
“A JOURNEY FROM ALCHEMIST TO AI CHEMIST: A STUDY”

Dr. Pratik E. P. Michael

Department of Chemistry,

Hislop College, Nagpur-440001

Email Id : pmichael28@gmail.com

Mob. No. : 91-9860104874

Crossref DOI - <https://doi.org/10.63665/rh.v7i2.74>

Abstract :

The objective of technological advancement has been to compliment chemists from heavy labour, assist human efforts to predict, design and execute a process to achieve the desired results and the increased accessibility to communication. It has always been believed that humans possessed the knowledge of providing answers in complicated situations and making decisions. However, the scenario seems to be changing fast with the advent of Artificial Intelligence (AI) which is becoming increasingly important in the field of chemistry.

Despite numerous applications of AI, it's still a relatively new concept that has a number of disadvantages and challenges that can affect the process if not researched enough and done right. In the present article, we will provide a general picture of how AI can help chemists be faster and more creative in their research.

Keywords : Artificial Intelligence (AI), AI chemist, Researchers, Computational chemistry

Introduction :

AI is a developing area and focuses on utilising computer systems to solve issues by executing algorithms that emulates the cognitive functions of the human brain. Most of the present day AI algorithms can establish connections between inputs and outputs, adjust their behaviour based on environmental indications and subsequently make decisions, enhancing the likelihood of delivering precise responses.

Artificial intelligence has numerous applications, including natural language processing, reasoning, and strategic decision-making. AI can modify objects based on specific requirements. Artificial intelligence is not limited to engineering but also has many applications in the chemical field too.

Chemistry has experienced a significant increase in data, which coincided with the advent of powerful computer technology. Chemists have primarily gained their knowledge by doing experiments and thus gather data. The various data acquired is then analysed by the chemists, fostering their understanding of chemistry. Chemistry have from the very beginning



derived knowledge from the available data. Chemists have run experiments to obtain data, may it be chemical or physical properties, on chemical reactions, or on biological activities. These data are then used to make predictions or to derive models for the principles that underly the data. The use of active learning as well as the integration of various models with AI are becoming common, allowing for more effective use of data to new systems. Rolf Huisgen has written a chapter “Mesomerie-Lehre” for a textbook on laboratory experiments benefiting the understanding of chemistry in students [1].

Researchers are transforming from awareness to usage of AI for the benefit of the society. AI has revolutionaried drug discovery [2, 3], the design of smart materials [4-6], drug safety [7] and synthesis of organic compounds [8]. The most fascinating techniques in ML today is large language models (LLMs) Brown, 2020 [9], Achiam et al. in 2023 [10], Touvron et al. [11], Zhao et al. [12], and meanwhile LLMs for chemistry is gradually gaining attention.

Significant advancements in technology and research over the years have led to substantial improvements in chemical datasets, providing a great support for machine learning applications in this field. Developed at Stanford University in the mid-1960s, “DENDRAL” was one of the earliest and most influential chemical expert systems Buchanan et al. in 1969, [13]. By encoding the knowledge of chemists into a series of rules and instructions, the system was successfully applied to identify molecular structures from mass spectrometry data. Several other expert systems emerged for predicting the reaction outcomes Salatin and Jorgensen [14], Funatsu and Sasaki in 1988 [15], Satoh and Funatsu, 1995 [16], but all relied on predefined rules and were limited in adapting to new or complex scenarios.

Machine learning algorithms can be utilised with less intricate data and can potentially help overcome challenges in analytical chemistry [17, 18-21].

The setup of artificial intelligence (AI) is changing rapidly and in order to ascertain that technology has a positive impact on research and other domains, it's important to monitor the views of those who are already using it.

The present review covers the evolution and impact of AI in chemistry from early computational methods to machine learning and deep learning developments; key milestones in chemical research and its applications with AI are discussed in areas such as prediction of reactions and properties, CASE, CASD, drug discover and materials science. The review also covers the scope of analysis on both the theoretical basis and practical implementation of AI in chemistry for researchers and practitioners in chemistry possible process that can be adapted and the underlining urgency to address associated challenges.

