Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

From Platform to Knowledge Graph: Evolution of Laboratory Automation

Jiaru Bai[1,†], Liwei Cao[1,†], Sebastian Mosbach[1,2], Jethro Akroyd[1,2], Alexei A. Lapkin[1,2], Markus Kraft[1,2,3,4]

released: November 10, 2021

[1] Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

[2] CARES
Cambridge Centre for Advanced Research and Education in Singapore
1 Create Way
CREATE Tower, #05-05
Singapore, 138602

[3] School of Chemical and Biomedical Engineering
Nanyang Technological University
62 Nanyang Drive
Singapore, 637459

[4] The Alan Turing Institute
London, NW1 2DB
United Kingdom

[†] J.B. and L.C. contributed equally to this work.

Preprint No. 284

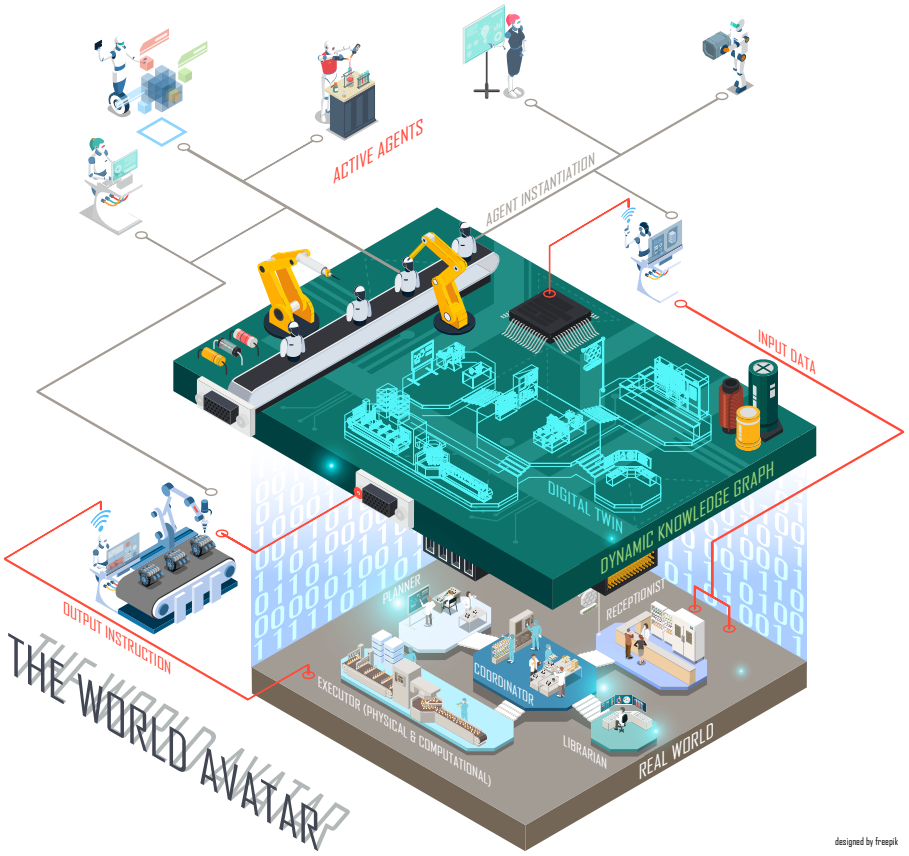
Keywords: Knowledge-graph technology, digital twin, chemistry digitalisation, closed-loop optimisation, laboratory automation

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

E-Mail: mk306@cam.ac.uk
World Wide Web: https://como.ceb.cam.ac.uk/

Abstract

High-fidelity computer-aided experimentation is becoming more accessible with the development of computing power and artificial intelligence tools. The advancement of experimental hardware also empowers researchers to reach a level of accuracy that was not possible in the past. Marching towards the next generation of self-driving laboratories, the orchestration of both resources lies at the focal point of autonomous discovery in chemical science. To achieve such a goal, algorithmically-accessible data representations and standardised communication protocols are indispensable. In this perspective, we recategorise the recently introduced approach based on Materials Acceleration Platforms into five functional components and discuss recent case studies that focus on the data representation and exchange scheme between different components. Emerging technologies for interoperable data representation and multi-agent systems are also discussed with their recent applications in chemical automation. We hypothesise that knowledge graph technology, orchestrating semantic web technologies and multi-agent systems will be the driving force to bring data to knowledge, evolving our way of automating laboratory.



Highlights

- Reviewed data flow within state-of-the-art chemical automation studies.
- Summarised current data representation and exchange protocols.
- Proposed a dynamic knowledge-graph-based approach towards automated closed-loop optimisation.

Contents

1 Introduction

The automation of laboratory involves linking the abstract concepts of chemical processes and the hardware responsible for the execution [55, 155]. It can be achieved by creating a fully connected virtual representation of the physical equipment and their status, *i.e.*, a 'digital twin' of the laboratory that bridges the gap between the virtual and the real world. By doing so, it enables the orchestration of physical and computational experimentation in cyberspace, facilitating the automation of chemical discovery [140]. Therefore, it shortens the time span from making a new chemical in the research environment to the delivery of its mass production to the end-users. This presents the opportunity to deliver a significant level of decarbonisation with reduced labour and energy consumption, making the digitalisation of chemical manufacturing one of the critical technology paths towards a more sustainable society [12, 65].

The first automated hardware for chemistry dates back to the late 1960s [95]. Since then, considerable advances have been made to expand the potentialities of such a tool, covering the field of chemical reactions [24, 25], drug discovery [130], and material discovery for clean energy [139]. As chemists' quest to achieve a universal organic compound synthesis machine, three key capabilities were identified [115], *i.e.*, access to database of chemical reaction knowledge, synthetic steps planning, and automated execution of proposed action sequence. For a detailed historical excursus, the readers refer to Dimitrov et al. [29]. In 2020, Flores-Leonar et al. [40] proposed materials acceleration platforms (MAP), a platform-based approach, as the paradigm to accelerate the material discovery process. In line with the three key capabilities that seem to be required to build a robochemist [115], Flores-Leonar et al. [40] envisaged integration of machine learning (ML) algorithms and robotics platforms, with further interfacing between humans and robots, is the way towards autonomous experimentation. The current practices of development towards laboratory automation is seen following this trend. Researchers adopt automation of chemical experiments and advances in ML to enable functional material discovery [83, 88], the discovery of chemical reactions [92], synthesis planning [23, 138], and optimisation of process conditions [6, 11, 38]. Despite the great success demonstrated by the community, the effort required to incorporate new equipment into an existing platform can be expensive. Tailored extraction-transformation-loading (ETL) tools and the specific data exchange scheme for establishing effective communication are to be developed for each piece of equipment added. Therefore, these platforms normally face difficulties in scalability and interoperability due to heterogeneous data formats as an obstacle to holistic integration. As a prerequisite condition towards digitalisation, the absence of standardised data representation and exchange protocols is seen as one of the critical challenges faced by the community [25].

A way forward may be offered by Semantic Web technologies [7]. It represents the vision of a fully linked web of data, demonstrating interoperability across scales and domains. It uses ontologies to describe the concepts and relationships within a given domain for communal understandings. An ontology normally consists of two components: a terminological box (TBox) and an assertional box (ABox) [151]. TBox refers to the description at a conceptual level, while ABox stores the data that is a realisation of the concepts defined by the TBox. Both levels can be accessed via internationalised resource identifiers

(IRIs) for unambiguous identification. In the context of experiment automation, this opens up the possibility of developing a fully linked data representation for the chemical processes and equipment status as a universal framework to facilitate concrete data exchange within and between platforms.

Besides the interoperable data representation, an effective way to communicate and share data must be addressed to achieve laboratory automation. In this regard, collective intelligent agents have been used to automate the tasks involved in crystal-structure phase mapping [50], material discovery [98], and reaction optimisation [14]. Considering the historical discussions of integrating the two technologies [60], we hypothesise that an ontological representation of a laboratory, linked with different data standards, would enable the rapid implementation of artificial intelligence (AI) tools for chemical discovery and development.

This perspective focuses on the current progress of data representation and data exchange towards the next generation of self-driving laboratories – chemical 'digital twins'. It does so by identifying the functional components required by such laboratories and then mapping the state-of-the-art studies into the classifications. The data representation and exchange between the components are thus reviewed and assessed. We review efforts in the community towards better data representations and exchanges. The current landscape can be divided into two parts, semantic and non-semantic, depending on whether chemistry ontologies are involved. We map out the links between different initiatives and find an emerging trend for using semantic representations of chemical knowledge to facilitate the development of reaction informatics. We review the realisation of some functional components in agents demonstrated by some studies. The possibility of combining the two technologies is discussed by introducing 'The World Avatar' project.

The presentation of this perspective is structured as follows: section 2 reviews the selected state-of-the-art studies of chemical automation; section 3 discusses community efforts towards standardised data representation and effective data exchange; section 4 conjectures that the knowledge graph technology, *i.e.*, a combination of ontologies and agents, will be the driving force towards laboratory automation and beyond; finally, section 5 concludes our proposal.

2 Platform-based approach

Detailed reviews of the applications of the closed-loop optimisation have been published by Cao et al. [13] and Coley et al. [24]. In this section, we focus on the data flow between the different components of such an automated experimentation platform as presented in the state-of-the-art studies. To have a clearer demonstration of the data flow between different parts, thus revealing how these functional components can be shifted into agents as in the knowledge-graph-based approach, we re-group the five key elements proposed by Flores-Leonar et al. [40] and recast them as illustrated in Fig. 1. The receptionist acts as a human-machine interface that receives, analyses, and translates the requests into machine-understandable objects, as well as enables real-time and interactive communication between user and data. The coordinator manages the workflow by locating resources given constraints, requesting data from the librarian, asking the planner for suggestions

over the next steps, and requesting experiment from the executor. The planner is a decision making entity that designs the experiment, plans retrosynthesis steps, also selects suitable surrogate models given use-cases. The librarian is responsible for data management, including maintenance of the database, data cleaning, data validation, and outlier detection. The executor performs the computational and physical experiments, both interfaced with the available experimental resources. We categorise the selected studies into the realisation of functional components and assess the data communication between each of them. It should be noted that we do not cover the specific internal realisation of the components, *i.e.*, we do not consider how the planner handles the input historical data and how it recommends the synthesis route, instead, we focus on the format of the recommendation output from the planner. Following the review, we list the limitations of the platform-based approach which lead to the quest to better data representation and exchange protocols.

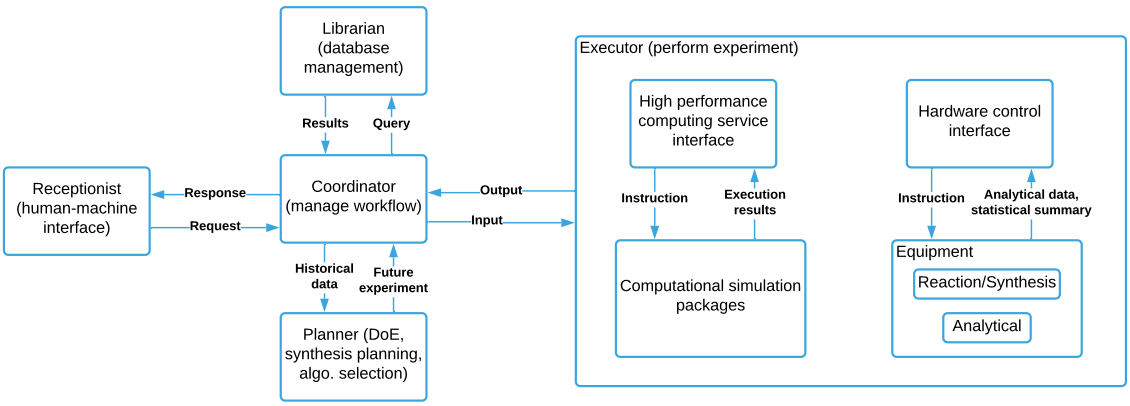


Figure 1: *Functional components of a platform-based approach towards chemical discovery, annotated with the communications between each component.*

2.1 Selected studies

There have been extensive reviews on developing each of the functional components [36, 40, 47, 56, 82]. In the context of chemical automation, Mateos et al. [89] reviewed the realisation of the components in selected continuous flow platforms. In this review, we selected the studies below to illustrate how the data is exchanged between the functional components in the platform-based approach. Specifically, we will review the data exchange protocols between the coordinator, librarian, planner and executor for further investigation on interoperability within one platform and between different platforms in the current setups. We identified three main types of data representation and storage in the automated experimentation platforms, namely, variables stored in a reserved memory location of programming languages, data stored in a file on a hard disk, and data stored in a database. Based on this classification, three types of data transfer and communication protocols were identified as assigning in-memory cache values during software programme run-time, file transfer protocol, and HTTP request/response. It should be noted that although both the latter two ways of communication belong to the application layer in the

TCP/IP model, they are distinguished herein to emphasise the format in which the data is stored and consequently transferred. To the best of our knowledge, the complete details are summarised in Tables 1 and 2.

2.1.1 Receptionist

The receptionist is a human-machine interface enabling the experimentalist to access data, request for experiment execution, and inspect experiment progress. Among different platforms, multiple ways of interaction have been reported. Knight et al. [75] present a voice-controlled user interface integrating voice, text, and visual dashboards. This increased the flexibility for the experimentalist to communicate and collaborate with the automated setups without the coding experience required. Web interfaces via HTTP requests/responses [37, 38, 67] is another way of interaction. The advantage of this approach is that authorised users can log in to the web page and access the platform from all over the world [36]. Moreover, the natural language processing (NLP) modules can build on top of the web interface as chatbots, which can further connect to existing messaging services such as Gmail, Twitter, Slack, and Dropbox [88, 122]. The graphical user interface (GUI) is a more intuitive way of interaction between the users and the automated experimental platforms. It can be built through different coding software, such as Matlab [70], Python [83, 138], and LabVIEW [6, 17, 97]. It should be noted that each receptionist can only work within its own operating system due to its bonded communication protocols as well as the coding language.

2.1.2 Coordinator

The coordinator manages the workflow in the closed-loop system, ranging from the information within the library, the update of the AI models, to the execution on the robotic platform. Different programming languages/tools have been employed to develop the coordinator. For Python-based coordinators, the Aspuru-Guzik Group proposed ChemOS [88, 122], a modular coordinator orchestrating the learning module (the AI-based planner), the communication module (server-based receptionist) and an operation module for remote control of the robotic platform. ChemOS demonstrated decision-making capabilities in managing the workflow for thin-film material discovery [88] and increasing the efficiency of organic photovoltaics [81]. Zhu's group presented MAOSIC [83], a coordinator upgraded from their previous system MAOS [84], allowing virtual-reality human-robot interaction. It was applied to the autonomous discovery of optically active chiral inorganic perovskite nanocrystals. LeyLab [37] is a coordinator orchestrating multiple users and equipment in different continents for the development of catalysts and process conditions in flow reactors. The firewall within the coordinator prevents malicious attacks from unauthorised users. Chemputer [138] is developed for organic synthesis optimisation within batch reactors. This coordinator brought together synthesis abstraction, chemical programming and hardware control, and tested the synthesis of three small pharmaceutical compounds with similar yields to those obtained by manual work. Moreover, by using a standardised format for reporting a chemical synthesis procedure within the coordinator, Chemputer captures synthetic protocols as digital code that can be further published, ver-

sioned and transferred flexibly. The Lapkin group presented a Matlab-based coordinator for multi-objective optimisation of the reaction conditions for SNAr and N-benzylation reactions [133]. It demonstrated its flexibility to a different chemical system with an aldol condensation reaction optimisation [70]. There are also coordinators based on LabVIEW. Given the user-friendly graphical programming interface in LabVIEW, building a receptionist module is not required in this setup. However, Matlab [97] or Python [17] are occasionally paired up with the LabVIEW to enable the planner module to suggest new experiments. It can be seen that coordinators followed different coding philosophies in different programming languages. For each case study, the reported coordinator indeed satisfied the specific need yet fail to extend to other systems.

