF. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci* 3: 434–443, 2017.
27. Szymkuc S, Gajewska EP, Klucznik T, et al. Computer-assisted synthetic planning: the end of the beginning. *Angew Chem Int Edit* 56:5904-5937, 2016.
28. Coley CW, Green WH, Jensen KF. Machine lerning in computer-aided synthetic planning. *Acc Chem Res* 51:1281-1289, 2018.
29. Davies IW. The digitization of organic synthesis. *Nature* 570:175-181, 2019.
30. Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. Predicting reaction performance in C–N cross-coupling using machine learning. Science 360 (issue 6385):185-190, 2018.
31. Peiretti F, Brunel JM. Artificial Intelligence: The future for organic chemistry? *ACS Omega* 3 (10): 13263-13266, 2018.
32. De Almeida AF, Moreira R, Rodrigues T. Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry* 3:589-604, 2019.
33. Collins KD, Gensch T, Glorius F. Contemporary screening approaches to reaction discovery and development. *Nat. Chem* 6**:** 859–871, 2014.
34. Warr WA. A short review of chemical reaction database systems, computer-aided synthesis design, reaction prediction and synthetic feasibility. *Mol Inform* **33**: 469–476, 2014.
35. Plata RE, Singleton DA. A case study of the mechanism of alcohol-mediated Morita Baylis-Hillman reactions. The importance of experimental observations. *J Am Chem Soc* 137: 3811–3826, 2015.
36. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 521: 436–444, 2015.
37. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260, 2015.
38. Raccuglia P, Elbert KC, Adler PD, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* 533**:** 73–76, 2016.

39. Graulich N, Hopf H, Schreiner PR. Heuristic thinking makes a chemist smart. *Chem Soc Rev* 39: 1503–1512, 2010.
40. Gil Y, Greaves M, Hendler J, Hirsh H. Amplify scientific discovery with artificial intelligence. *Science* 346:171–172, 2014.
41. Trobe M, Burke MD. The molecular industrial revolution: automated synthesis of small molecules. *Angew Chem Int Ed* **57**:4192-4214, 2018.
42. Granda JM, Donina L, Dragone V, Long D-L, Cronin L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* 559:377-381, 2018.
43. Coley CW, Thomas II DA, Lummiss JAM, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* 365(No. 64653):565-570, 2019.
44. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555: 604-610, 2018.
45. Schwaller P, Laino T. Data-Driven Learning Systems for Chemical Reaction Prediction: An Analysis of Recent Approaches. November 2019. In *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*. American Chemnical Society Symposium Series Vol. 1326, Cahpter 4: 61-79, 2019.[DOI**:**10.1021/bk-2019-1326.ch004].
46. Matel AC, Coote ML. Deep learning in chemistry. *J Chem Inform Model* 59(6):2545-2559, 2019. [https://doi.org/10.1021/acs.jcim.9b00266].
47. Foscato M, Jensen VR. Automated in silico design of homogeneous catalysts. *ACS Catalysis 10*(3): 2354-2377, 2020.
48. Yan W, Fidelis TT, Sun W-H. Machine Learning in Catalysis, From Proposal to Practicing. *ACS Omega 5*(1): 83-88, 2020.
49. Cova TFG, Pais ACC. Deep learning for deep chemistry: optimizing the prediction of chemical patterns. *Frontiers in Chemistry* **2019,** *7*.
50. Lo Y-C, Ren G, Honda H, Davis KL. Artificial Intelligence-Based Drug Design and Discovery. *ChemInformatics and Its Applications. Drug Discovery Today*, 2019, DOI: 10.5772/intechopen.89012.