Purpose and Scope of the Review :

Even though AI has numerous applications in the area of technology, it also has many applications in the field of chemical science [22]. Searching the features of molecules included in existing databases, Artificial Learning can find combinations that may be promising as drugs. ML techniques are able to transform the search for novel medications as they can operate faster and are cheaper than people [23]. The analysis of sequence of data, it's



experience and learning constitute machine learning. The sequence and reasoning constitutes AI [24], meaning that machine learning serves AI to achieve the tasks.

A. Learnings in Chemistry :

AI is becoming a focus in chemistry, effecting greater probabilities to innovations and strengthening research capabilities. Key AI methods which are primarily used in chemistry are Machine Learning (ML)/Data Learning (DL), and Natural Language Processing (NLP).

1. Machine Learning :

a. Supervised Learning :

Supervised learning is one of the most commonly adopted AI technique in chemistry, where a model is trained against a labeled dataset in which input-output pairs are already known. It is very common approach in regression and also classification of tasks. Regression, on the other hand, involves predicting an output that is continuous based on the input features. These models of supervised learning are able to predict those properties of molecules such as solubility, boiling points, and toxicity. Classic examples include QSAR models, which, establishes a correlation between chemical structure and biological activity.

b. Unsupervised Learning :

Unsupervised learning is different from supervised learning and deals with data without labeled outcomes. The key task is to discover a pattern or structure existing within the data. There are two main approaches having useful application in chemistry under unsupervised learning are clustering and dimensionality reduction. Some other people have employed k-means and hierarchical clustering algorithms in grouping similar compounds together by properties and are useful in the design of a compound library and also in optimisation during lead identification in drug discovery. Commonly, PCA and t-SNE techniques are applied; these use up the high- dimensional chemical data space of its associated complexity thereby making visualisation and interpretation easier. Chemical activity can be determined with molecular descriptors and identification of important features by using such analysis.

c. Reinforcement Learning :

Reinforcement learning (RL) in AI is an approach where one agent interacts with the environment to make a decision on the actions. Whether it will receive rewards or penalties will be based on the actions taken. Lately, there is an increasing interest in the exploration of RL in chemistry like molecular design and reaction optimisation. To the problem where and agents goal is to find an optimal sequence that could be able to synthesise a given compound viz. retrosynthesis, RL algorithms have been applied. Example, DQN can be considered, which was trained by proposing reaction pathways and exploring big chemical spaces.

2. Natural Language Processing (NLP) :

One can use NLP processing methodology (AI methods) on human language to



process and analyze it. For instance, natural language processing has been traditionally used in chemistry to mine and extract knowledge from the scientific literature available in plenty. To automatically identify chemical entities, reactions, and properties, Specifically text mining tools namely entity recognition and relation extraction can be applied.

B. Data base :

Diverse datasets availability determines the success of AI models in computational chemistry. These datasets allow AI algorithms, particularly machine learning and deep learning models, to identify patterns within chemical systems, enabling them to make predictions, simulate reactions, and discover new materials. The datasets used are:

1. Quantum Chemistry Datasets :

Quantum chemistry datasets are crucial for modelling molecular properties based on quantum mechanics. These datasets contain atomic configurations, molecular geometries, and electronic structures, which AI models use to predict energy states, electronic properties, and molecular reactivity. Examples such as QM7 and QM9, provide quantum mechanical properties for small organic molecules, this also includes atomisation energies and polarizabilities, which help AI models in simulating molecular properties without resorting to computationally intensive quantum mechanical calculations [25].

a. Chemical Reaction Datasets :

Chemical reaction datasets comprises of experimental and simulated reaction data, viz. reaction pathways, transition states, and kinetics. The AI models use this data to predict reaction outcomes or optimize reaction conditions. The widely used datasets for reaction prediction tasks are USPTO Reaction Dataset which contains millions of chemical reactions from U.S. patent filings.

b. Materials Property Datasets :

Critical information in materials chemistry about material compositions, crystal structures, and properties such as band gaps, elastic moduli, and thermal conductivity is provided by datasets. These datasets are used for designing new materials and optimizing existing ones. Important examples involves the Materials Project Database, which provides data on thousands of inorganic compounds and their computed properties like formation energies and band structures.

c. Molecular Dynamics Datasets (MD) :

MD datasets are essential for studying time-dependent behaviours of molecules. These datasets provide information regarding molecular motion, interactions and forces over time, enabling AI models to simulate dynamic molecular behaviour and predict future states. An important example is the MD17 dataset as it contains molecular trajectories of small molecules, generated from ab initio molecular dynamics simulations and is valuable for training AI models that aim to predict molecular interactions and forces in real-time applications.