2.1.3 Coordinator - Librarian

The interaction between the coordinator and librarian focuses on reading historical data and writing new data for data storage. Depending on the operating system of the coordinator, as well as the structure of the librarian in each platform, the data communication protocols between the coordinator and librarian are various. An intuitive approach is to store and transfer the data as variables in the memory of the operating system. Jeraal et al. [70] stored and transferred data as Matlab variables. Similarly, Christensen et al. [19] used Python variables for communication. This approach is lightweight and independent of the database structure. However, it is vulnerable as there is no backup for the data obtained. Moreover, the data stored are hard-coded and picked beforehand, meaning the variables will be reassigned during the iterations.

File transfer is an approach to overcome this issue. Cao et al. [12, 13] used CSV files as the bridge for communication. Other studies used MAT files in a similar fashion [6, 154]. In this approach, the experimental results were exported and stored as a file that can be loaded later for suggesting the next experiments. Compared to storing data as in-memory cache variables, the file transfer approach gives a way to back up the data on a separate machine or online server with flexible access and secure storage. However, the files can still be hard to track and classify when the number of experiments is high or more than one type of experiment is run on the platform.

Some groups deployed database queries for the experimental data reading and writing between the coordinator and the librarian. Li et al. [83] stored long-term data through SQLAlchemy which supports a database management system (DBMS), with databases such as MySQL, Postgres, Oracle, and SQLite as the back-end. The coordinator MAOSIC can read and write new entries to the server-based database via API. In Roch et al. [122], the coordinator ChemOS was connected to SQLite, and the information was stored in four distinct databases (requestDB, parameterDB, robotDB, feedbackDB) on SQLite to better classify the data and retrieve them in the later stage.

2.1.4 Coordinator - Planner

To avoid an exhaustive search of the chemical space, the planner needs to decide which new experiments should be conducted. Depending on the purpose of the platform, the

planner algorithm can be classified into discovery and optimisation. Detailed reviews of the existing algorithms for planner have already been published; interested reviewers refer to Garud et al. [44] and Clayton et al. [20]. The coordinator passes the planner data it requires, then the planner works out the next steps and passes the instruction back to the coordinator. The communication between these two is mainly done in two ways: variable stored in memory [6, 14, 88], and file transfer [12, 23, 133, 138]. It is worth mentioning that the communication protocols are not necessarily the same over one platform. Li et al. [83] used database queries for the interaction between the coordinator and librarian, yet they depend on Python variables for the communication between the coordinator and planner. It can be seen that the platform-based approach can adapt to different ways of data exchange, yet modifications that are case sensitive will be needed.

2.1.5 Coordinator - Executor

The executor is the module that runs the experiments, computationally or physically, and sends back the experimental results. The coordinator sends the high-level instruction to the executor for the experiment to be executed, either computational or physical, the executor returns the results to the coordinator. The interaction between the coordinator and executor module highly depends on the operating system for the instrument, as the actual experiment resources within the executor are normally surrounded by a layer of the interface. Therefore we review the communication protocols of the physical and computational experimental platforms separately.

Physical experiment interface Robotic platform have their origins in instances such as peptide synthesis [95] and the pharmaceutical industry [86, 158]. Some existing commercially available semi- and fully-automated platforms in chemistry have emerged as powerful tools and can be embedded into the closed-loop optimisation system [40].

Commercial platforms provide various high-throughput workflow solutions, ranging from single bench-top/standalone automated workstations up to complete and integrated product development workflows for the entire product development processes in chemical material science [63, 91]. Greenaway et al. [52] applied the Chemspeed Accelerator SLT-100 synthesiser platform in the discovery of porous organic cages and the optimisation of the cage formation conditions. This platform can carry out up to 96 reactions in parallel, highly speeding up the testing of the proposed experimental conditions that are sent to the platform via file transferring within the Chemspeed custom software. Vapourtec delivers automated flow reaction platform with multiple choices for pumps, and flow reactors. Successful examples of using the Vapourtec system in the closed-loop optimisation setup include drug discovery [45], scale-up development [149], and reaction condition optimisation [70, 133]. It is worth mentioning that commercially available mobile robots and robotic arms have been used in complex and multi-step operations [11, 23]. Communication between the coordinator and the robots was achieved using various communication protocols (TCP/IP over WIFI/LAN, RS-232, websocket, *etc.*). Although commercial systems developed by various vendors are easily implemented with a user-friendly user interface, it limits the experimental choice across platforms, and it is hard to configure the platform to the existing workflow architecture and setups in the lab.

To enable a modular-based plug-and-play platform, single-board controllers, *e.g.*, Raspberry Pi and Arduino, were used to act as the interface layer connecting the coordinator to the actual experiment executor, *i.e.*, sample preparation, analytics *etc.* This is favoured by the academic community due to its flexibility and compatibility with different experimental instruments at a relatively low cost. The communication protocols between the coordinator, single-board controller and experiment executors are various. A TCP/IP protocol was used in the cases where a Raspberry Pi was applied. Fitzpatrick et al. [38] used a VLAN to control around lab equipment, also an SSH tunnel between the virtual environment and the remote control server. Similarly, Roch et al. [121] controlled the pump system using the Raspberry Pi and interacted via an SCP with the executor codes. In Chemputer designed by Steiner et al. [138], an Arduino was designed as the microcontroller. Instances of experiment executors are created as Python instances at the initialisation stage and the coordinator reads related information stored in a GraphML file. Li et al. [83] conducted their high-throughput experiments via an Arduino control board as well but followed the JSON-RPC 2.0 protocol used for robots and characterisation equipment control. A detailed review of microcontrollers and their applications in automated experimental systems can be found in Fitzpatrick et al. [39]. The in-house built platform can connect to different lab equipment based on the users' need and existing lab setup, yet different communication protocols prevent it from extending to other lab/systems.

Computational experiment interface With the rapid development of computational power and simulation methods, computational experiments are playing a more vital role in catalyst design and optimisation [148], synthesis planning [142] and catalyst discovery [146]. By using theoretical, fully automated screening methods combining ML and optimisation to guide density functional theory (DFT) calculations, Tran and Ulissi [145] screened across intermetallics for the discovery of electrocatalysts for CO_2 reduction and H_2. The main executor for computational experiments is the high-performance computer (HPC). However, the interaction between the HPC and the coordinator on local computers is different from case to case. The scheduler is the interface for the users on the login nodes to send work to the compute nodes on the HPC, as the users cannot run their calculations directly and interactively (as they do on their personal workstations or laptops), instead they need to submit non-interactive batch jobs to the scheduler. The scheduler stores the batch jobs, evaluates their resource requirements and priorities, and distributes the jobs to suitable compute nodes. There is quite a few open-source scheduling software depending on the setup of HPC, among which SLURM is widely used in research computing services [118]. Rosen et al. [124] developed and used the PyMOFScreen Python package to manage and carry out the automated DFT calculations, leading to new electronic structure database constructions and accelerate new materials discovery [123]. Multiple software packages were developed to enable high-throughput screening on the HPC, such as Python Materials Genomics (pymatgen) [111], FireWorks [69], custodian [111], Atomate [90], and GASpy [145, 146]. Depending on the user's need as well as the DFT calculation software, the structure and the output file of those Python packages are different and non-transferable.

2.2 Current limitations

Despite the huge improvements made in the literature, a few limitations remain. The platform-based approach presented heavily relies on the coordinator. This increases the possibility of data loss during transmission, and it will become unsustainable soon with further expansion of the ecosystem. Direct communication between functional components is one potential approach to mitigate this issue, as demonstrated by Fitzpatrick et al. [38] in letting the planner directly communicate with lab equipment via TCP/IP.

Another limitation is the *ad hoc* data representation and storage. This is particularly important as there is no standard method of representing results or recipes for chemical experiments, despite several competing standards of representing molecules co-exist. The heterogeneous data format lacks interoperability that precludes the full utilisation of the embedded information. This problem is further exacerbated when the collaboration between different groups is considered; potentially data generated from one group will be shared and tested on the platform of another group for reproducibility and further experimentations. Moreover, the consequent various data transfer and communication protocols result in low extensibility issues as a considerable amount of time is often required when new hardware or software is integrated, also noted by Breen et al. [10].

Unbalanced chemical data is another limitation to be addressed [25]. In ML applications, historical data from reaction databases are normally applied as the training set to guide the learning of the planner models. However, only 'good' experiment results are published and stored in these databases, limiting the opportunity of learning from 'bad' examples [117]. Not to mention those platforms generating experimental data from scratch, without utilising the prior chemical knowledge at all. A further issue lies in several examples where users are required to manually input chemical data [70, 137]. This is error-prone and limits the potential of full automation.

In brief, improving the interoperability within one platform and between different platforms is a key step in lowering the entry barrier of digitalising chemistry and promoting a fully automated laboratory. It is thus important for us, as a community, to know how far we are from meeting the prerequisite condition – a fully interconnected data representation capturing the data generated within the experimentation.

3 Data representation and exchange protocols

As promoted by various researchers [25, 56, 61, 155], the digitalisation of chemistry facilitates the collaboration between research groups. Figure 2 reviews data representation and exchange from the different perspectives of a chemical experiment, namely, molecule, reaction, analytical data and method, procedure and hardware, and finally holistic data capture and exchange. Importantly, we distinguish the community efforts into non-semantic and semantic paradigms depending on whether chemical ontologies are involved, and lay out the connection between them. The agent-based approaches towards standardised and effective communication between each of the components involved are discussed.

3.1 Non-semantic representation

In this review, we broadly distinguish non-semantic efforts into four parts: a representation of cheminformatics formats, a schema for constrained encoding of data, a collection of data stored in a database, and finally a holistic architecture that aims to capture all data generated within an experiment.

Since the discovery of the periodic table of the elements, chemical knowledge is built on structures with competing representations [162]. The most commonly used representation is string and line notation, including SMILES [153], InChI [59], SMARTS [27], SELFIES [77], *etc.* for molecules, and RInChI [53], SMIRKS [28], *etc.* for reactions. Chemical table files express molecules and reactions in terms of *x*-*y*-*z* coordinates of atoms and bonds. For a more visual representation, molecules and reactions can be illustrated with 2D line drawings (or 2.5D including stereochemistry), and 3D conformers. These formats are interchangeable with the help of cheminformatics tools, *e.g.*, Open Babel [109] and RDKit [80]. An ML application normally starts with encoding structural representations in the form of high-dimensional vectors to map the implicit chemistry to either physicochemical properties of one molecule or reactivity between different molecules.

Popular chemical databases and registry systems normally store various representations of the above with registry numbers, *e.g.*, IUPAC name, CAS number and PubChem CID, for unique and unambiguous identification within themselves and cross-reference between repositories. PubChem [72] is the largest open-source structural chemical information repository. For reaction informatics, the scale of open-source databases is much smaller. The USPTO database [87] is one of the seminal databases in the community that contains 3.7 million reactions extracted from US patents. It was commercialised as Pistachio [104] containing more than 13 million reactions with annotated reaction classifications using named reaction ontology (RXNO [131]) and expanded coverage to other patent offices, *i.e.*, World Intellectual Property Organization (WIPO) and European Patent Office (EPO). Despite the public availability of the USPTO database, its representation schema, *i.e.*, Chemical Markup Language (CML) in eXtensible Markup Language (XML), requires extra efforts of format transferring for ML applications. This results in different versions of the USPTO subset that were derived and adapted by various researchers for their applications, as compared in Schwaller et al. [132]. As the tailored database can be kept private to the research group, it could be difficult for bench-marking new algorithms.

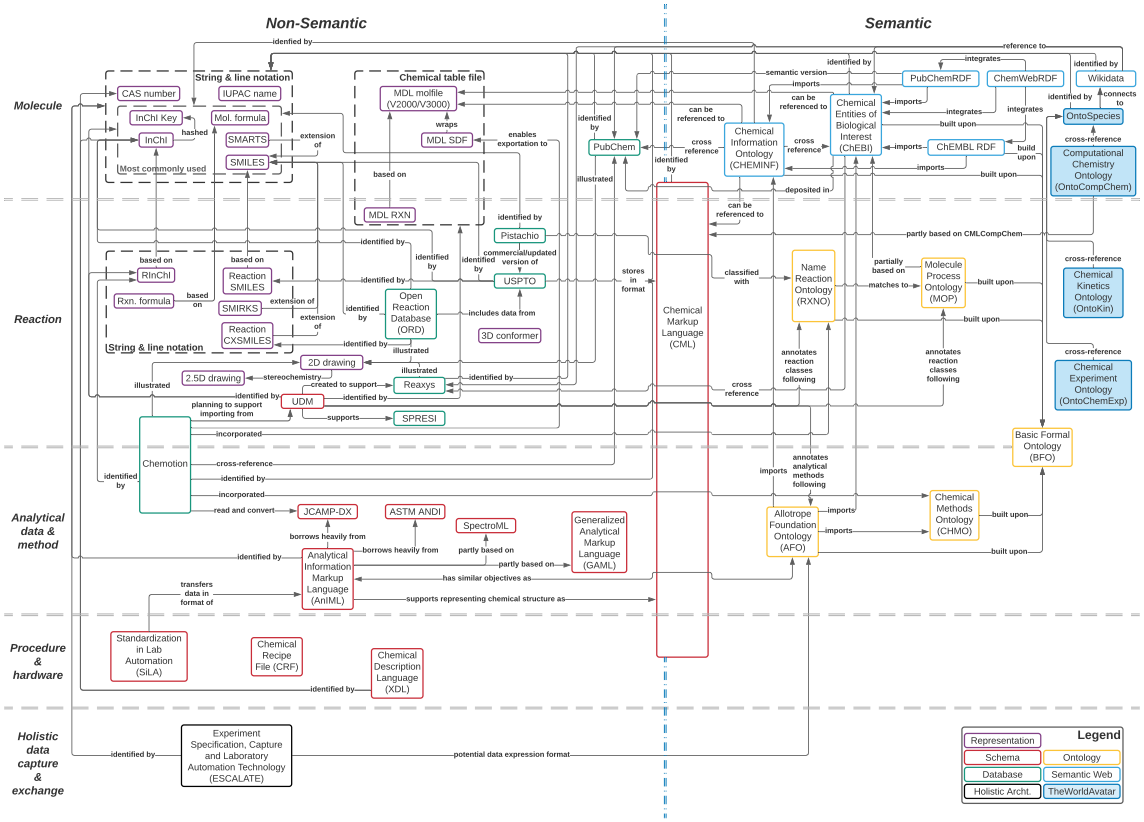


Figure 2: *The community landscape towards a better data representation and exchange in chemical digitalisation. The focus of each category: (a) molecule: chemical structure, physicochemical properties, spectral information of a given species; (b) reaction: chemical reaction scheme, conditions, description of procedures, and statistic summary of the reaction outcome; (c) analytical data & method: analytical data collected and the methods applied within the experimentation, this is distinct from the spectral information of a given species as this focuses on the data collection process; (d) procedure & hardware: the operational procedure in an experiment in the format that can be directly executed by hardware; (e) holistic data capture & exchange: the initiatives to capture all the experimental information generated within the experiment and the exchange of data between different hardware/software. For those on the fence between two categories, we meant they cover both areas. Chemical Markup Language (CML) was labelled as both semantic and non-semantic since it preserves hard-coded and rule-based semantics but not ontologies following semantic web standards [151]. Basic Formal Ontology (BFO) is an upper-level ontology as the basis of other ontologies and it does not capture any domain-specific information.*

To facilitate the development of ML in chemistry, Open Reaction Database (ORD) [112] was formed to encourage precompetitive data sharing in a standardised format. It records how the reaction was performed, including reaction inputs, conditions, outcome, *etc*. Notably, ORD uses a protocol buffer as its data structure, instead of the commonly used XML schema. It deliberately avoids the use of ontologies due to insufficient ML applications with ontologies seen in the community [106, "The Open Reaction Database"]. Despite ORD storing the operation sequence in a machine-readable format, the authors declared it a non-goal at present to make it compatible with programmatic execution on automated synthesis hardware. For more complex operations, ORD only supports a free-text description of the procedure. In terms of the reaction outcome, it focuses more on the statistical summary of the reaction, *e.g.*, conversion and yield, and unprocessed analytical data if available. At present, ORD contains 2 million reactions [106, "The Open Reaction Database"], including part of the USPTO dataset that was converted from CML.

