From Platform to Knowledge Graph: Evolution of Laboratory Automation

Jiaru Bai,[†,‖] Liwei Cao,[†,‖] Sebastian Mosbach,[†,‡] Jethro Akroyd,[†,‡]

Alexei A. Lapkin,[*,†,‡] and Markus Kraft[*,†,‡,¶,§]

†*Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, United Kingdom*

‡*Cambridge Centre for Advanced Research and Education in Singapore (CARES), CREATE Tower #05-05, 1 Create Way, Singapore 138602*

¶*School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore 637459*

§*The Alan Turing Institute, London NW1 2DB, United Kingdom*

‖*J.B. and L.C. contributed equally to this work.*

E-mail: aal35@cam.ac.uk; mk306@cam.ac.uk

Phone: +44 (0)1223 762784

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.

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

Introduction

The automation of laboratory involves linking the abstract concepts of chemical processes and the hardware responsible for the execution.[1,2] 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.[3] 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.[4,5]

The first automated hardware for chemistry dates back to the late 1960s.[6] Since then, considerable advances have been made to expand the potentialities of such a tool, covering the field of chemical reactions,[7,8] drug discovery,[9] and material discovery for clean energy.[10,11] As chemists' quest to achieve a universal organic compound synthesis machine, three key

capabilities were identified,[12] *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.[13]. In 2018, Aspuru-Guzik and Persson[14] proposed materials acceleration platforms (MAP), a platform-based approach, as the paradigm to accelerate the material discovery process, which was further adopted and expanded by Flores-Leonar et al.[15]. In line with the three key capabilities that seem to be required to build a robo-chemist,[12] Flores-Leonar et al.[15] 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,[16,17] the discovery of chemical reactions,[18] synthesis planning,[19,20] and optimisation of process conditions.[21–23] 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. Especially when it comes to the vision of a globally integrated collaboration network.[11] 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.[8]

A way forward may be offered by Semantic Web technologies,[24] which present a 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. In this article, we refer to ontologies developed to describe knowledge in the chemistry domain, and more importantly, those implemented in a way that compatible with the semantic web standards,[25] as chemical ontologies. One prominent example is ChEBI.[26,27]

An ontology normally consists of two components: a terminological box (TBox) and an assertional box (ABox).[25] 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), essentially generalised uniform resource identifiers (URIs), for unambiguous identification. In the context of automating experiments, 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,[28] material discovery,[29] and reaction optimisation.[30] Considering the historical discussions of integrating the two technologies,[31] 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 aims to review the potential for arising technologies to enhance how we approach laboratory automation. The presentation of this perspective is structured as follows. First, we review the state-of-the-art in laboratory automation practice with a focus on data infrastructure. Based on the limitations of current approaches, we assess community efforts towards standardised data representation and effective data exchange. We identify dynamic knowledge graphs, *i.e.*, a combination of ontologies and agents, as an interesting technology option. This approach allows the intelligent automation of experiments to be linked with chemical knowledge resources and aligned with other AI techniques. It is suggested that this will play a key role in the next generation of laboratory automation.

Platform-based approach

Detailed reviews of the applications of the closed-loop optimisation have been published by Cao et al.[32] and Coley et al.[7]. 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.[15] 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.