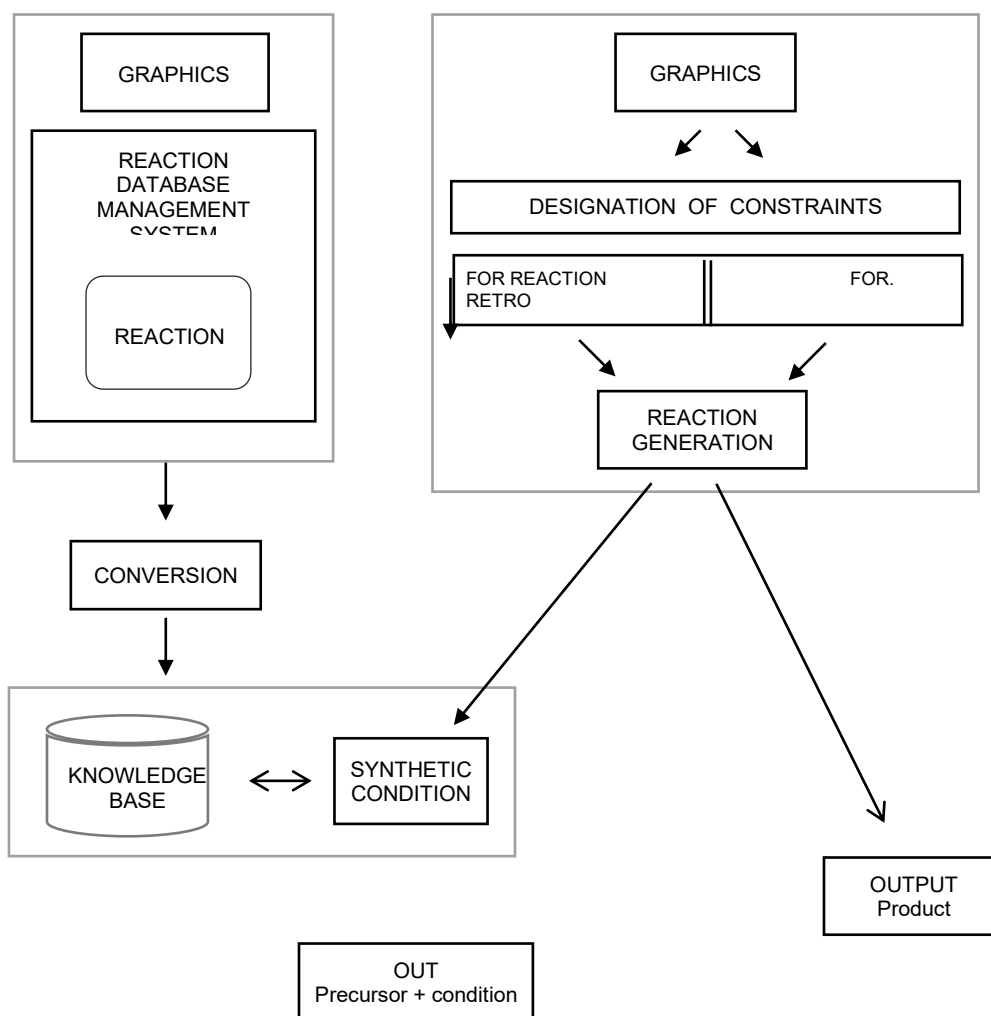
d. Toxicity and Bioactivity Datasets :

It is imperative that AI models focus on drug discovery, datasets containing bioactivity, toxicity, and pharmacological properties of small molecules. These datasets aid predicting compound's interaction with biological targets or its potential toxicity. ChEMBL Database an example of this type of datasets, which contains bioactivity data for thousands of drug-like molecules, including their interactions with various biological targets.