Unified Data Model (UDM) [116] is another initiative aiming at capturing and integrating the experimental information generated during the chemical synthesis. UDM was originally developed by Roche as a transfer model of MDL RD file format for integrating data from various sources into Reaxys database [51]. It has since evolved to an XML schema with three main elements, namely, *citations*, *molecules* and *reactions*. In addition to recording the molecule and reaction identifiers, UDM annotates its data with semantic vocabularies. The reaction classification is based on the molecular processes (MOP [31]) and RXNO ontologies, demonstrated by its sample data taken from Reaxys. The analytical method and results type are based on a working draft version of Allotrope Foundation Ontology (AFO [96]) where duplicate entries exist. However, it should be noted that the way UDM integrates the ontologies is by enumerating the ontological classes as a subschema of UDM and tagging them to the XML elements as attributes. One general issue with this type of enumeration and attribution is that the relationships declared in the ontologies are not retained in the XML schema, *e.g.*, class and subclass relationship between concepts in MOP and RXNO, and the corresponding relationship between result types and analytical methods in the AFO. By looking at the publicly available resources, there are no programmatic constraints over how ontological axioms are enforced in a UDM file. Moreover, UDM allows any type of format for the analytical data recording, at least by XML schema itself, tailored tools would be necessary for better utilisation of the data. In its latest release, UDM extends its support to the SPRESI database [125]. Moving forward, UDM aims to provide fully captured representations of reaction predictions and optimisations for multi-step reactions. Additional support for environmental health and safety data is also of interest [106, "The Unified Data Model (UDM)"].

Similar to ORD, Chemotion [147] aims to build a community-driven repository to better publish reaction data generated across different laboratories. In practice, despite containing less data, a key distinguisher of Chemotion is its level of interoperability in enabling programmatic transfer of raw analytical measurements for integration of electronic lab notebook (ELN) from individual laboratories. It does so by supporting reading and converting analytical data in the widely-used JCAMP-DX format [79]. Each published reaction in Chemotion has a semi-machine-readable format with a digital object identifier (DOI). It cross-references compound entries in PubChem. Like UDM, Chemotion incorporates ontologies (RXNO and chemical methods (CHMO [32])) for semantic annotations at a vocabulary level. On the data validation front, Chemotion automates curation

of some types of analytical data, *e.g.* plausibility checks of nuclear magnetic resonance (NMR) data. Human inputs are still required to ensure data quality for publication. To enable more data resources, Chemotion is planning to support reactions stored in a UDM format. Chemotion is also planning to connect ELN to robotics to establish an automated platform for chemical synthesis [106, "Documentation and publication of reactions with Chemotion ELN and Repository"].

As mentioned, JCAMP-DX is a data standard widely-used for recording and sharing analytical data. However, one drawback to its utilisation is the lack of validation tools making it difficult for data generated from different software to adhere to the standard terms [144]. One approach to alleviate this problem is modernising the standard terms with an XML schema, such as Analytical Information Markup Language (AnIML) [2]. AnIML is partly based on SpectroML [126] and Generalized Analytical Markup Language (GAML) [144], also draws from JCAMP-DX and ASTM ANDI. On the chemical structure side, AnIML supports the CML format together with other commonly used line notations. AnIML aims to provide vendor-neutral analytical and biological data representations that are designed for manufacturers to install and maintain. For the same reason, AnIML provides audit trials and other metadata for reporting information in regulatory processes. At present, AnIML supports most common analytical equipment with detailed documentation for ultraviolet–visible spectrophotometry (UV/Vis), chromatography, and indexing.

Up to this point, reviewed efforts are standardising the data generated during the experiment. Initiatives exist to standardise the instrumentation interface, *e.g.* Standardization in Lab Automation (SiLA) [136]. SiLA is a micro-service architecture using gRPC and HTTP/2 protocols with a protocol buffer as its payload. It adopts a client/server view to describe the devices in the lab environment, where entities expose (multiple) services as SiLA Features accessible to others. SiLA Features are expressed in a predefined XML-based schema and stored in an online repository for service discovery. Each feature is assigned with a unique identifier to enable peer-to-peer interactive communication, status queries, and reactions to events. As SiLA is a communication protocol for equipment control, it utilises AnIML as the medium for the bidirectional transfer of analytical data between laboratory information management systems (LIMS) and chromatography data systems (CDS) in a file-less fashion [129]. The combination of SiLA and AnIML represents a promising direction: standardised interfaces for instrumentation and unified machine-readable data representations. This results in a complete data package after completion of the analytical experiment, including all the process steps and the generated data.

Whilst SiLA standardises equipment interface, chemical recipe file (CRF) [23] and chemical description language (XDL) [93] are initiatives to automate experiment execution. They both focus on translating the operational procedures from unstructured descriptions to robot execution commands.

CRF [23] is a CSV-based schema developed for flow synthesis. Since the instructions are generated based on batch reaction data, human modification is required to enable continuous processes. One notable aspect of their setup is their modularised reaction hardware, making it robotically self-reconfigurable, as demonstrated by the back-to-back synthesis of medicinally relevant small molecules.

XDL [93] is an XML schema focusing on batch synthesis. It contains three main com-

ponents as the apparatus to be employed and manually configured, chemicals to be used, and robotic steps abstracted from operations used by chemists in the lab. An ontology is proposed to map the command and hardware executions, however, it is not published in semantic web standards [151]. Before the instructions are sent to execution, researchers can modify the conditions to benefit human intuitions.

Both CRF and XDL focused on providing a flexible framework to conduct synthesis for multiple molecules. However, neither of them included an automated analysis step. The statistic summary of the chemical synthesis is thus not provided in a standardised format as done by other reaction schemas.

Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE) is an attempt towards holistic data capture and exchange [114]. It proposed an ontological framework for experimentation, supporting data collection, reporting and experiment generation. However, the claimed ontological framework was realised by implementing a systematic naming scheme for the files storing the captured data, *e.g.*, CSV and textfile. It is still a heterogeneous data representation and semi-structured data storage without semantic features. Despite human researchers still being required to conduct part of the experiment operations to close the loop, acknowledged by the authors, this framework captures and reports all the reactions conducted, including "bad reactions" – in line with the cultural change promoted by the community [61]. At present, similar to ChemOS, ESCALATE uses a file-sharing folder infrastructure (Google Drive). Their initial implementation was used to perform algorithmically-controlled metal halide perovskite crystallisation experiments. Further discovery of the formation of two new perovskite phases was demonstrated [85]. In the future, this framework might be upgraded into a full ontological framework as the authors acknowledge that the Allotrope Foundation Data Standard can be incorporated into their data lake.

In general, the non-semantic efforts are closely connected to each other. Multiple representations are normally used within schemas or databases to meet the needs of different applications. Databases cross-reference to each other using registry numbers. Most of the schemas annotate their elements to ontological taxonomies for classification but without supporting the full ontologies. Despite Chemotion demonstrating interoperability to some extent, it generally remained an issue to be addressed.

Another notable trend is the adoption of XML schema as data structures. XML is a machine-readable format for algorithmic operations. It relies on string parsing when automating some of the processing steps. For example, the automated unit conversion provided by XDL, where the case-insensitive conversion to a standard unit was performed. However, XML is not designed to host large sets of data as querying between different files can be challenging. The linkage between entries in XML is implicit and requires tailored codes to handle. A solution to this problem could be hosting data in a database and exposing that as the query interface. Yet as demonstrated in the platform-based approach, the same scalability issue would emerge.

3.2 Semantic representation

Since the landmark publication by Berners-Lee et al. [7], the semantic web field has envisioned the next generation of the web in both a human- and machine-readable format for better data sharing among mankind and faster data processing using computers. Through ups and downs, the semantic web community has pivoted from ontologies to linked data, and further to knowledge graphs, which are gaining attention again in recent years. For a comprehensive review of developments in the semantic web field, interested readers are referred to Hitzler [62]. The focus herein is the uptake of such technologies in the chemistry domain, as illustrated in the right half of Fig. 2. For initiatives where only TBox are available, we labelled them as "Ontology", whereas ABox are published are labelled "Semantic Web". Those under "TheWorldAvatar" will be introduced in the next section.

Chemical informatics has a long history of utilising semantic web technologies. The chemical semantic web [21, 46, 103] is one of such early attempts by Murray-Rust and co-workers, contemporaneously to Berners-Lee's proposal of the semantic web [7]. In their work, CML was employed to host the data, prior to Web Ontology Language (OWL) becoming the semantic web standard. CML schema covers concepts related to atoms, molecules, computational chemistry, crystallography, spectra, chemical reactions, and polymers. It greatly influenced the development of reaction informatics, especially, it is the molecule representation implicitly used by various cheminformatics software [102].

Since OWL became more and more popular in modelling ontologies, more activities of ontology development have been demonstrated in the scientific domain. Despite the authors of CML holding the view that ontologies following the semantic web standards [151] are "too complex for the chemical community to take on board, and provides little effective added value" [101] compared to their approach, the benefit of semantics motivated the development of chemical ontologies to a great extent, especially work at Royal Society of Chemistry (RSC) [5], *i.e.*, CHMO [32], RXNO [33], and MOP [31]. These ontologies are sophisticated and carefully curated. As demonstrated in the non-semantic efforts, they are widely-used for annotating reaction classes and analytical methods.

Another driving force of ontology development in the chemistry and biology domain is the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI). In contrast to RSC ontologies that only provide concepts, EBI ontologies provide knowledge at both a terminological and assertional level, covering small molecules (ChEBI [58]) and cheminformatics (CHEMINF [57]) in a cross-referenced fashion. CHEMINF supports molecular structure representations in the CML format, it also partly transformed data from PubChem into a knowledge base together with cross-reference to their PubChem entries. ChEBI deposited its data in PubChem entries and cross-referenced to Reaxys entries. These ontologies complement other ontologies in the field. For example, CHMO intends to describe the physical and practical methods, whereas CHEMINF covers the computational and theoretical ones.

Ontologising existing databases was demonstrated in the community, including ChEMBL RDF [157] and PubChemRDF [41], the semantic version of the current largest open-source chemical information repository – PubChem [72]. However, the Resource Description Framework (RDF) version of these databases did not come with an officially supported SPARQL Protocols and RDF Query Language (SPARQL) endpoint. Galgonek

and Vondrášek [42] recently addressed this issue by integrating PubChem, ChEMBL and ChEBI datasets as a PostgreSQL database and exposing that to support SPARQL queries. This enabled fast access to chemical data from different sources.

Allotrope Foundation is a collaborative effort from the pharmaceutical industry [96]. Similar to AnIML, it aims to propose a common data exchange format to unify the laboratory information technology (IT) landscape. It started from realising the vision of Roberts et al. [119, 120] where an XML schema was envisaged to provide a holistic data format. It later decided to store data based on HDF5 and RDF formats that were controlled by ontologies for semantic capabilities. The foundation now contains three ontologies, namely, AFO, Allotrope Data Format (ADF), and Allotrope Data Model (ADM). AFO is the ontology at the TBox level representing the knowledge in the chemistry domain and it borrows heavily from CHMO. ADF refers to the ontology ABox classified by AFO, extended with more features on data structure and provenance for long-term archiving. ADM is the constraint for how data in ADF should be modelled following AFO. However, only AFO is freely accessible to the public, with the remaining resources restricted to community members.

Compared to non-semantic efforts, a key distinguishing factor of the semantic approach is its fully-linked concepts and data instances. This is particularly true for the ontologies reviewed above, as their concepts follow the classification of the Basic Formal Ontology (BFO). The instances stored under each ontology are inherently linked and consistent in logic. This enables interoperability between domains and easy access to data from different sources via SPARQL queries. Moreover, the linked nature made it possible to reduce duplication of information by providing unique identification to the entities, whereas in XML it would be more likely that the same information would appear in different files, *e.g.*, when the same molecules are involved in different reactions.

The biology community has demonstrated the population of data is the key to a broader impact with well-defined ontologies [4]. However, classifying and annotating data into ontologies while maintaining logical consistency is a challenging task, especially with complex ontologies. It is costly to adopt and creates a high entry barrier. This is reflected in reaction informatics, as ontological data is still very much limited to chemical species information, and there is currently no semantic version of reaction data available. This further exacerbated the problem of insufficient adoption of semantic web technologies in ML and other practical engineering applications, as noted by the developers of ORD [106, "The Open Reaction Database"]. Not to mention to actually control the equipment execution and automate the data exchange framework is even more challenging. A trade-off between engineering practices and comprehensive representation is thus important. A potential solution to this would be to convert existing databases [94] into RDF.

The same issue was acknowledged by the Allotrope Foundation [96] that there is a trend of making simpler data models for practical applications. One of their partner companies, TetraScience, developed an Intermediate Data Schema (IDS) – a JSON-based schema of analytical data as the precursor of the AFO format. Using an agent, data generated from the analytical equipment was collected and converted to ADF for further analytics. Despite of being proprietary, it enlightens the way forward to standardise data conversion and integration while it is generated. A perspective from Godfrey et al. [48] backed this idea, *i.e.*, data stored in an ontological framework would very much facilitate the proliferation of interoperable standards, also keep the flexibility of introducing new methodologies.

3.3 Agent-based approaches

With the ontological data representation, the way of data generation and consumption is another issue needing to be addressed. By definition, an agent is a piece of 'automated' software programme capable of acting towards achieving its objectives [127]. In such a process, they can communicate and coordinate, *i.e.*, exchange information with each other, in a standardised format. As aforementioned, TetraScience utilises agents to standardise data generation, this section focuses on agent applications in standardising the data utilisation.

In the context of chemical automation, agent-based approaches can be adapted to replace the functional components within a platform-based approach. Montoya et al. [98] wrapped different algorithms as agents to suggest the next experiments for DFT calculations on stable materials discovery. Gomes et al. [50] standardised various tasks as agents (bots) in a platform for crystal-structure phase mapping. Caramelli et al. [14] applied agent-based model simulations to showcase the effectiveness of multi-threaded networking principles in searching for the optimal solution in the chemical space.

In the above studies, a step was made to turn functional components into modularised agents and standardise the data exchange between them. However, the communication was done by passing in-memory programming variables [50, 98], or posting plain-text on a human messaging platform (Twitter) [14]. As discussed in earlier sections, the same drawbacks such as lack of scalability and interoperability will emerge when scaling up the framework and integrating computational and physical experimentation.

Following the introduction of ontological data representations, a natural question would be if we can merge these two technologies into one to make good use of each strength. Relationships between semantic web technologies and agent systems have been complex since they were put together back in the 2000s [60]. In theory, the ontology can help agents with more flexible operations, whereas agents can help the ontology for better data utilisation. The Foundation for Intelligent Physical Agents [143] (FIPA) proposed a set of specifications focusing on communication and interoperability between agents. In one of their specifications, *i.e.*, FIPA Ontology Service Specification, an idea of having an ontology agent to support the message interpretation between agents was elaborated in detail. However, it never made it to the standard stage. In the following years, JADE [68], a Java-based software platform that simplifies the implementation of FIPA-compatible multi-agent systems, attempted to provide an ontology in its realisation of FIPA standards, but they only provided the ontology as part of the Java code, without connecting to a knowledge base. Attempts to merge the two technologies have been seen in other domains, but not much in chemistry until very recently – 'The World Avatar' project.