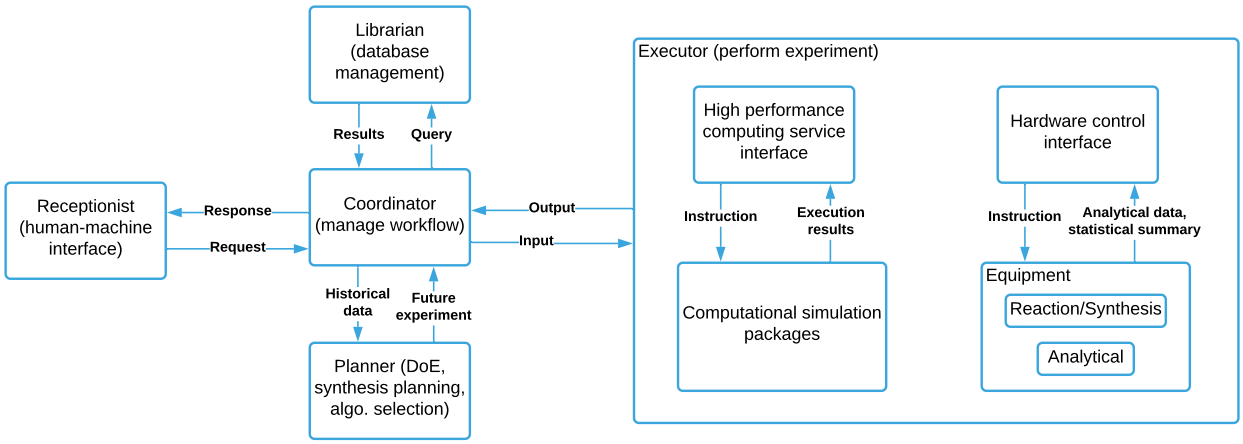


Figure 1: Functional components of a platform-based approach towards chemical discovery, annotated with the communications between each component.

Selected studies

There have been extensive reviews on developing each of the functional components.[15,33–36] In the context of chemical automation, Mateos et al.[37] 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 in the Supporting Information.

Receptionist

The receptionist acts as the human-machine interface. Among different platforms, multiple ways of interaction have been reported. Knight et al.[38] 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[21,39,40] 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.[35] 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.[16,41] 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,[42] Python,[17,19] and LabVIEW.[22,43,44] 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.

Coordinator

The coordinator manages the workflow in the closed-loop system. Among the different programming languages/tools that have been employed to develop the coordinator, Python is perhaps the most widely adopted. The Aspuru-Guzik Group proposed ChemOS,[16,41] 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[16] and increasing the efficiency of organic photovoltaics.[45] It has now been commercialised as Atinary SDLabs[46] with a Scientia version freely available for academics. Zhu's group presented MAOSIC,[17] a coordinator upgraded from their previous system MAOS,[47] which was applied to the autonomous discovery of

optically active chiral inorganic perovskite nanocrystals. Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE) has a coordinator acting as a bridge to connect the experimental workflow.[48] Its initial implementation was designed for the exploratory synthesis of single-crystal metal-halide perovskites. Further discovery of the formation of two new perovskite phases was demonstrated.[49] Chemputer[19] was developed for organic synthesis optimisation in 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, versioned and transferred flexibly.

LeyLab[39] is a PHP-based 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.

The Lapkin group presented a Matlab-based coordinator for multi-objective optimisation of the reaction conditions for SNAr and N-benzylation reactions.[50] It demonstrated its flexibility to a different chemical system with an aldol condensation reaction optimisation.[42]

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[43] or Python[44] are occasionally paired up with the LabVIEW to enable the planner module to suggest new experiments.

Another notable development is C#-based ARES OS,[51] an open-source software released by Air Force Research Laboratory (AFRL) following their autonomous research system (ARES). As the first reported autonomous experimentation system for materials development, ARES demonstrated its capability in carbon nanotube synthesis experiments,[52,53] and additive manufacturing applications.[54]

It can be seen that coordinators followed different coding philosophies in different pro-

gramming languages. For each case study, the reported coordinator indeed satisfied the specific need yet fail to extend to other systems.

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.[42] stored and transferred data as Matlab variables. Similarly, Christensen et al.[55] 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.[5],[32] used CSV files as the bridge for communication. Other studies used MAT files in a similar fashion.[22,56] 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.

Databases provide a solution to efficiently manage large amounts of experimental data. Li et al.[17] 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.[41], 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. Materials Experiment and Analysis Database (MEAD)[57] consists of both raw data and metadata from high-throughput experimentation. By instantiating an event-sourced architecture for materials provenances (ESAMP),[58] the MEAD database enabled the ML algorithm to utilise the material state within its experimental workflow for accelerating materials discovery.

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.[59] and Clayton et al.[60]. The communication between the coordinator and the planner is mainly done in two ways: variable stored in memory,[16,22,30] and file transfer.[5,19,20,50] It is worth mentioning that the communication protocols are not necessarily the same over one platform. Li et al.[17] 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.

Coordinator - Executor

The executor runs the experiments, computationally or physically, and sends back the experimental results. 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 interface. Therefore we review the communication protocols of the physical and computational experimental platforms separately.

Physical experiment interface Robotic platforms have their origins in instances such as peptide synthesis[6] and the pharmaceutical industry.[61,62] 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.[15]

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.[63,64] Greenaway et al.[65] 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. The hardware from Chemspeed is also used by IBM's RoboRxn,[66] a remotely-accessible automated organic synthesis platform utilising various Transformer[67]-based ML algorithms for chemical reaction prediction,[68] retrosynthetic pathway planning,[69] synthesis action extraction,[70] and chemistry grammar extraction.[71] 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,[72] scale-up development,[73] and reaction condition optimisation.[42,50] It is worth mentioning that commercially available mobile robots and robotic arms have been used in complex and multi-step operations.[20,23] 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.[21] 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.[74] controlled the pump system using the Raspberry Pi and interacted via an SCP with the executor codes. In Chemputer designed by Steiner et al.[19], an Arduino was designed as the micro-controller. 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.[17] 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.[75]. 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.