C. Prediction of Reactions :

One of the system present here has been named "AIPHOS" (Artificial Intelligence for Planning and Handling Organic Synthesis).

AIPHOS is a system designed to help the chemist in planning a synthetic route, or predicting the most realistic reaction path for a chemical compound. A coarse division of



AIPHOS, illustrated in Fig. 1.

Fig. 1 Block Diagram of AIPHOS

The diagram shows three parts of the system :



1. Strategy is the central algorithm of the system which is oriented to the reaction generation process in the system.
2. A data base of individual and specific reactions and its management system.
3. The two former functions are brought together, generating automatically a knowledge base from the data base of specific reactions of part 2.

Recent advances in the application of machine learning to synthetic chemistry, divided in three categories Fig. 2: [26]

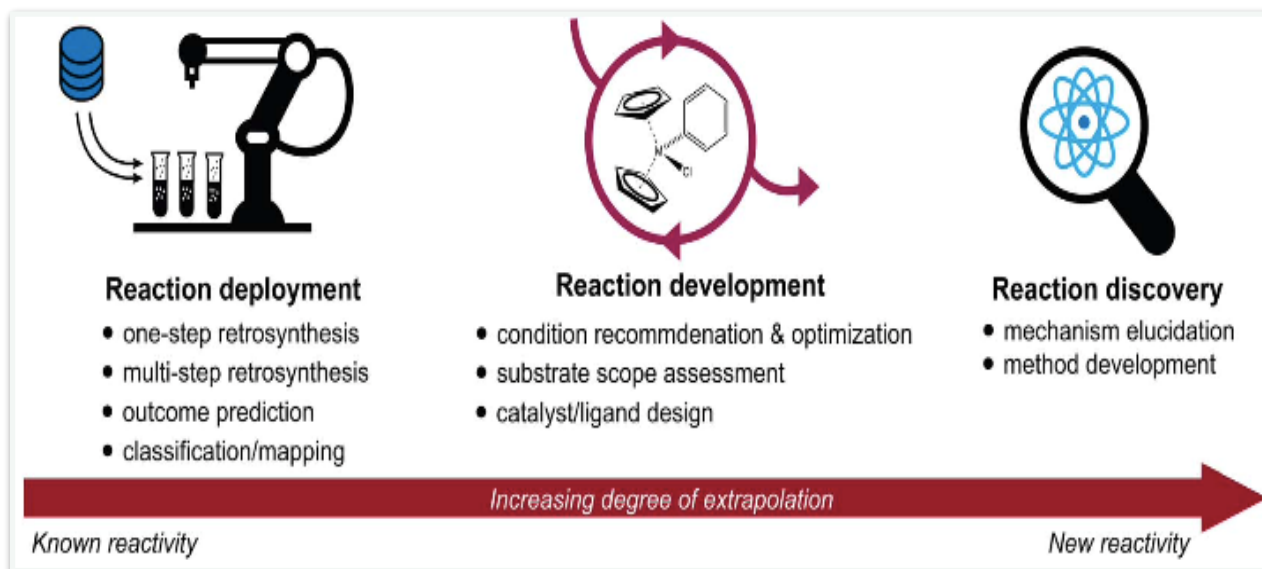


Fig. 2 Overview of the three main categories of predictive chemistry tasks discussed throughout this review: reaction deployment, development, and discovery. It is useful to consider the extent to which each task represents an extrapolation from known reactivity to new reactivity.

- (1) Reaction deployment : learning from reaction corpora to identify trends and predict when known reactions apply to novel substrates or combinations thereof.
- (2) Reaction development : accelerating the improvement or optimisation of an existing chemical process, often in an interactive setting incorporating experimental feedback.
- (3) Reaction discovery : creating new knowledge through the elucidation of reaction mechanisms or the discovery of unprecedented synthetic methods.