4 Dynamic knowledge-graph-based approach: The World Avatar

In this section, we advocate a combination of semantic web technologies and multi-agent systems – a dynamic knowledge-graph-based approach – to be the driving force towards a complete digital and self-driving laboratory, *i.e.*, a chemical digital twin. In particular, we introduce 'The World Avatar' project, its potential application towards laboratory automation, and its connection to a wider context. Before diving into the details, below is a glossary of terms that are heavily used in 'The World Avatar' context. These definitions are definitions of terms as we use them, whilst acknowledging that they may have different meanings in different contexts – we make no attempt at general definitions here.

Knowledge graph: a collection of data and software agents expressed as a directed graph controlled by ontologies, where the nodes and edges refer to concepts and relationships correspondingly. This has broader coverage than the knowledge graph as commonly used in semantic web studies [62], where only data are modelled as a directed graph. This is also different from the knowledge graph built based on Reaxys by Segler and Waller [134] for reaction discovery problems, which expressed molecules as nodes and binary reactions as edges.

Digital twin: a virtual replica of real-world entities in the form of a knowledge graph. It is usually created for the real-time monitoring and controlling of real entities, thus should be synchronous with its physical counterpart.

Autonomous agent: a semantic web service that acts upon the knowledge graph to achieve predefined goals. Importantly, agents themselves are part of the knowledge graph and represented using the ontology for the agent. While active, agents communicate with each other and interact with the knowledge graph for data retrieval and operation. In the sense of a multi-agent system, the knowledge graph is the 'environment' of the agents. The communication between the active agents is conducted via an HTTP request/response. They use ontologies to establish a common understanding of the topic of interest.

Dynamic knowledge graph: a knowledge graph that is constantly modified by agents with the latest status of the real world. It controls and influences the real world by updating the specifications of the digital twin and actuating that with agents.

4.1 Current state

'The World Avatar' (http://theworldavatar.com/) is an all-encompassing modelling framework aiming to create a universal 'digital twin' of the world including anything that potentially can be conceptualised [1, 30]. It does so by developing a dynamic knowledge graph, based on an ontological representation of the physical entities and interoperable agents that keep the digital replica in sync with the real world. Starting from the industrial scale – J-Park Simulator, a precursor of 'The World Avatar'– the developed framework was applied to utilise the waste energy [160], and network optimisation [161] of an eco-industrial park on Jurong Island, Singapore [113].

At the chemical processes level, as shown in Fig. 2, 'The World Avatar' covers ontologies for quantum chemistry (OntoCompChem [76]), chemical reaction kinetics (OntoKin [34]), chemical species (OntoSpecies [35]) and combustion experiments (OntoChemExp [3]). One key feature of the ontological approach herein is that OntoSpecies links three other ontologies to provide unambiguous identification of the chemicals, enabling translation of chemical names when integrating chemical data gathered from different sources [3]. The ontologies can be linked to those described in previous sections to promote further interoperability. In fact, the development of OntoCompChem is partly based on the CompChem terms as described in the CML and the Gainesville Core (GNVC) ontology [18]. To facilitate the automated data utilisation within the knowledge graph, an agent ontology (OntoAgent [163]) was developed as the design pattern of interoperable agents. This allows a standardised and modularised way of agent development. Each atomic agent is capable of predefined simple tasks with their input/output (I/O) signature linked to the concepts in the domain ontologies. This enabled I/O-based service discoveries to form the agent composition for complex tasks [163]. Notably, by using OntoAgent to express the agents as part of the knowledge graph, the activities of agents are easily trackable so that provenance can be recorded to document the changes of the knowledge graph over time.

Various agents were developed to provide service in the chemistry domain, *e.g.*, automated DFT calculations to address inconsistent thermodynamic data [100], automated mechanism calibration to improve the alignment between kinetic models and experimental data [3], and a question answering system enabling intuitive human data interaction – natural language queries of chemical data covering data from different sources [164].

A core strength of the knowledge graph technology is its interoperability, empowered by the linked nature and the semantic representation of concepts. This allows one to combine data, hardware interfaces, and software from different sources in a standardised way, facilitating automation and allowing cross-domain communication of agents [3, 100].

Another key feature of the knowledge graph is its open-world assumption. This enables the scalability of a knowledge graph system. Once the skeleton ontology is set, extending knowledge coverage and tailoring against specific applications is easy to manage. It should work just like adding new features to a computational library. Moreover, once the code of conduct is defined for each of the agents, they can act autonomously and modify the knowledge graph as time elapses. By doing so, the dynamic knowledge graph reflects and influences the ever-evolving status of the real world.

4.2 Automated closed-loop optimisation

The characteristics aforementioned open a new horizon to develop a dynamic knowledge-graph-based approach towards closed-loop optimisation by transforming a case study previously demonstrated in the platform-based approach [70].

Figure 3 illustrates the whole framework consisting of three layers, namely, the real world, the dynamic knowledge graph, and active agents. Reaction data are expressed in ontologies and hosted in the knowledge graph, together with the 'digital twin' of the lab equipment and interoperable agents. Once activated, these agents act autonomously over the

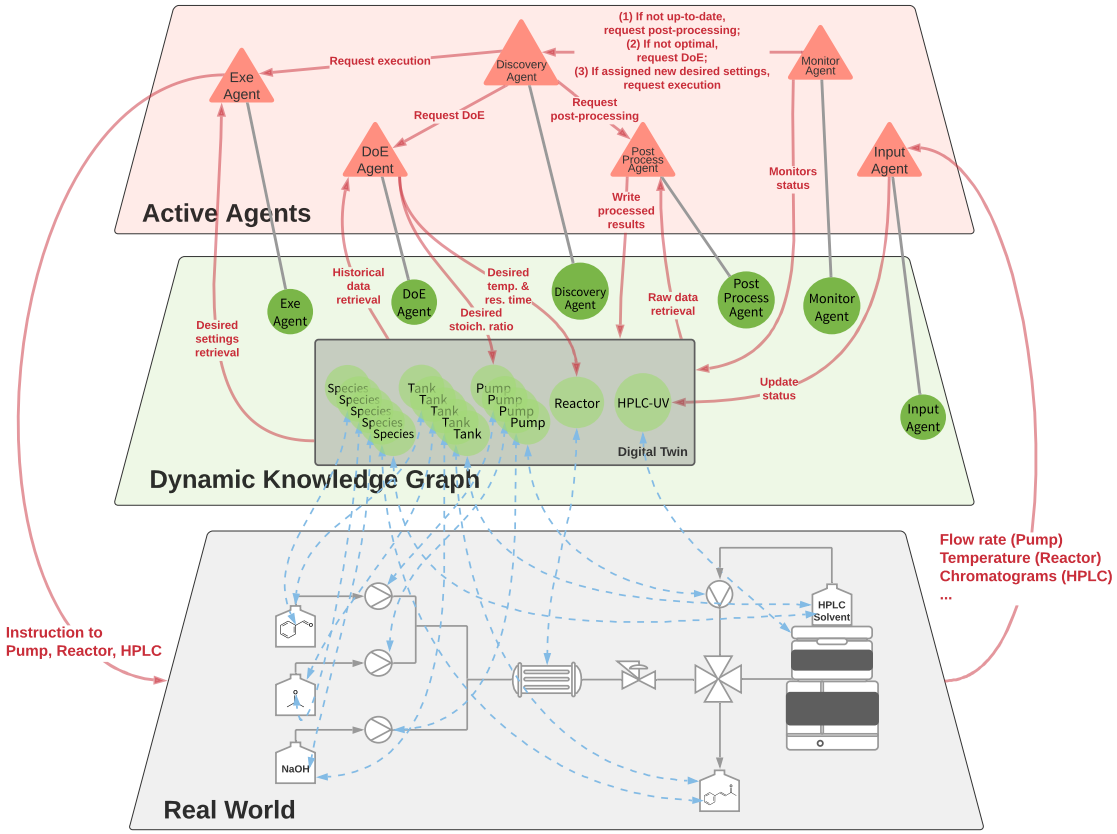


Figure 3: *Dynamic knowledge-graph-based approach towards automated closed-loop optimisation. The real world layer demonstrates the existing physical entities, adapting from the experimentation setup of Jeraal et al. [70]. The dynamic knowledge graph layer hosts all the data generated during the experimentation and a 'digital twin' of the experimentation apparatus. This layer is dynamic as it reflects and influences the status of the real world in real-time. This synchronisation is enforced by the agents in the active agents layer which are instantiated from their ontological representation in the knowledge graph.*

knowledge graph and keep the cyber- and the real-world synchronised. The update of the 'digital twin' is based on the readings from the equipment. This is not only limited to the reaction and analytical equipment but environmental sensors located in the laboratory. Each device has its corresponding input agent transmitting the data into the knowledge graph. The monitor agent is responsible for monitoring the status of the 'digital twin' and assessing if further optimisation is required. If needed, it invokes the design of experiment (DoE) agent to suggest new experiments and update the configurations of the 'digital twin'. The actuation of such settings is the responsibility of the execution agent to reflect the changes made in the knowledge graph. This loop of self-optimisation continues until the monitor agent decides the optimal condition is reached. Importantly, with agents expressed in the OntoAgent format, this framework supports agent discovery service to enable agent-agnostic execution requests.

Compared to the platform-based approach, one distinguishing feature of the dynamic knowledge-graph-based approach is that everything is connected, scalable, unambiguous, distributed, multi-domain, interoperable, accessible, and most importantly evolving in time. Ontologies unify the format of any resources expressed in the knowledge graph, *e.g.*, all the digital replicas of the hardware are expressed in the same way. Therefore, once new equipment is instantiated in the knowledge graph, it can be immediately accessed by any existing software. The same applies when adding new ML algorithms. Once they are wrapped following OntoAgent specifications, standardised interactions with data and HPC services can be established in no time [100]. This enables the rapid integration of the most advanced algorithms and equipment. The orchestration between physical and computational experimentation is made possible. Due to the modularised nature, in contrast to heavily intertwined coding logic within a monolithic application, the duty of development of each component is separated, improving the maintainability of the entire system.

Another advantage of the dynamic knowledge-graph-based approach is its future-proof nature, *e.g.*, its interoperability when integrating with other ontological initiatives in the community. At the species level, OntoSpecies acts like a register system that covers most of the chemical identifiers, making it possible to match with PubChemRDF or other molecular databases. In terms of chemical reactions, OntoKin is already able to describe the kinetic mechanisms of gas-phase chemistry, with OntoChemExp covering the statistical summary of combustion reactions. These concepts can be expanded to describe other chemistry domains of interest. A further opportunity lies in linking the reactions with concepts as defined in RXNO and MOP, embracing their full semantic capabilities. Similar expansion can be made with CHMO or AFO to describe the analytical data and method employed in the experimentation.

4.3 Towards a digital laboratory and beyond

Beyond closed-loop optimisation, various researchers have pictured the future towards the next-generation of autonomous laboratories [10, 25, 40, 56, 67, 119, 120, 155]. Jointly, we listed below a few key challenges and how we see the knowledge-graph-based approach helping.

Data generation, integration, and sharing This challenge lies in the data management practice in the platform-based approach [25, 56]. Going towards a full digitalisation, the ability of to capture all generated data within an experiment (even a 'bad' reaction), integrating it with literature data, and sharing with the community is crucial for navigating in the chemical space. As aforementioned, the knowledge-graph-based approach is designed to be a holistic data capture and exchange framework. With a consensual description of the experiment, literature data stored in the open-source databases can be converted into the ontological format, integrated with the newly generated data.

Roberts et al. [120] envisioned a combination of XML and relational databases to achieve the same goal. However, the authors acknowledged that a database is difficult for a non-specialist to explore without clear documentation. To enable data-agnostic queries within 'The World Avatar', a prototype question answering system was developed to convert natural language into SPARQL queries [164]. Researchers can thus interact with their data intuitively from anywhere at any time, aligning with FAIR principles [156]. The semantic-rich nature incorporates prior knowledge into the data, presenting the potential to explore informed ML applications [150].

Orchestration of physical and computational experiment This challenge lies in the emerging trend of physically synthesising the compounds identified by computational high-throughput screening [10, 22, 25, 26, 52]. Based on historical data, the synthesis steps proposed for the chemical structures are executed by the physical hardware. In a platform-based approach, this requires a heavy workload on the coordinator to manage the information flow and to orchestrate the software and hardware from different vendors. SiLA and AnIML are the initiatives to provide standardised interfaces and data reporting for proprietary hardware, adopting a mindset of peer-to-peer information exchange that is similar to the platform-based approach.

Whereas in a vision by Roberts et al. [119, 120] and our knowledge graph, we promote that information should be accessible to all stakeholders within a laboratory environment, flattening the structural design. Active agents in 'The World Avatar' share the same world-view. The communication between them only serves as a pointer to the correct resources (IRIs). This enables asynchronous communication to accommodate time-consuming activities. Moreover, the communication itself is stored in the knowledge graph and accessible to all agents – everything is transparent and FAIR. By further introducing dependency between different concepts, both data and instructions to the instrument will act like a flow of information travelling in the knowledge graph, analogous to an adaptive organism. Integrating physical entities into the cyber space, knowledge graph technology promotes better utilisation of the plethora of computational power in our efforts towards a sustainable future [49].

Democratisation of chemical automation As previously discussed, different approaches towards chemical automation coexist. Choices are to be made for groups upgrading from a common lab environment. Ideally, an off-the-shelf solution should be available that is compatible with any platform to lower the entry barrier. Therefore, interoperability is key towards the democratisation of chemical automation.

By design, the knowledge graph approach is able to connect to any laboratory. As it is based on ontologies abstracted from the laboratory entities, it is possible to instantiate a new lab into the knowledge graph and utilise the framework. Developing such a usable and reusable ontology is an iterative process and requires the consensus of the domain. It is envisioned to be a community effort in developing and maintaining its life-cycle. As demonstrated by the general semantic web community [62], and particular application experience in the chemical engineering community (OntoCAPE [99]), trial-and-error will be inevitable in the coming decade. However, it is reasonable to be positive given the successful adoption of these technologies by giant IT companies [108]. In that regard, 'The World Avatar' is an open project with all resources available on Github and welcomes contributions from the community.

Role of human researchers Despite the advantage of chemical automation, there has been scepticism that the automation of chemistry will replace the bench chemist [9]. In our view, the transition from manual operations to a fully digitalised and automated laboratory liberates the researchers from labour-intensive work. They can spend more time on more creative things, without worrying about the exact physical steps required to achieve their goals. This is similar to how the computer changed our way of working and increased productivity. Since the data in the knowledge graph is easy to query, researchers can focus on interpreting the experimental data and finding insights in historical knowledge generated from mankind [105]. There exists an opportunity for researchers to encode their chemistry intuition into the knowledge graph, essentially making a 'digital twin' of themselves. It would be possible for researchers from different laboratories to exchange views and establish collaborations previously unfeasible. It would be interesting to see what human intuition can achieve when empowered by greater computing abilities.

Moreover, the linked nature of semantic web technologies can bring us further to smart factories, smart buildings, and smart grids [128], as has already been demonstrated by the application of 'The World Avatar' in smart city planning [15], and the UK national grid [1]. By constructing a digital laboratory and linking it to the wider context, we believe it will facilitate multi-scale and cross-domain interactions between scientists, engineers, and policymakers to investigate how research done in the lab would affect the whole world. Equipped with scenario analysis, this will help to identify the direction science advances.

5 Conclusions and outlook

In this contribution, we proposed a dynamic knowledge-graph-based approach towards chemistry digitalisation. This proposal was motivated by the absence of standardised data representations and communication protocols which precludes further development towards the next generation of self-driving laboratories.

To identify the pain point of the current practices, we performed a thorough review of the data flow between the different functional components within state-of-the-art studies on chemical automation. We found the common platform-based approach employs *ad hoc* data representations and subsequently different data transfer protocols based on their utilities. This results in scalability issues when integrating new hardware and software, and interoperability issues when collaborating among different platforms. It is thus crucial to improve the interoperability within and between these platforms with better data representation and exchange.