Robot Operating System (ROS)[76] is the *de facto* standard middleware in the robotics field for orchestrating multi-robot systems. In 2019, Marquez-Gamez and Maffetton[77] proposed a ROS architecture for laboratory robotics motivated by Burger et al.[23], envisaging a 'cobot' future where human researchers and robots work collaboratively in the chemistry lab using modular and reconfigurable lab equipment interfaced via ROS. A recent paper from Fakhruldeen et al.[78] shows proof-of-concept towards this direction.

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,[79] synthesis planning[80] and catalyst discovery.[81] By using

theoretical, fully automated screening methods combining ML and optimisation to guide density functional theory (DFT) calculations, Tran and Ulissi [82] 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 submit batch jobs 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). The scheduler stores the batch jobs, evaluates their resource requirements and priorities, and distributes the jobs to suitable compute nodes.

There are quite a few open-source scheduling software depending on the setup of HPC, among which SLURM is widely used in research computing services. [83] Rosen et al. [84] developed the PyMOFScreen Python package to manage automated DFT calculations, leading to new electronic structure database constructions and accelerate new materials discovery. [85] Multiple software packages were developed to enable high-throughput screening on the HPC, such as Python Materials Genomics (pymatgen), [86] FireWorks, [87] custodian, [86] Atomate, [88] GASpy, [81,82] and ChemEco. [89,90] 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. A notable effort in addressing this issue is MolSSI QCArchive, [91] which offers open access to millions of quantum chemistry calculations done with different software, as well as on-demand computation.

Current limitations

Despite the huge improvements made in the literature, a few limitations remain to be overcome before it is possible to achieve a global collaborative network. [11] 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.[21] 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.[92].

Unbalanced chemical data is another limitation to be addressed.[8] 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.[93] 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.[42,94] 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.

Data representation and exchange protocols

As promoted by various researchers,[1,8,36,95] 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.

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.[96] The most commonly used representation is string and line notation, including SMILES,[97] InChI,[98] SMARTS,[99] SELFIES,[100] *etc.* for molecules, and RInChI,[101] SMIRKS,[102] *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[103] and RDKit.[104] 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.