D. Computer Assisted Structure Elucidation (CASE) :

NMR spectroscopy the most widely used analytical method for structure elucidation and identification of organic chemical compounds. The process is to lookup in to the database for an already known compound (dereplication) or for structure fragments has been included in several existing CASE systems using the spectral fingerprint of one-dimensional (1D) NMR experiments. The collection of NMR data used for computer-assisted structure elucidation (CASE) may contain 1D spectra such as ^1H , ^{13}C and DEPT as well as the 2D HSQC, HMBC and COSY spectra. Together with a given molecular formula (MF) mainly determined by mass spectrometry (MS), the substructures and restrictions derived from 1D and 2D NMR spectra form the cornerstone of structure generation by CASE systems. These

constraints and the ones resulting from fragment search shrink the chemical search space that would be otherwise too wide for practical applications [27-34]. The Computer assisted structure elucidation (CASE) software is readily available for all users, bringing a new pedagogical perspective to the next generation of chemists [35]. With CASE, user are able to identify and learn from their logical errors, promoting autonomous learning in the process.

E. Reaction Prediction and Computer Assisted Synthesis Design (CASD) :

Organic reactions present a great challenge for computer program as they are not like the process of chess or Sudoku games, because they are full of exceptions and rarely have fixed rules. With introduction of artificial intelligence (AI), scientists realised if AI and synthetic planning are combined they would probably provide general trend in this field. One cannot guarantee the correctness of computer-designed synthetic routes, AI may probably come up with incredible new ideas and its comprehension of complex reaction patterns.

There is a need for such an application to be made apparent by the fact that a complete, logic centered synthetic analysis of a complex organic structure may require much time for even a skilled chemist.

Reaction databases are in widespread use in the organic chemistry and the quality of these systems depends not only on the chemical information contained within the databases but also on the retrieval system supplied with the databases. It is of vital importance to create user-friendly systems which allow for high-level queries and AI techniques have to be effectively employed.

A typical search in a reaction database using standard systems now available and described by :

Standard Existing Technique for Reaction Search :

The search for a set of suitable retro-synthetic reactions serving as suggestions to be actually tried in the laboratory, normally consists of various stages of interaction with a database system. The first stage is to confirm whether the target is a product in any of the reactions. If not, then the target must be retro-synthetically analysed for important groups and reactive centers. By trial and error process one can find a suitable substructure that yields a reasonable number of suggestions. If too many "hits" are found, then the substructure must be expanded to include more of the target so that less reactions will be included. If no hits are found, then groups of atoms must be eliminated from the substructure to make the search more general.

Retro-synthetic Search Using RETROSYN :

A system developed by RISC-Linz, RETROSYN for reaction database search is a complete CAOS (computer- aided organic synthesis) systems. The RETROSYN module is part of an expert chemical system, written in LISP, which is used to extract, calculate, and organise chemical information (with emphasis on organic synthesis) from the raw information contained in chemical databases⁹ The system is now in usable prototype form.



F. Drug Delivery :

Drug discovery, an highly intricate process requires the identification of potential drug candidates that can effectively treat various diseases. The use of AI has brought a significant shift in the approach to drug discovery and has fundamentally transformed the pharmaceutical industry by speeding up the drug discovery process, improving precision, and decreasing costs.

Drug Delivery System (DDS) may be defined as a system comprising of drug formulation. DDS is a medical device or dosage form/technology to carry the drug inside the body mechanism for its release. Conventional drug delivery involves the formulation of the drug into a suitable form (compressed tablet for oral administration or a solution for intravenous administration). Some new drug delivery systems have been developed or are being developed to meet the limitation of the conventional drug delivery systems.

G. Material Science :

The rapid growth of AI technologies and the increased demands on materials databases have shifted significantly, evolving from simple data hosting platforms to more data-driven systems. It has translated into a new approach to database design and management, as traditional methods no longer meet the needs of heterogeneous and multimodal materials data.

The user-defined data are being developed at a rapid pace that allows for user-defined data descriptions, gradually becoming the foundation for the open interconnection and seamless sharing of multiple data resources and applications.

The Materials Data Curation System (MDCS) [36], developed by National Institute of Standards and Technology (NIST) in the U.S., offers reusable general base types, including double-precision floating-point types, unit types, and uncertainty types. The users is allowed to combine these base types to create new complex data types, forming customised materials data templates for data uploads.