We reviewed both semantic and non-semantic efforts in the community and outlined the connections between initiatives. Besides the existence of a pattern in utilising semantic representations to promote better descriptions of chemical knowledge, studies emerging to use agent-based approaches for modularised tasks to enable standardised generation and consumption of data.

With our past experience in closed-loop optimisation and knowledge-graph development, we conjecture that an ontological representation of a laboratory would enable rapid integration of data and AI-based agents for chemical discovery and development. The complete framework will be an all-encompassing dynamic knowledge-graph that faithfully reflects and influences the ever-evolving status of the real world with below value propositions.

- All data are fully connected, consistent and logically coherent, aligning with FAIR principles. For ML, this demonstrates a great opportunity for transfer learning and improving the interpretability of models.
- The existing agent infrastructure makes it easy to automate repetitive tasks with standardised communication protocols. Based on blockchain for assessing the quality of service, it offers a trade-off towards achieving goals in the form of the price and time required to invoke an agent.
- As agent communications are recorded with provenance, any activities done to modify the knowledge graph are tractable, explainable, and reproducible.
- With the further development of intuitive human-data interactions empowered by NLP and reasoning modules, 'The World Avatar' will be able to process and interpret the high-level discovery goals defined by human researchers and execute them automatically.
- 'The World Avatar' links heterogeneous data, software, and hardware interfaces. It thus orchestrates physical experiments with simulation in the cyber-space and enables combining newly acquired data with historical data. With the base toolkit and documentation available to the public, it can be rolled-out at a low cost.

In light of the Industry 4.0 revolution, as well as the current COVID situation, this perspective combines the review of common practices in data representation/exchange, community landscape in the development of better data for reaction informatics, also an outlook towards the holistic integration of automation, AI, and chemistry. The topic of this perspective is timely and we believe it will start thought-provoking conversations over our way towards fully digitalised chemistry as a community.

Moreover, following the knowledge graph approach, hopefully in the not too distant future, we will see the realisation of a complete digital laboratory. Consumables and chemicals will be monitored and automatically ordered. Human researchers will be visualised in the lab with constantly updated health and safety status. The laboratory itself would be connected to the building management system, urban planning, even the national grid. We envisage it would allow more interdisciplinary studies to be conducted for a better understanding of the research activities of mankind. With such further advancements to knowledge graph technology, we are looking forward to a sustainable future in the commencing decade.

Acknowledgements

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme, and Pharma Innovation Platform Singapore (PIPS) via grant to CARES Ltd "Data2Knowledge, C12". The authors are grateful to EPSRC (grant number: EP/R029369/1) and ARCHER for financial and computational support as a part of their funding to the UK Consortium on Turbulent Reacting Flows (www.ukctrf.com). This work was co-funded by EPSRC (grant number: EP/R009902/1) "Combining Chemical Robotics and Statistical Methods to Discover Complex Functional Products". The authors thank Dr Andrew C. Breeson for his help with proofreading. The authors thank Yiqun Bian and Guanhua Li for their help with preparing the graphical abstract, which was designed using resources designed by macrovector/Freepik.com. J. Bai acknowledges financial support provided by CSC Cambridge International Scholarship from Cambridge Trust and China Scholarship Council. M. Kraft gratefully acknowledges the support of the Alexander von Humboldt Foundation.

A Supporting Information

Table 1: *Functional component realisation of selected state-of-the-art studies in chemical automation. For computational model development applications, the model generated as planner was trained on executor with user-defined optimisers. The executors are physical hardware unless otherwise stated. HPC: high-performance computing. MS: mass spectrometer. IR: infrared spectroscopy. BPR: back pressure regulator. HPLC: high-performance liquid chromatography. NMD-M3: Nanotechnology Materials Data Mining, Modeling & Management. VASP: Vienna ab initio Simulation Package. ASE: Atomic Simulation Environment. ICSD: Inorganic Crystal Structure Database.*

Reference	Application	System	Receptionist	Coordinator	Librarian	Planner	Executor
Fitzpatrick et al. [37]	Reaction condition optimisation of a three-dimensional heterogeneous catalytic reaction and a five-dimensional Appel reaction	Flow reactor	Web browser with credentials	LeyLab: PHP-based	Database	Complex Method [71]	Vapourtec R2/R4, online MS, webcam, and inline IR
Ingham et al. [67]	Multi-step synthesis of 2-aminoadamantane-2-carboxylic acid	Flow reactor	Web browser	Octopus: Python-based	CSV file	N/A	Uniqsis, Vapourtec R2+, Knauer machine (HPLC pump), and webcams
Fitzpatrick et al. [38]	Reaction condition optimisation for three active pharmaceutical ingredients (APIs)	Flow reactor	Web browser	LeyLab: PHP-based	Database: MySQL	Complex Method [71]	Flow reactor, HPLC, inline IR, BPR, computer vision module, *etc.*
Nikolaev et al. [107]	Automated designing, executing, and evaluating carbon nanotube growth experiments	Batch reactor	NMD-M3 software	NMD-M3 software	Database within NMD-M3 software	Random forest and genetic algorithm	In-house built ARES platform: inverted Raman microscope, laser, pressure gauge, vaccum pump and gas mass glow controller
Wigley et al. [154]	Process condition optimisation of the production of Bose-Einstein condensates (BEC)	Exponential evaporation ramp	N/A	Python code	MAT file	MLOO [154]: Gaussian process-based	Cooling ramp, evaporation process
Greenaway et al. [52]	Automated porous organic cages discovery by high-throughput screening	Batch reactor	Chemspeed software	Python code	CSV from Chemspeed	Predefined algorithmic workflow in Fig. 6 in Greenaway et al. [52]	**Computational:** HPC; **Physical:** Chemspeed Accelerator SLT-100 automated synthesis platform, NMR, HRMS, HPLC, *etc.*
Bédard et al. [6]	High-yielding implementations of C-C and C-N cross-coupling, olefination, reductive amination, nucleophilic aromatic substitution (S_NAr), photoredox catalysis, and a multistep sequence	Flow reactor	LabVIEW-based GUI	Matlab code	MAT file	SNOBFIT algorithm	Reactor, pump, pressure sensor, flow meter, phase sensor, IR based temperature sensor, camera, HPLC, IR, raman spectroscopy, MS

Table 1: *(Continued)*

Reference	Application	System	Receptionist	Coordinator	Librarian	Planner	Executor
Caramelli et al. [14]	Collaborative chemical space exploration by two internet-connected robots on multiple chemical processes	Batch reactor	N/A	Python code	Twitter feed	Random search, grid search, and Monte-Carlo	Peristaltic pumps, webcam, and reaction flask
Skilton et al. [137]	Etherification of n-propanol in supercritical CO_2 over a γ-Al_2O_3 catalyst, optimized for the formation of di-n-propyl ether	Flow reactor	GLC Solutions	GLC Solutions	Stored within GLC Solutions	GLC Solutions software	Require local human operator for filling stock
Coley et al. [23]	Automated synthesis of 15 medicinal relevant small molecules	Flow reactor	Web browser (for ASKCOS) and Python-based GUI (for robotic execution)	Human researcher	Database (USPTO and Reaxys)	ASKCOS package askcos.mit.edu	Modularised flow reactor, six axis UR3 Universal Robot, BPR, MS, HPLC, NMR, *etc.*
Roch et al. [122]	Automated chemical recipe discovery	Batch reactor	NLP-based chatbot (Twitter, Slack, Gmail)	ChemOS: Python-based	SQLite database	Random search, Spearmint, SMAC, and Phoenics [122]	Various lab equipment based on user needs: pumps, HPLC, *etc.*
Montoya et al. [98]	End-to-end computational system for autonomous materials discovery	Binary/ternary inorganic chemicals	N/A	Python code	JSON	Algorithms in Fig. 2 in Montoya et al. [98]	**Computational:** DFT simulation on AWS EC2
Li et al. [83]	Discovery of optically active chiral inorganic perovskite nanocrystals	Flow reactor	Web browser with SSH login and Python-based GUI	MAOSIC: Python-based	DBMS, *e.g.* MySQL	SNOBFIT	Microfludic reactor, collaborative robot, syringe pump, environment sensor, *etc.*
Xue et al. [159]	Materials discovery for very low thermal hysteresis (ΔT) multicomponent NiTi-based shape memory alloys	Batch reactor	N/A	Python code	CSV (based on supplementary information)	Efficient global optimisation and knowledge gradient	**Computational:** QUANTUM ESPRESSO planewave pseudopotential package; **Physical:** synthesis performed manually
Cao et al. [12]	Formulated product recipe optimisation	Batch reactor	N/A	Python code	CSV and textfile	TSEMO [8]	Robotic platform, syringe pumps, pH analyser, turbidity analyser, and viscometer
Jeraal et al. [70]	Multi-objective reaction condition optimisation for aldol condensation reaction	Flow reactor	Matlab-based GUI	Matlab application	CSV file	TSEMO [8]	Vapourtec R2/R4, Agilent 1260 HPLC, and BPR

Table 1: *(Continued)*

Reference	Application	System	Receptionist	Coordinator	Librarian	Planner	Executor
King et al. [74]	Identification of genes encoding orphan enzymes in yeast *Saccharomyces cerevisiae*	Mobile robotic platform	N/A	Adam: robot scientist	KEGG (prior), MySQL database (new)	Bioinformatic methods (hypotheses generation); two-factor DoE	Freezer, liquid handler, incubators, *etc.*
King et al. [73]	Functional genomics synthesis optimisation for aromatic amino acid synthesis pathway in yeast	Mobile robotic platform	N/A	Laboratory Information Management System	Database within LIMS	Bayesian analysis of decision-tree learning	Liquid handling, pipetting and mixing liquids on microtitre plates
Ingham et al. [66]	Multi-step reaction condition optimisation for pyrazine-2-carboxamide and piperazine-2-carboxamide	Flow reactor	Web browser	Octopus: Python-based	CSV file	Complex Method [71]	Vapourtec R2+/R4, HPLC, computer vision module, *etc.*
Schweidtmann et al. [133]	Multi-objective reaction condition optimisation for S_NAr reaction and N-benzylation	Flow reactor	N/A	Matlab code	CSV file	TSEMO [8]	JASCO PU980 pumps, Vapourtec, Agilent 1100 HPLC, and BPR
HamediRad et al. [54]	Biosynthetic pathway optimisation of lycopene	Continuous workflow	N/A	BioAutomata: Python-based	DAT file	Bayesian optimisation	iBioFAB automated platform [16]
MacLeod et al. [88]	Self-driving laboratory for accelerated discovery of thin-film materials	Mobile robotic platform	Python-based GUI	ChemOS: Python-based	Database	Phoenics within ChemOS	North Robotics N9 robots and liquid handler
Segler et al. [135]	Computational model development for synthesis planning of small molecules	Single-step reactions	N/A	Python code	Extracted from ZINC and Reaxys	Neural network combined with Monte Carlo tree search	**Computational:** single NVIDIA K80 graphics processing unit
Steiner et al. [138]	Automated synthesis of three pharmaceutical compounds: diphenhydramine hydrochloride, rufinamide, and sildenafil	Batch reactor	N/A	Chemputer: Python-based	N/A	Synthesis route set by human	Reactor, pumps, filter, separator, and rotary evaporator
Mehr et al. [93]	Automated syntheses of 12 compounds from the literature, including the analgesic lidocaine, the Dess-Martin periodinane oxidation reagent, and the fluorinating agent AlkylFluor	Batch reactor	Web browser (ChemIDE), allow editing proposed synthesis steps	Chemputer: Python-based	Literature operation procedure in textfile (free-text format)	SynthReader: NLP-based synthesis action sequence planner	Reactor, pumps, filter, separator, rotary evaporator, vacuum, stirrer, conductivity sensor, heater *etc.*
Kusne et al. [78]	Automated phase mapping and property optimisation for accelerating materials discovery with high-throughput X-ray diffraction	Ge–Sb–Te ternary system	GUI (details not provided)	CAMEO: Matlab-based	Preloaded database (ICSD and AFLOW.org)	Bayesian-based active learning method	Not implemented yet

Table 1: *(Continued)*

Reference	Application	System	Receptionist	Coordinator	Librarian	Planner	Executor
Burger et al. [11]	Photocatalysts material discovery for hydrogen production from water	Mobile robotic chemist	Java-based GUI	Java-based	CSV file	Bayesian optimiser [64]	KUKA mobile robot to conduct experiments workflow in the lab, *e.g.*, gas chromatograph measurements, solid dispensing, *etc.*
Tran and Ulissi [145]	Computational screening for electrocatalysts discovery of CO_2 reduction and H_2 evolution	Electrochemical reduction	N/A	Python code	MongoDB database	ML method in TPOT package [110]	**Computational:** High-throughput DFT by VASP using ASE on HPC
Christensen et al. [19]	Autonomous process optimisation of a palladium-catalysed stereoselective Suzuki-Miyaura coupling	Batch reactor	N/A	ChemOS: Python-based	Database	Phoenics and Gryffin within ChemOS	Chemspeed SWING robotic system, Agilent 1100
Gao et al. [43]	Computational model development for optimal reaction condition recommendation of organic synthesis reactions	Single-product and single-step reactions	N/A	Python code	Reaxys database	Nerual Network based condiction prediction tool	**Computational:** single NVIDIA GeForce GTX 1080 GPU
Rosen et al. [123]	Accelerating chemical space exploration of metal–organic frameworks with quantum-chemical calculations and machine learning	Crystalline solids	N/A	Python code	Database	Crystal graph convolutional neural network (CGCNN)	**Computational:** high-throughput DFT by VASP using ASE on HPC
Taylor et al. [141]	Automated determination of reaction models and kinetic parameters	Flow reactor	Matlab-based GUI	Matlab code	Reaction model database in Matlab	MILP optimising reaction kinetics	Tubular reaction vessel built in-house, HPLC pumps, auto-sampler, and HPLC
Waldron et al. [152]	Rapid kinetic model identification	Flow reactor	LabVIEW-based GUI	Python code	CSV file	MBDoE algorithm	Flow reactor, pumps, sampler-dilutor, and HPLC

Table 2: *Data flow and communication protocols between functional components of the selected state-of-the-art studies in chemical automation. The workflow indicates the data flow exchanged within the platform that managed by the coordinator. EP: executor (physical). EC: executor (computational). MS: mass spectrometer. IR: infrared spectroscopy. NMD-M3: Nanotechnology Materials Data Mining, Modeling & Management. HPC: high-performance computing. CRF: chemical recipe file. XDL: chemical description language. VASP: Vienna ab initio Simulation Package. ASE: Atomic Simulation Environment.*

Reference	Receptionist - Coordinator	Coordinator - Librarian	Coordinator - Planner	Coordinator - Executor	Inner Executor (physical)	Inner Executor (computational)	Workflow (R C L P EP EC)
Fitzpatrick et al. [37]	TCP/IP	MySQL database query (TCP/IP)	In-memory cache	TCP-IP: MS readings transmitted by Arduino-based analogue to serial converter (RS232 serial communication to Ethernet); webcam liquid level position computed by Raspberry Pi in JSON format	Arduino and Raspberry Pi worked as interface for controlling the equipment	N/A	C-[R-P-L-EP-P]
Ingham et al. [67]	HTTP	TCP/IP	Python variables	(1) USB root hub: two webcams; (2) TCP/IP: Uniqsis, Knauer machines and Vapourtec were connected via Brainboxes ES-701 and ES-257 ethernet-to-serial adapters	Raspberry Pi worked as interface for controlling the equipment	N/A	C-[R-P-L-EP-P]
Fitzpatrick et al. [38]	HTTPS	Database query (TCP/IP)	In-memory cache	TCP/IP (with RS232 serial to Ethernet adaptor): equipment was placed within a VLAN that connected to LeyVM server via SSH tunnel	Raspberry Pi worked as interface for controlling the equipment	N/A	C-[R-P-EP-L-P]
Nikolaev et al. [107]	Handled by NMD-M3 software	Handled by NMD-M3 software	Handled by NMD-M3 software	Handled by NMD-M3 software	Through in house built software in C#/.NET	N/A	C-[R-P-EP-L-P]
Wigley et al. [154]	N/A	File transfer	Python variables	File transfer: MAT and textfile	Not specified	N/A	C-[L-P-EP-L]
Greenaway et al. [52]	Within Chemspeed software	File transfer	SSH to HPC	Set the input variables through Chemspeed software	Controlled through Chemspeed software	N/A	C-[EC-L-R-P-EP]
Bédard et al. [6]	Matlab variables	Matlab varialbes	Matlab variable	LabVIEW: through national instrument, serial modem cable, USB cable	Serial command through LabVIEW	N/A	R-P-EP-L-P

Table 2: *(Continued)*

Reference	Receptionist - Coordinator	Coordinator - Librarian	Coordinator - Planner	Coordinator - Executor	Inner Executor (physical)	Inner Executor (computational)	Workflow (R C L P EP EC)
Caramelli et al. [14]	N/A	Plaintext as Python variables	Python variables	Python variables handled by gpio and opencv Python libraries	Pumps and webcam are interfaced via pcDuino3 board	N/A	C-[L-P-EP-L]
Skilton et al. [137]	Passed within GLC Solutions	Passed within GLC Solutions	Passed within GLC Solutions	Remote computer control through GLC Solutions	Handled by GLC Solutions	N/A	C-[L-P-E-L]
Coley et al. [23]	CRF file transfer by human researcher	MongoDB database query	Planner generates CRF file to be modified by human researcher	Human researcher pass the CRF file to robotic execution GUI	Universal process bays provide sealing and alignment mechanisms for the fluidic, electrical, and pneumatic process connections	SLURM scheduling software	C-[R(of EC)-L-P(resulted from EC)-R(of EP)-EP]
Roch et al. [122]	JSON, parsed by Python	Database query	Python array	Python pickle object	Raspberry Pi as controller of pumping system, communicated via SCP with the executor codes; Dropbox for synchronising the characterisation equipment	N/A	C-[R-L-P-EP-L]
Montoya et al. [98]	N/A	Python variables	Python variables	Python variables	N/A	AWS Batch API	C-[L-P-EC-L]
Li et al. [83]	TLS encrypted file transfer	SQLAlchemy database query	Python variables (within MAOSIC)	JSON-RPC	Interfaced via high-level and low-level instructions based on JSON-RPC2.0 protocol	N/A	C-[R-L-P-EP-L]
Xue et al. [159]	N/A	Not specified	Python variables	File transfer for DFT	Synthesis performed manually	Handled by Quantum ESPRESSO package	C-[L-P-EC-L]
Cao et al. [12]	N/A	File transfer	File transfer	File transfer	File transfer	N/A	C-[L-P-EP-L]
Jeraal et al. [70]	In-memory cache of Matlab variables	File transfer	Matlab variables	File transfer: CSV file	Interfaced via FlowCommander, a software provided by Vapourtec	N/A	C-[R-L-P-EP-L]
King et al. [74]	N/A	Database query	Not specified	File transfer: LABORS (OWL-DL format)	Closed-source software from Caliper Life Sciences	N/A	C-[L-P-EP-L]
King et al. [73]	N/A	Database query	Not specified	Prolog commands through TCP/IP	Robot operations controlled by tool command language translated from Prolog commands	N/A	C-[L-P-EP-L]

Table 2: *(Continued)*

Reference	Receptionist - Coordinator	Coordinator - Librarian	Coordinator - Planner	Coordinator - Executor	Inner Executor (physical)	Inner Executor (computational)	Workflow (R C L P EP EC)
Ingham et al. [66]	HTTP	TCP/IP	Python variables	(1) USB root hub: two webcams; (2) TCP/IP: Uniqsis, Knauer machines and Vapourtec were connected via Brainboxes ES-701 and ES-257 ethernet-to-serial adapters	Raspberry Pi worked as interface for controlling the equipments	N/A	C-[L-P-EP-L]
Schweidtmann et al. [133]	N/A	File transfer	File transfer: CSV file	File transfer: CSV file	Interfaced via FlowCommander, a software developed by Vapourtec	N/A	C-[EP-L-P-EP-L]
HamediRad et al. [54]	N/A	File transfer	Python variables	File transfer: CSV file (iScheduler code)	Managed by iScheduler scheduling software on iBioFAB platform	N/A	C-[L-P-EP-L]
MacLeod et al. [88]	Python variables	Database query	Python variables	Python variables	Driven by North Robotics C9 controller, which also provides auxiliary controls for third-party instruments and components used by the robot	N/A	C-[R-L-P-EP-L]
Segler et al. [135]	N/A	Python variables	Python variables	Python variables	N/A	Theano-backend Keras	C-[L-P-EC]
Steiner et al. [138]	N/A	N/A	XDL (XML-based file) for synthesis route	XDL file	Arduino as micro-controller	N/A	C-[P-EP]
Mehr et al. [93]	In-memory cache & XDL	Textfile	XDL file	XDL file	Arduino as micro-controller	N/A	C-[R-L-R-EP]
Kusne et al. [78]	Not specified	File transfer: MAT file	Matlab variables	Programmatically generated script via SPEC for the SLAC high-throughput system or a GADDS script for the Bruker system	Not specified	N/A	C-[L-P-EP-L]
Burger et al. [11]	N/A	CSV read/write	File transfer	Various communication protocols (TCP/IP over WIFI/LAN; RS-232)	Simultaneous localisation and mapping (SLAM) was used for robot allocation; Arduino designed as micro-controller	N/A	C-[L-P-EP-L]
Tran and Ulissi [145]	N/A	Database query	Python variables	Atom object converted from JSON	N/A	Managed by Luigi and Fireworks	C-[L-P-EC-L]

Table 2: *(Continued)*

Reference	Receptionist - Coordinator	Coordinator - Librarian	Coordinator - Planner	Coordinator - Executor	Inner Executor (physical)	Inner Executor (computational)	Workflow (R C L P EP EC)
Christensen et al. [19]	N/A	Database query	Python variables	File transfer (done by a Python script): CSV, textfile & Python pickle object	Chemspeed AutoSuite acts as the control interface	N/A	C-[R-L-P-EP-L]
Gao et al. [43]	N/A	Reaxys API	Python variables	Python variables	N/A	Theano-backend Keras	C-[L-P-EC-L]
Rosen et al. [123]	N/A	Python variables	Python variables	File transfer (ASE argument to VASP)	N/A	Managed by PyMOFScreen	C-[L-P-EC-L]
Taylor et al. [141]	Matlab variables	Matlab variables	Matlab variables	Matlab variables	Matlab variables	N/A	C-[L-P-EP-L]
Waldron et al. [152]	Python variables	File transfer	Python variables	Python variables	LabVIEW acts as the interface	N/A	C-[R-L-P-EP-L]

References

[1] J. Akroyd, S. Mosbach, A. Bhave, and M. Kraft. Universal Digital Twin – A Dynamic Knowledge Graph. *Data-Centric Engineering*, 2:e14, 2021. doi:10.1017/dce.2021.10.

[2] AnIML Working Group. AnIML: Overview. URL https://www.animl.org/overview. Accessed 31 July 2021.

[3] J. Bai, R. Geeson, F. Farazi, S. Mosbach, J. Akroyd, E. J. Bringley, and M. Kraft. Automated Calibration of a Poly(oxymethylene) Dimethyl Ether Oxidation Mechanism Using the Knowledge Graph Technology. *J. Chem. Inf. Model.*, 61(4):1701–1717, 2021. doi:10.1021/acs.jcim.0c01322.

[4] J. B. L. Bard and S. Y. Rhee. Ontologies in Biology: Design, Applications and Future Challenges. *Nat. Rev. Genet.*, 5(3):213–222, 2004. doi:10.1038/nrg1295.

[5] C. Batchelor and P. Corbett. Semantic Enrichment of Journal Articles Using Chemical Named Entity Recognition. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 45–48, 2007. URL https://aclanthology.org/P07-2012. Accessed 31 July 2021.

[6] A.-C. Bédard, A. Adamo, K. C. Aroh, M. G. Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen, and T. F. Jamison. Reconfigurable System for Automated Optimization of Diverse Chemical Reactions. *Science*, 361(6408):1220–1225, 2018. doi:10.1126/science.aat0650.

[7] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Sci. Am.*, 284(5):34–43, 2001. doi:10.1038/scientificamerican0501-34. URL https://www.jstor.org/stable/10.2307/26059207.

[8] E. Bradford, A. M. Schweidtmann, and A. Lapkin. Efficient Multiobjective Optimization Employing Gaussian Processes, Spectral Sampling and A Genetic Algorithm. *J. Glob. Optim.*, 71(2):407–438, 2018. doi:10.1007/s10898-018-0609-2.

[9] R. Brazil. Automation in the Chemistry Lab., 2021. URL https://www.chemistryworld.com/careers/automation-in-the-chemistry-lab/4012832.article. Accessed 31 July 2021.

[10] C. P. Breen, A. M. K. Nambiar, T. F. Jamison, and K. F. Jensen. Ready, Set, Flow! Automated Continuous Synthesis and Optimization. *Trends Chem.*, 2021. doi:10.1016/j.trechm.2021.02.005.

[11] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, and A. I. Cooper. A Mobile Robotic Chemist. *Nature*, 583(7815):237–241, 2020. doi:10.1038/s41586-020-2442-2.

[12] L. Cao, D. Russo, K. Felton, D. Salley, A. Sharma, G. Keenan, W. Mauer, H. Gao, L. Cronin, and A. A. Lapkin. Optimization of Formulations Using Robotic Experiments Driven by Machine Learning DoE. *Cell Rep. Phys. Sci.*, 2(1):100295, 2021. doi:10.1016/j.xcrp.2020.100295.

[13] L. Cao, D. Russo, and A. A. Lapkin. Automated Robotic Platforms in Design and Development of Formulations. *AIChE J.*, page e17248, 2021. doi:10.1002/aic.17248.

[14] D. Caramelli, D. Salley, A. Henson, G. A. Camarasa, S. Sharabi, G. Keenan, and L. Cronin. Networking Chemical Robots for Reaction Multitasking. *Nat. Commun.*, 9(1):1–10, 2018. doi:10.1038/s41467-018-05828-8.

[15] A. Chadzynski, N. Krdzavac, F. Farazi, M. Q. Lim, S. Li, A. Grisiute, P. Herthogs, A. von Richthofen, S. Cairns, and M. Kraft. Semantic 3D City Database - An Enabler for a Dynamic Geospatial Knowledge Graph. *Energy and AI*, 6:100106, 2021. doi:10.1016/j.egyai.2021.100106.

[16] R. Chao, Y. Yuan, and H. Zhao. Building Biological Foundries for Next-Generation Synthetic Biology. *Sci. China: Life Sci.*, 58(7):658–665, 2015. doi:10.1007/s11427-015-4866-8.

[17] S. Chatterjee, M. Guidi, P. H. Seeberger, and K. Gilmore. Automated Radial Synthesis of Organic Molecules. *Nature*, 579(7799):379–384, 2020. doi:10.1038/s41586-020-2083-5.

[18] Chemical Semantics. GNVC: Gainesville Core Ontology - Standard for Publishing Results of Computational Chemistry, 2015. URL `http://ontologies.makolab.com/gc/gc07.owl`. Accessed 21 September 2021.

[19] M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, and J. Hein. Data-Science Driven Autonomous Process Optimization. *Commun. Chem.*, 4(1):1–12, 2021. doi:10.1038/s42004-021-00550-x.

[20] A. D. Clayton, J. A. Manson, C. J. Taylor, T. W. Chamberlain, B. A. Taylor, G. Clemens, and R. A. Bourne. Algorithms for the Self-Optimisation of Chemical Reactions. *React. Chem. Eng.*, 4(9):1545–1554, 2019. doi:10.1039/C9RE00209J.

[21] S. J. Coles, N. E. Day, P. Murray-Rust, H. S. Rzepa, and Y. Zhang. Enhancement of the Chemical Semantic Web through the Use of InChI Identifiers. *Org. Biomol. Chem.*, 3(10):1832–1834, 2005. doi:10.1039/B502828K.

[22] C. W. Coley. Defining and Exploring Chemical Spaces. *Trends Chem.*, 3(2):133–145, 2021. doi:10.1016/j.trechm.2020.11.004.

[23] C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison, and K. F. Jensen. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science*, 365(6453):eaax1566, 2019. doi:10.1126/science.aax1566.

[24] C. W. Coley, N. S. Eyke, and K. F. Jensen. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem., Int. Ed.*, 59(51):22858–22893, 2020. doi:10.1002/anie.201909987.

[25] C. W. Coley, N. S. Eyke, and K. F. Jensen. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem., Int. Ed.*, 59(52):23414–23436, 2020. doi:10.1002/anie.201909989.

[26] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.*, 12(3):191–201, 2013. doi:10.1038/nmat3568.

[27] Daylight. SMARTS - A Language for Describing Molecular Patterns, 2014. URL https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed 27 May 2021.

[28] Daylight. SMIRKS - A Reaction Transform Language, 2014. URL https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html. Accessed 27 May 2021.

[29] T. Dimitrov, C. Kreisbeck, J. S. Becker, A. Aspuru-Guzik, and S. K. Saikin. Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces*, 11(28): 24825–24836, 2019. doi:10.1021/acsami.9b01226.

[30] A. Eibeck, M. Q. Lim, and M. Kraft. J-Park Simulator: An Ontology-Based Platform for Cross-domain Scenarios in Process Industry. *Comput. Chem. Eng.*, 131: 106586, 2019. doi:10.1016/j.compchemeng.2019.106586.

[31] EMBL-EBI. Molecular Process Ontology, 2014. URL https://www.ebi.ac.uk/ols/ontologies/mop. Accessed 14 June 2021.

[32] EMBL-EBI. Chemical Methods Ontology, 2019. URL https://www.ebi.ac.uk/ols/ontologies/chmo. Accessed 14 June 2021.

[33] EMBL-EBI. Name Reaction Ontology, 2021. URL https://www.ebi.ac.uk/ols/ontologies/rxno. Accessed 14 June 2021.

[34] F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca, and M. Kraft. OntoKin: An Ontology for Chemical Kinetic Reaction Mechanisms. *J. Chem. Inf. Model.*, 60(1):108–120, 2020. doi:10.1021/acs.jcim.9b00960.

[35] F. Farazi, N. B. Krdzavac, J. Akroyd, S. Mosbach, A. Menon, D. Nurkowski, and M. Kraft. Linking Reaction Mechanisms and Quantum Chemistry: An Ontological Approach. *Comput. Chem. Eng.*, 137:106813, 2020. doi:10.1016/j.compchemeng.2020.106813.

[36] D. E. Fitzpatrick and S. V. Ley. Engineering Chemistry for the Future of Chemical Synthesis. *Tetrahedron*, 74(25):3087–3100, 2018. doi:10.1016/j.tet.2017.08.050.

[37] D. E. Fitzpatrick, C. Battilocchio, and S. V. Ley. A Novel Internet-Based Reaction Monitoring, Control and Autonomous Self-Optimization Platform for Chemical Synthesis. *Org. Process Res. Dev.*, 20(2):386–394, 2016. doi:10.1021/acs.oprd.5b00313.

[38] D. E. Fitzpatrick, T. Maujean, A. C. Evans, and S. V. Ley. Across-the-World Automated Optimization and Continuous-Flow Synthesis of Pharmaceutical Agents Operating through a Cloud-Based Server. *Angew. Chem., Int. Ed.*, 57(46):15128–15132, 2018. doi:10.1002/anie.201809080.

[39] D. E. Fitzpatrick, M. O'Brien, and S. V. Ley. A Tutored Discourse on Microcontrollers, Single Board Computers and Their Applications to Monitor and Control Chemical Reactions. *React. Chem. Eng.*, 5(2):201–220, 2020. doi:10.1039/C9RE00407F.