Conclusion :

AI has truly transformed how research in chemistry is conducted and applied from early computational methods to sophisticated machine learning and deep learning models. AI techniques are capable of handling Big data chemical data for the prediction of molecular properties with high accuracy. Machine learning, both supervised and unsupervised, and even reinforcement learning, has made chemical models increasingly predictive. Applications so far in drug discovery, materials science, chemical synthesis, and spectroscopy illustrate the potential of AI to bring change in these fields. The impact of AI on the chemical industry and academia is unprecedented. Some of the characteristics include increased efficiency, cost reduction, shifting research methodologies, and educational practice.

Despite the potential of AI in computational chemistry, including its application in materials chemistry, several challenges and limitations still exist. These challenges



encompass the complexity of quantum systems, the interpretability of AI models, and the necessity for high-quality data. Understanding and predicting the behaviour of quantum systems are essential in materials chemistry.

Hence, one can confidently say that the impact of AI on both the chemical industry and academia is nothing short of revolutionary. In the near future, AI will continue to rapidly improve in chemistry.

References :

- Huisgen R. in *Die Praxis des organischen Chemikers*, L. Gattermann, H. Wieland, Th. Wieland, Walter de Gruyter & Co, Berlin, 1959, pp. 377-395.
- Richardson A, Signor BM, Lidbury BA, Badrick T. Clinical chemistry in higher dimensions: machine-learning and enhanced prediction from routine clinical chemistry data. *Clin Biochem.* 2016 Nov 1;49(16-17):1213-20. doi: 10.1016/j.clinbiochem.2016.07.013, PMID 27452181.
- Kalayil NV, D'Souza SS, Khan SY, Paul P. AI in pharmacy drug design. *Artif Intell.* 2022;15(4).
- Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for detecting malaria. *Transl Res.* 2018 Apr 1;194:36-55. doi: 10.1016/j.trsl.2017.12.004, PMID 29360430.
- Nayak J, Vakula K, Dinesh P, Naik B, Pelusi D. Intelligent food processing: journey from artificial neural network to deep learning. *Comput Sci Rev.* 2020 Nov 1;38:100297. doi: 10.1016/j.cosrev.2020.100297.
- Engkvist O, Norrby PO, Selmi N, Lam YH, Peng Z, Sherer EC. Computational prediction of chemical reactions: current status and outlook. *Drug Discov Today.* 2018 Jun 1;23(6):1203-18. doi: 10.1016/j.drudis.2018.02.014, PMID 29510217.
- Sujith T, Chakradhar T, Marpaka S, Sowmini K. Aspects of utilization and limitations of AI in drug safety. *Asian J Pharm Clin Res.* 2021;14(8):34-9.
- Panteleev J, Gao H, Jia L. Recent applications of machine learning in medicinal chemistry. *Bioorg Med Chem Lett.* 2018 Sep 15;28(17):2807-15. doi: 10.1016/j.bmcl.2018.06.046, PMID 30122222.
- Brown Tom B, Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.
- Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M. Anne, Lacroix T., Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.
- Zhao H., Zhao Q., Zhao X., Fen-Fang Zhou, Zhou Z., Zhuo J., Yi-Ling Zou, Qiu X., Yu Qiao, and Dahua Lin. Internl. technical report. ArXiv, abs/2403.17297, 2024.
- Buchanan B., Sutherland G., and Feigenbaum E. A., Heuristic dendral: A program for generating explanatory hypotheses. *Organic Chemistry*, page 30, 1969.