[40] M. M. Flores-Leonar, L. M. Mejía-Mendoza, A. Aguilar-Granda, B. Sanchez-Lengeling, H. Tribukait, C. Amador-Bedolla, and A. Aspuru-Guzik. Materials Acceleration Platforms: On the Way to Autonomous Experimentation. *Curr. Opin. Green Sustain. Chem.*, page 100370, 2020. doi:10.1016/j.cogsc.2020.100370.

[41] G. Fu, C. Batchelor, M. Dumontier, J. Hastings, E. Willighagen, and E. Bolton. PubChemRDF: Towards the Semantic Annotation of PubChem Compound and Substance Databases. *J. Cheminf.*, 7(1):1–15, 2015. doi:10.1186/s13321-015-0084-4.

[42] J. Galgonek and J. Vondrášek. IDSM ChemWebRDF: SPARQLing Small-Molecule Datasets. *J. Cheminf.*, 13(1):1–19, 2021. doi:10.1186/s13321-021-00515-1.

[43] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen. Using Machine Learning to Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.*, 4(11):1465–1476, 2018. doi:10.1021/acscentsci.8b00357.

[44] S. S. Garud, I. A. Karimi, and M. Kraft. Design of Computer Experiments: A Review. *Comput. Chem. Eng.*, 106:71–95, 2017. doi:10.1016/j.compchemeng.2017.05.010.

[45] K. Gilmore, D. Kopetzki, J. W. Lee, Z. Horváth, D. T. McQuade, A. Seidel-Morgenstern, and P. H. Seeberger. Continuous Synthesis of Artemisinin-Derived Medicines. *Chem. Commun.*, 50(84):12652–12655, 2014. doi:10.1039/C4CC05098C.

[46] G. V. Gkoutos, P. Murray-Rust, H. S. Rzepa, and M. Wright. Chemical Markup, XML, and the World-Wide Web. 3. Toward a Signed Semantic Chemical Web of Trust. *J. Chem. Inf. Comput. Sci.*, 41(5):1124–1130, 2001. doi:10.1021/ci000406v.

[47] A. G. Godfrey, T. Masquelin, and H. Hemmerle. A Remote-Controlled Adaptive Medchem Lab: An Innovative Approach to Enable Drug Discovery in the 21st Century. *Drug Discovery Today*, 18(17-18):795–802, 2013. doi:10.1016/j.drudis.2013.03.001.

[48] A. G. Godfrey, S. G. Michael, G. S. Sittampalam, and G. Zahoránszky-Köhalmi. A Perspective on Innovating the Chemistry Lab Bench. *Front. Rob. AI*, 7:24, 2020. doi:10.3389/frobt.2020.00024.

[49] C. Gomes, T. Dietterich, C. Barrett, J. Conrad, B. Dilkina, S. Ermon, F. Fang, A. Farnsworth, A. Fern, X. Fern, D. Fink, D. Fisher, A. Flecker, D. Freund, A. Fuller, J. Gregoire, J. Hopcroft, S. Kelling, Z. Kolter, W. Powell, N. Sintov, J. Selker, B. Selman, D. Sheldon, D. Shmoys, M. Tambe, W.-K. Wong, C. Wood, X. Wu, Y. Xue, A. Yadav, A.-A. Yakubu, and M. L. Zeeman. Computational Sustainability: Computing for a Better World and a Sustainable Future. *Commun. ACM*, 62(9):56–65, 2019. doi:10.1145/3339399.

[50] C. P. Gomes, J. Bai, Y. Xue, J. Björck, B. Rappazzo, S. Ament, R. Bernstein, S. Kong, S. K. Suram, R. B. van Dover, and J. M. Gregoire. CRYSTAL: A Multi-Agent AI System for Automated Mapping of Materials' Crystal Structures. *MRS Commun.*, 9(2):600–608, 2019. doi:10.1557/mrc.2019.50.

[51] J. Goodman. Computer Software Review: Reaxys. *J. Chem. Inf. Model.*, 49(12): 2897–2898, 2009. doi:10.1021/ci900437n.

[52] R. L. Greenaway, V. Santolini, M. J. Bennison, B. M. Alston, C. J. Pugh, M. A. Little, M. Miklitz, E. G. B. Eden-Rump, R. Clowes, A. Shakil, H. J. Cuthbertson, H. Armstrong, M. E. Briggs, K. E. Jelfs, and A. I. Cooper. High-Throughput Discovery of Organic Cages and Catenanes Using Computational Screening Fused with Robotic Synthesis. *Nat. Commun.*, 9(1):1–11, 2018. doi:10.1038/s41467-018-05271-9.

[53] G. Grethe, G. Blanke, H. Kraut, and J. M. Goodman. International Chemical Identifier for Reactions (RInChI). *J. Cheminf.*, 10(1):1–9, 2018. doi:10.1186/s13321-018-0277-8.

[54] M. HamediRad, R. Chao, S. Weisberg, J. Lian, S. Sinha, and H. Zhao. Towards a Fully Automated Algorithm Driven Platform for Biosystems Design. *Nat. Commun.*, 10(1):1–10, 2019. doi:10.1038/s41467-019-13189-z.

[55] A. J. S. Hammer, A. I. Leonov, N. L. Bell, and L. Cronin. Chemputation and the Standardization of Chemical Informatics. *JACS Au*, 2021. doi:10.1021/jacsau.1c00303.

[56] F. Häse, L. M. Roch, and A. Aspuru-Guzik. Next-Generation Experimentation with Self-Driving Laboratories. *Trends Chem.*, 1(3):282–291, 2019. doi:10.1016/j.trechm.2019.02.007.

[57] J. Hastings, L. Chepelev, E. Willighagen, N. Adams, C. Steinbeck, and M. Dumontier. The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web. *PloS One*, 6(10):e25513, 2011. doi:10.1371/journal.pone.0025513.

[58] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck. The ChEBI Reference Database and Ontology for Biologically Relevant Chemistry: Enhancements for 2013. *Nucleic Acids Res.*, 41(D1):D456–D463, 2012. doi:10.1093/nar/gks1146.

[59] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.*, 7(1):1–34, 2015. doi:10.1186/s13321-015-0068-4.

[60] J. Hendler. Agents and the Semantic Web. *IEEE Intell. Syst.*, 16(2):30–37, 2001. doi:10.1109/5254.920597.

[61] S. Herres-Pawlis, O. Koepler, and C. Steinbeck. NFDI4Chem: Shaping a Digital and Cultural Change in Chemistry. *Angew. Chem., Int. Ed.*, 58(32):10766–10768, 2019. doi:10.1002/anie.201907260.

[62] P. Hitzler. A Review of The Semantic Web Field. *Commun. ACM*, 64(2):76–83, 2021. doi:10.1145/3397512.

[63] R. Hoogenboom, M. W. M. Fijten, C. Brändli, J. Schroer, and U. S. Schubert. Automated Parallel Temperature Optimization and Determination of Activation Energy for the Living Cationic Polymerization of 2-Ethyl-2-Oxazoline. *Macromol. Rapid Commun.*, 24(1):98–103, 2003. doi:10.1002/marc.200390017.

[64] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Parallel Algorithm Configuration. In *International Conference on Learning and Intelligent Optimization*, pages 55–70. Springer, 2012. doi:10.1007/978-3-642-34413-8_5.

[65] O. Inderwildi, C. Zhang, X. Wang, and M. Kraft. The Impact of Intelligent Cyber-Physical Systems on the Decarbonization of Energy. *Energy Environ. Sci.*, 13(3): 744–771, 2020. doi:10.1039/C9EE01919G.

[66] R. J. Ingham, C. Battilocchio, J. M. Hawkins, and S. V. Ley. Integration of Enabling Methods for the Automated Flow Preparation of Piperazine-2-Carboxamide. *Beilstein J. Org. Chem.*, 10(1):641–652, 2014. doi:10.3762/bjoc.10.56.

[67] R. J. Ingham, C. Battilocchio, D. E. Fitzpatrick, E. Sliwinski, J. M. Hawkins, and S. V. Ley. A Systems Approach Towards an Intelligent and Self-Controlling Platform for Integrated Continuous Reaction Sequences. *Angew. Chem., Int. Ed.*, 127 (1):146–150, 2015. doi:10.1002/anie.201409356.

[68] JADE. Java Agent DEvelopment Framework: Jade Site, 2021. URL `https://jade.tilab.com/`. Accessed 27 May 2021.

[69] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, and K. A. Persson. FireWorks: A Dynamic Workflow System Designed for High-Throughput Applications. *Concurr. Comput. Pract. Exp*, 27(17):5037–5059, 2015. doi:10.1002/cpe.3505.

[70] M. I. Jeraal, S. Sung, and A. A. Lapkin. A Machine Learning-Enabled Autonomous Flow Chemistry Platform for Process Optimization of Multiple Reaction Metrics. *Chem. Methods*, 1(1):71–77, 2021. doi:10.1002/cmtd.202000044.

[71] R. F. Kazmierczak Jr. Optimizing Complex Bioeconomic Simulations Using an Efficient Search Heuristic. 1996. doi:10.2139/ssrn.15071.

[72] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.*, 47(D1):D1102–D1109, 2019. doi:10.1093/nar/gky1033.

[73] R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver. Functional Genomic Hypothesis Generation and Experimentation by a Robot Scientist. *Nature*, 427(6971):247–252, 2004. doi:10.1038/nature02236.

[74] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, and A. Clare. The Automation of Science. *Science*, 324(5923):85–89, 2009. doi:10.1126/science.1165620.

[75] N. J. Knight, S. Kanza, D. Cruickshank, W. S. Brocklesby, and J. G. Frey. Talk2Lab: The Smart Lab of the Future. *IEEE Internet Things J.*, 7(9):8631–8640, 2020. doi:10.1109/JIOT.2020.2995323.

[76] N. Krdzavac, S. Mosbach, D. Nurkowski, P. Buerger, J. Akroyd, J. Martin, A. Menon, and M. Kraft. An Ontology and Semantic Web Service for Quantum Chemistry Calculations. *J. Chem. Inf. Model.*, 59(7):3154–3165, 2019. doi:10.1021/acs.jcim.9b00227.

[77] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn.: Sci. Technol.*, 1(4):045024, 2020. doi:10.1088/2632-2153/aba947.

[78] A. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, A. V. Davydov, R. Agarwal, L. A. Bendersky, M. Li, A. Mehta, and I. Takeuchi. On-the-Fly Closed-Loop Materials Discovery via Bayesian Active Learning. *Nat. Commun.*, 11(1):1–11, 2020. doi:10.1038/s41467-020-19597-w.

[79] P. Lampen, J. Lambert, R. J. Lancashire, R. S. McDonald, P. S. McIntyre, D. N. Rutledge, T. Fröhlich, and A. N. Davies. An Extension to the JCAMP-DX Standard File Format, JCAMP-DX V. 5.01. *Pure Appl. Chem.*, 71(8):1549–1556, 1999. doi:10.1351/pac199971081549.

[80] G. Landrum et al. RDKit: Open-Source Cheminformatics. URL `https://www.rdkit.org/`. Accessed 27 May 2021.

[81] S. Langner, F. Häse, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik, and C. J. Brabec. Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems. *Adv. Mater.*, 32(14):1907801, 2020. doi:10.1002/adma.201907801.

[82] S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, and R. M. Myers. Organic Synthesis: March of the Machines. *Angew. Chem., Int. Ed.*, 54(11):3449–3464, 2015. doi:10.1002/anie.201410744.

[83] J. Li, J. Li, R. Liu, Y. Tu, Y. Li, J. Cheng, T. He, and X. Zhu. Autonomous Discovery of Optically Active Chiral Inorganic Perovskite Nanocrystals through an Intelligent Cloud Lab. *Nat. Commun.*, 11(1):1–10, 2020. doi:10.1038/s41467-020-15728-5.

[84] J. Li, Y. Tu, R. Liu, Y. Lu, and X. Zhu. Toward "On-Demand" Materials Synthesis and Scientific Discovery through Intelligent Robots. *Adv. Sci.*, 7(7):1901957, 2020. doi:10.1002/advs.201901957.

[85] Z. Li, M. A. Najeeb, L. Alves, A. Z. Sherman, V. Shekar, P. Cruz Parrilla, I. M. Pendleton, W. Wang, P. W. Nega, M. Zeller, J. Schrier, A. J. Norquist, and E. M. Chan. Robot-Accelerated Perovskite Investigation and Discovery. *Chem. Mater.*, 32(13):5650–5663, 2020. doi:10.1021/acs.chemmater.0c01153.

[86] J. S. Lindsey. A Retrospective on the Automation of Laboratory Synthetic Chemistry. *Chemom. Intell. Lab. Syst.*, 17(1):15–45, 1992. doi:10.1016/0169-7439(92)90025-B.

[87] D. Lowe. Chemical Reactions from US Patents (1976-Sep2016). 2017. doi:10.6084/m9.figshare.5104873.v1.

[88] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, and C. P. Berlinguette. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials. *Sci. Adv.*, 6(20):eaaz8867, 2020. doi:10.1126/sciadv.aaz8867.

[89] C. Mateos, M. J. Nieves-Remacha, and J. A. Rincón. Automated Platforms for Reaction Self-Optimization in Flow. *React. Chem. Eng.*, 4(9):1536–1544, 2019. doi:10.1039/C9RE00116F.

[90] K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I.-h. Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z. K. Liu, J. Neaton, S. P. Ong, K. Persson, and A. Jain. Atomate: A High-Level Interface to Generate, Execute, and Analyze Computational Materials Science Workflows. *Comput. Mater. Sci.*, 139:140–152, 2017. doi:10.1016/j.commatsci.2017.07.030.

[91] A. McNally, C. K. Prier, and D. W. MacMillan. Discovery of an α-Amino C–H Arylation Reaction Using the Strategy of Accelerated Serendipity. *Science*, 334 (6059):1114–1117, 2011. doi:10.1126/science.1213920.

[92] A. McNally, B. Haffemayer, B. S. L. Collins, and M. J. Gaunt. Palladium-Catalysed C–H Activation of Aliphatic Amines to Give Strained Nitrogen Heterocycles. *Nature*, 510(7503):129–133, 2014. doi:10.1038/nature13389.

[93] S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan, and L. Cronin. A Universal System for Digitization and Automatic Execution of the Chemical Synthesis Literature. *Science*, 370(6512):101–108, 2020. doi:10.1126/science.abc2986.

[94] A. Menon, N. B. Krdzavac, and M. Kraft. From Database to Knowledge Graph—Using Data in Chemistry. *Curr. Opin. Chem. Eng.*, 26:33–37, 2019. doi:10.1016/j.coche.2019.08.004.

[95] R. B. Merrifield, J. M. Stewart, and N. Jernberg. Instrument for Automated Synthesis of Peptides. *Anal. Chem.*, 38(13):1905–1914, 1966. doi:10.1021/ac50155a057.

[96] T. Millecam, A. J. Jarrett, N. Young, D. E. Vanderwall, and D. Della Corte. Coming of Age of Allotrope: Proceedings from the Fall 2020 Allotrope Connect. *Drug Discovery Today*, 2021. doi:10.1016/j.drudis.2021.03.028.

[97] Y. Mo, G. Rughoobur, A. M. K. Nambiar, K. Zhang, and K. F. Jensen. A Multifunctional Microfluidic Platform for High-Throughput Experimentation of Electroorganic Chemistry. *Angew. Chem., Int. Ed.*, 59(47):20890–20894, 2020. doi:10.1002/anie.202009819.

[98] J. H. Montoya, K. T. Winther, R. A. Flores, T. Bligaard, J. S. Hummelshøj, and M. Aykol. Autonomous Intelligent Agents for Accelerated Materials Discovery. *Chem. Sci.*, 11(32):8517–8532, 2020. doi:10.1039/D0SC01101K.

[99] J. Morbach, A. Yang, and W. Marquardt. OntoCAPE - a Large-scale Ontology for Chemical Process Engineering. *Eng. Appl. Artif. Intell.*, 20(2):147–161, 2007. doi:10.1016/j.engappai.2006.06.010.

[100] S. Mosbach, A. Menon, F. Farazi, N. Krdzavac, X. Zhou, J. Akroyd, and M. Kraft. Multiscale Cross-Domain Thermochemical Knowledge-Graph. *J. Chem. Inf. Model.*, 60(12):6155–6166, 2020. doi:10.1021/acs.jcim.0c01145.

[101] P. Murray-Rust. CML - Frequently Asked Questions. URL `http://www.xml-cml.org/documentation/FAQ.html#chemistry`. Accessed 31 July 2021.

[102] P. Murray-Rust. Chemistry for Everyone. *Nature*, 451(7179):648–651, 2008. doi:10.1038/451648a.

[103] P. Murray-Rust, H. S. Rzepa, S. M. Tyrrell, and Y. Zhang. Representation and Use of Chemistry in the Global Electronic Age. *Org. Biomol. Chem.*, 2(22):3192–3203, 2004. doi:10.1039/B410732B.

[104] NextMove Software. Pistachio. URL `https://www.nextmovesoftware.com/pistachio.html`. Accessed 15 July 2021.

[105] M. C. Nicklaus. NIH Virtual Workshop on Ultra-Large Chemistry Databases, Dec 1-3, 2020. URL https://cactus.nci.nih.gov/presentations/NIHBigDB_2020-12/NIHBigDB.html. Accessed 31 July 2021.

[106] M. C. Nicklaus. NIH Virtual Workshop on Reaction Informatics, May 18-20, 2021. URL https://cactus.nci.nih.gov/presentations/NIHReactInf_2021-05/NIHReactInf.html. Accessed 31 July 2021.

[107] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, and B. Maruyama. Autonomy in Materials Research: A Case Study in Carbon Nanotube Growth. *npj Comput. Mater.*, 2(1):1–6, 2016. doi:10.1038/npjcompumats.2016.31.

[108] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. Industry-Scale Knowledge Graphs: Lessons and Challenges. *Commun. ACM*, 62(8):36–43, 2019. doi:10.1145/3331166.

[109] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An Open Chemical Toolbox. *J. Cheminf.*, 3(1):1–14, 2011. doi:10.1186/1758-2946-3-33.

[110] R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd, and J. H. Moore. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. In *European Conference on the Applications of Evolutionary Computation*, pages 123–137. Springer, 2016. doi:10.1007/978-3-319-31204-0_9.

[111] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder. Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.*, 68:314–319, 2013. doi:10.1016/j.commatsci.2012.10.028.

[112] Open Reaction Database Project Authors. Welcome to the Open Reaction Database!, 2021. URL https://docs.open-reaction-database.org/en/latest/. Accessed 27 May 2021.

[113] M. Pan, J. Sikorski, C. A. Kastner, J. Akroyd, S. Mosbach, R. Lau, and M. Kraft. Applying Industry 4.0 to the Jurong Island Eco-industrial Park. *Energy Procedia*, 75:1536–1541, 2015. doi:10.1016/j.egypro.2015.07.313.

[114] I. M. Pendleton, G. Cattabriga, Z. Li, M. A. Najeeb, S. A. Friedler, A. J. Norquist, E. M. Chan, and J. Schrier. Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): A Software Pipeline for Automated Chemical Experimentation and Data Management. *MRS Commun.*, 9(3):846–859, 2019. doi:10.1557/mrc.2019.72.

[115] M. Peplow. Organic Synthesis: The Robo-Chemist. *Nature*, 512(7512):20, 2014. doi:10.1038/512020a.

[116] Pistoia Alliance. Unified Data Model, 2020. URL https://github.com/PistoiaAlliance/UDM. Accessed 27 May 2021.

[117] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature*, 533(7601):73–76, 2016. doi:10.1038/nature17439.

[118] A. Reuther, C. Byun, W. Arcand, D. Bestor, B. Bergeron, M. Hubbell, M. Jones, P. Michaleas, A. Prout, A. Rosa, and J. Kepner. Scalable System Scheduling for HPC and Big Data. *J. Parallel Distrib. Comput.*, 111:76–92, 2018. doi:10.1016/j.jpdc.2017.06.009.

[119] J. M. Roberts, M. F. Bean, S. R. Cole, W. K. Young, and H. E. Weston. Informatics in the Analytical Laboratory: Vision for a New Decade. *Am. Pharm. Rev.*, 13(6):60, 2010. URL https://www.americanpharmaceuticalreview.com/Featured-Articles/115071-Informatics-in-the-Analytical-Laboratory-Vision-for-a-New-Decade/.

[120] J. M. Roberts, M. F. Bean, S. R. Cole, W. K. Young, and H. E. Weston. The Adaptable Laboratory: A Holistic Informatics Architecture. *Am. Pharm. Rev.*, 14 (1):12, 2011. URL https://www.americanpharmaceuticalreview.com/Featured-Articles/37098-The-Adaptable-Laboratory-A-Holistic-Informatics-Architecture/.

[121] L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, and A. Aspuru-Guzik. ChemOS: Orchestrating Autonomous Experimentation. *Sci. Robot.*, 3(19), 2018. doi:10.1126/scirobotics.aat5559.

[122] L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, and A. Aspuru-Guzik. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. *PLoS One*, 15(4):e0229862, 2020. doi:10.1371/journal.pone.0229862.

[123] A. Rosen, S. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. Notestein, and R. Q. Snurr. Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery. *Matter*, 4(5): 1578–1597, 2021. doi:10.1016/j.matt.2021.02.015.

[124] A. S. Rosen, J. M. Notestein, and R. Q. Snurr. Identifying Promising Metal–Organic Frameworks for Heterogeneous Catalysis via High-Throughput Periodic Density Functional Theory. *J. Comput. Chem.*, 40(12):1305–1318, 2019. doi:10.1002/jcc.25787.

[125] D. L. Roth. SPRESIweb 2.1, a Selective Chemical Synthesis and Reaction Database. *J. Chem. Inf. Model.*, 45(5):1470–1473, 2005. doi:10.1021/ci050274b.

[126] M. A. Rühl, R. Schäfer, and G. W. Kramer. Spectro ML-A Markup Language for Molecular Spectrometry Data. *JALA: J. Assoc. Lab. Autom.*, 6(6):76–82, 2001. doi:10.1016/S1535-5535-04-00168-6.

[127] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2010.

[128] M. Sabou, S. Biffl, A. Einfalt, L. Krammer, W. Kastner, and F. J. Ekaputra. Semantics for Cyber-Physical Systems: A Cross-Domain Perspective. *Semantic Web*, 11 (1):115–124, 2020. URL http://semantic-web-journal.net/content/semantics-cyber-physical-systems-cross-domain-perspective-0. Accessed 13 June 2021.

[129] B. Schäfer. Data Exchange in the Laboratory of the Future – A Glimpse at AnIML and SiLA, 2018. URL https://analyticalscience.wiley.com/do/10.1002/gitlab.17270/full/. Accessed 15 July 2021.

[130] G. Schneider. Automating drug discovery. *Nat. Rev. Drug Discovery*, 17(2):97–113, 2018. doi:10.1038/nrd.2017.232.

[131] N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli, and G. A. Landrum. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. *J. Med. Chem.*, 59(9):4385–4402, 2016. doi:10.1021/acs.jmedchem.6b00153.

[132] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.*, 5(9):1572–1583, 2019. doi:10.1021/acscentsci.9b00576.

[133] A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne, and A. A. Lapkin. Machine Learning Meets Continuous Flow Chemistry: Automated Optimization Towards the Pareto Front of Multiple Objectives. *Chem. Eng. J.*, 352: 277–282, 2018. doi:10.1016/j.cej.2018.07.031.

[134] M. H. S. Segler and M. P. Waller. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. Eur. J.*, 23(25):6118–6128, 2017. doi:10.1002/chem.201604556.

[135] M. H. S. Segler, M. Preuss, and M. P. Waller. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature*, 555(7698):604–610, 2018. doi:10.1038/nature25978.

[136] SiLA. SiLA Rapid Integration | Standardization in Lab Automation, 2021. URL https://sila-standard.com/. Accessed 27 May 2021.

[137] R. A. Skilton, R. A. Bourne, Z. Amara, R. Horvath, J. Jin, M. J. Scully, E. Streng, S. L. Y. Tang, P. A. Summers, J. Wang, E. Pérez, N. Asfaw, G. L. P. Aydos, J. Dupont, G. Comak, M. W. George, and M. Poliakoff. Remote-Controlled Experiments with Cloud Chemistry. *Nat. Chem.*, 7(1):1–5, 2015. doi:10.1038/nchem.2143.

[138] S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone, and L. Cronin. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. *Science*, 363(6423):eaav2211, 2019. doi:10.1126/science.aav2211.

[139] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, and A. Aspuru-Guzik. Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation. *Nat. Rev. Mater.*, 3 (5):5–20, 2018. doi:10.1038/s41578-018-0005-z.

[140] F. Tao and Q. Qi. Make More Digital Twins. *Nature*, 573:490–491, 2019. doi:10.1038/d41586-019-02849-1.

[141] C. J. Taylor, M. Booth, J. A. Manson, M. J. Willis, G. Clemens, B. A. Taylor, T. W. Chamberlain, and R. A. Bourne. Rapid, Automated Determination of Reaction Models and Kinetic Parameters. *Chem. Eng. J.*, 413:127017, 2021. doi:10.1016/j.cej.2020.127017.

[142] A. Thakkar, S. Johansson, K. Jorner, D. Buttar, J.-L. Reymond, and O. Engkvist. Artificial Intelligence and Automation in Computer Aided Synthesis Planning. *React. Chem. Eng.*, 6(1):27–51, 2021. doi:10.1039/D0RE00340A.

[143] The Foundation for Intelligent Physical Agents. Welcome to the Foundation for Intelligent Physical Agents, 2020. URL `http://www.fipa.org/`. Accessed 27 May 2021.

[144] Thermo Fisher Scientific (Informatics). An XML-Based File Format for Archival Storage of Analytical Instrument Data, 2001. URL `http://www.gaml.org/Documentation/XML%20Analytical%20Archive%20Format.pdf`. Accessed 31 July 2021.

[145] K. Tran and Z. W. Ulissi. Active Learning Across Intermetallics to Guide Discovery of Electrocatalysts for CO_2 Reduction and H_2 Evolution. *Nat. Catal.*, 1(9):696–703, 2018. doi:10.1038/s41929-018-0142-1.

[146] K. Tran, A. Palizhati, S. Back, and Z. W. Ulissi. Dynamic Workflows for Routine Materials Discovery in Surface Science. *J. Chem. Inf. Model.*, 58(12):2392–2400, 2018. doi:10.1021/acs.jcim.8b00386.

[147] P. Tremouilhac, C.-L. Lin, P.-C. Huang, Y.-C. Huang, A. Nguyen, N. Jung, F. Bach, R. Ulrich, B. Neumair, A. Streit, and S. Bräse. The Repository Chemotion: Infrastructure for Sustainable Research in Chemistry. *Angew. Chem., Int. Ed.*, 59(50): 22771–22778, 2020. doi:10.1002/anie.202007702.

[148] J. J. Varghese, L. Cao, C. Robertson, Y. Yang, L. F. Gladden, A. A. Lapkin, and S. H. Mushrif. Synergistic Contribution of the Acidic Metal Oxide–Metal Couple and Solvent Environment in the Selective Hydrogenolysis of Glycerol: A Combined Experimental and Computational Study Using ReO_X–Ir as the Catalyst. *ACS Catal.*, 9(1):485–503, 2018. doi:10.1021/acscatal.8b03079.

[149] T. Vieira, A. C. Stevens, A. Chtchemelinine, D. Gao, P. Badalov, and L. Heumann. Development of a Large-Scale Cyanation Process Using Continuous Flow Chemistry En Route to the Synthesis of Remdesivir. *Org. Process Res. Dev.*, 24(10): 2113–2121, 2020. doi:10.1021/acs.oprd.0c00172.

[150] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage, and J. Schuecker. Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Trans. Knowl. Data Eng.*, 2021. doi:10.1109/TKDE.2021.3079836.

[151] W3C. Semantic Web, 2015. URL https://www.w3.org/standards/semanticweb/. Accessed 1 June 2021.

[152] C. Waldron, A. Pankajakshan, M. Quaglio, E. Cao, F. Galvanin, and A. Gavriilidis. An Autonomous Microreactor Platform for the Rapid Identification of Kinetic Models. *React. Chem. Eng.*, 4(9):1623–1636, 2019. doi:10.1039/C8RE00345A.

[153] D. Weininger. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.*, 28(1): 31–36, 1988. doi:10.1021/ci00057a005.

[154] P. B. Wigley, P. J. Everitt, A. van den Hengel, J. W. Bastian, M. A. Sooriyabandara, G. D. McDonald, K. S. Hardman, C. D. Quinlivan, P. Manju, C. C. N. Kuhn, I. R. Petersen, A. N. Luiten, J. J. Hope, N. P. Robins, and M. R. Hush. Fast Machine-Learning Online Optimization of Ultra-Cold-Atom Experiments. *Sci. Rep.*, 6(1): 1–6, 2016. doi:10.1038/srep25890.

[155] L. Wilbraham, S. H. M. Mehr, and L. Cronin. Digitizing Chemistry Using the Chemical Processing Unit: From Synthesis to Discovery. *Acc. Chem. Res.*, 54(2): 253–262, 2021. doi:10.1021/acs.accounts.0c00674.

[156] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data*, 3(1):1–9, 2016. doi:10.1038/sdata.2016.18.

[157] E. L. Willighagen, A. Waagmeester, O. Spjuth, P. Ansell, A. J. Williams, V. Tkachenko, J. Hastings, B. Chen, and D. J. Wild. The ChEMBL Database as Linked Open Data. *J. Cheminf.*, 5(1):1–12, 2013. doi:10.1186/1758-2946-5-23.

[158] H. Winicov, J. Schainbaum, J. Buckley, G. Longino, J. Hill, and C. Berkoff. Chemical Process Optimization by Computer—A Self-Directed Chemical Synthesis System. *Anal. Chim. Acta*, 103(4):469–476, 1978. doi:10.1016/S0003-2670(01)83110-X.

[159] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman. Accelerated Search for Materials with Targeted Properties by Adaptive Design. *Nat. Commun.*, 7(1):1–9, 2016. doi:10.1038/ncomms11241.

[160] C. Zhang, A. Romagnoli, L. Zhou, and M. Kraft. Knowledge Management of Eco-industrial Park for Efficient Energy Utilization Through Ontology-Based Approach. *Appl. Energy*, 204:1412–1421, 2017. doi:10.1016/j.apenergy.2017.03.130.

[161] L. Zhou, M. Pan, J. J. Sikorski, S. Garud, L. K. Aditya, M. J. Kleinelanghorst, I. A. Karimi, and M. Kraft. Towards an Ontological Infrastructure for Chemical Process Simulation and Optimization in the Context of Eco-industrial Parks. *Appl. Energy*, 204:1284–1298, 2017. doi:10.1016/j.apenergy.2017.05.002.

[162] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, and S.-C. Zhang. Learning Atoms for Materials Discovery. *Proc. Natl. Acad. Sci.*, 115(28):E6411–E6417, 2018. doi:10.1073/pnas.1801181115.

[163] X. Zhou, A. Eibeck, M. Q. Lim, N. B. Krdzavac, and M. Kraft. An Agent Composition Framework for the J-Park Simulator - a Knowledge Graph for the Process Industry. *Comput. Chem. Eng.*, 130:106577, 2019. doi:10.1016/j.compchemeng.2019.106577.

[164] X. Zhou, D. Nurkowski, S. Mosbach, J. Akroyd, and M. Kraft. Question Answering System for Chemistry. *J. Chem. Inf. Model.*, 61(8):3868–3880, 2021. doi:10.1021/acs.jcim.1c00275.