- Salatin T. D. and Jorgensen W. L., Computer-assisted mechanistic evaluation of organic reactions. 1. overview. *The Journal of Organic Chemistry*, 45(11):2043–2051, 1980.
- Funatsu K. and Shin-Ichi Sasaki, Computer-assisted organic synthesis design and reaction prediction system, “aiphos”. *Tetrahedron Computer Methodology*, 1(1):27–37, 1988.
- Satoh H. and Funatsu K., A knowledge base-guided reaction prediction system- utilization of a knowledge base derived from a reaction database. *Journal of chemical information and computer sciences*, 35(1):34–44, 1995.
- Richardson A, Signor BM, Lidbury BA, Badrick T. Clinical chemistry in higher dimensions: machine-learning and enhanced prediction from routine clinical chemistry data. *Clin Biochem.* 2016 Nov 1;49(16-17):1213-20. doi: 10.1016/j.clinbiochem.2016.07.013, PMID 27452181.
- Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for detecting malaria. *Transl Res.* 2018 Apr 1;194:36-55. doi: 10.1016/j.trsl.2017.12.004, PMID 29360430.
- Nayak J, Vakula K, Dinesh P, Naik B, Pelusi D. Intelligent food processing: journey from artificial neural network to deep learning. *Comput Sci Rev.* 2020 Nov 1;38:100297. doi: 10.1016/j.cosrev.2020.100297.
- Engkvist O, Norrby PO, Selmi N, Lam YH, Peng Z, Sherer EC. Computational prediction of chemical reactions: current status and outlook. *Drug Discov Today.* 2018 Jun 1;23(6):1203-18. doi: 10.1016/j.drudis.2018.02.014, PMID 29510217.
- Panteleev J, Gao H, Jia L. Recent applications of machine learning in medicinal chemistry. *Bioorg Med Chem Lett.* 2018 Sep 15;28(17):2807-15. doi: 10.1016/j.bmcl.2018.06.046, PMID 30122222.
- Sokolova B. Biotech Startups, *BiopharmaTrend.com*, <https://www.biopharmatrend.com/contributor/734>. (accessed August 22, 2022).
- Rawat S, 5 AI Applications in Chemistry, May 24, 2021 <https://www.analyticssteps.com/blogs/5-ai-applications-chemistry> (accessed 19 August, 2022).
- Bajwa A. Artificial Intelligence vs Robotics vs Machine Learning vs Deep Learning vs Data Science, *DataDrivenInvestor*, 2021. <https://medium.datadriveninvestor.com> (accessed July 18, 2023).
- Dral P. O. (2024) AI in computational chemistry through the lens of a decade-long journey. *Chem Commun* 60(24):3240–3258.
- 26. Zhengkai Tu,†^a Thijs Stuyver†^b and Connor W. Coley, *Chem. Sci.*, 2023, 14, 226.
- Elyashberg M., Identification and structure elucidation by NMR spectroscopy. *TrAC Trends Anal. Chem.* **2015**, 69, 88–97. [CrossRef]
- Elyashberg, M.; Argyropoulos, D. Computer Assisted Structure Elucidation (CASE): Current and future perspectives. *Magn. Reson. Chem.* **2021**, 59, 669–690. [CrossRef]
- Steinbeck, C. LUCY—A Program for Structure Elucidation from NMR Correlation Experiments. *Angew. Chem. Int. Ed. Engl.* **1996**, 35, 1984–1986. [CrossRef]
- Burns, D.C.; Mazzola, E.P.; Reynolds, W.F. The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products.



Nat. Prod. Rep. **2019**, *36*, 919–933. [CrossRef].

- Elyashberg, M.; Williams, A. ACD/Structure Elucidator: 20 Years in the History of Development. *Molecules* **2021**, *26*, 6623. [CrossRef].
- Steinbeck, C. Recent developments in automated structure elucidation of natural products. *Nat. Prod. Rep.* **2004**, *21*, 512–518. [CrossRef].
- Steinbeck, C. Computer-Assisted Structure Elucidation. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Volume 3, pp. 1378–1406. [CrossRef].
- Elyashberg, M.E.; Williams, A.; Blinov, K. *Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation*; Royal Society of Chemistry: London, UK, 2015; Available online: <https://play.google.com/store/books/details?id=fmsoDwAAQBAJ> (accessed on 30 September 2022).
- Moser, A.; Pautler, B. G. The Fundamentals behind Solving for Unknown Molecular Structures Using Computer-Assisted Structure Elucidation: A Free Software Package at the Undergraduate and Graduate Levels. *Magn. Reson. Chem.* 2016, *54* (9), 701–704.
- Materials Data Curation System, (n.d.). <https://github.com/usnistgov/MDCS>.