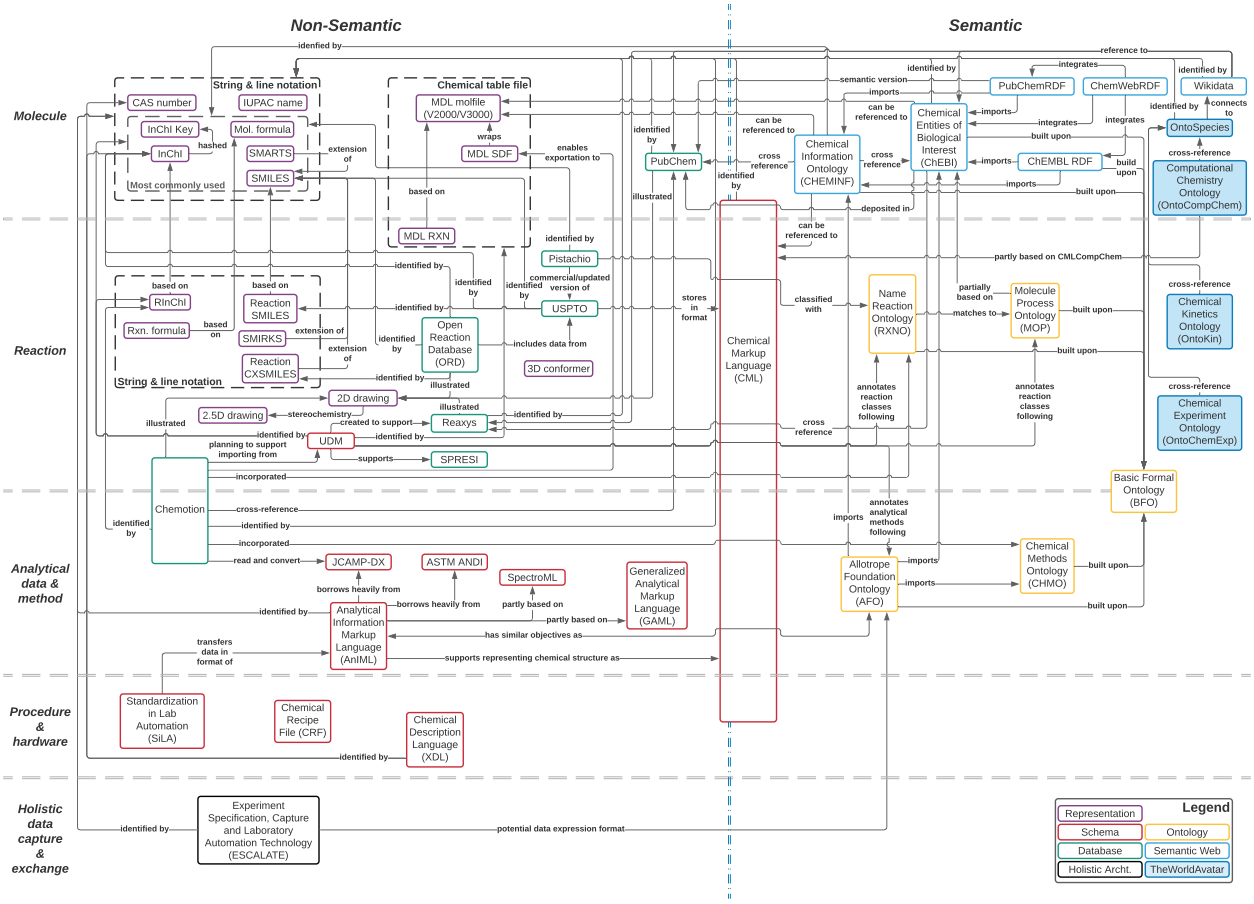


Figure 2: The community landscape towards a better data representation and exchange in chemical digitalisation. The focus of each category: (a) molecule: chemical structure, physicochemical properties, spectral information of a given species; (b) reaction: chemical reaction scheme, conditions, description 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.[25] Basic Formal Ontology (BFO) is an upper-level ontology as the basis of other ontologies and it does not capture any domain-specific information.

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[105] is the largest open-source structural chemical information repository. For reaction informatics,[106] the scale of open-source databases is much smaller. The USPTO database[107] is one of the seminal databases in the community that contains 3.7 million reactions extracted from US patents. It was commercialised as Pistachio[108] containing more than 13 million reactions with annotated reaction classifications using named reaction ontology (RXNO[109]) 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.[68,110–112] As the tailored database can be kept private to the research group, it could be difficult for bench-marking new algorithms.

To facilitate the development of ML in chemistry, Open Reaction Database (ORD)[113,114] 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.[115] 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,[115] 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.[117] 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[118]) 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[119]) where duplicate entries exist. However, it should be noted that the way UDM integrates the ontologies is by enumerating the ontological classes as a sub-schema 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.[120] 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.[121]

Similar to ORD, Chemotion[122] 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 analyt-

ical data in the widely-used JCAMP-DX format.[123] 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[124])) 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.[125]

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.[126] One approach to alleviate this problem is modernising the standard terms with an XML schema, such as Analytical Information Markup Language (AnIML).[127] AnIML is partly based on SpectroML[128] and Generalized Analytical Markup Language (GAML),[126] 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).[129] 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.[130] 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)[20] and chemical description language (XDL)[131] are initiatives to automate experiment execution. They both focus on translating the operational procedures from unstructured descriptions to robot execution commands.

CRF[20] 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[131] is an XML schema focusing on batch synthesis. It contains three main components 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.[25] 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.

ESCALATE is an attempt towards holistic data capture and exchange.[48] It proposed an ontological framework for experimentation, supporting data collection, reporting and experiment generation. This framework captures and reports all the reactions conducted, including "bad reactions" – in line with the cultural change promoted by the community.[95] In its first release,[48] the claimed ontological framework was realised by implementing template-based files to store the experimental information, *e.g.*, CSV and text files in a file-sharing folder infrastructure (Google Drive). The authors additionally acknowledge that the Allotrope Foundation Data Standard could be incorporated into this data lake. Despite uniform resource locators (URLs) being employed as pointers to some data, the data representation remains heterogeneous and only semi-structured, without the semantic features required by semantic web standards.[25] In a more recent development,[132] an ESCALATE REST API[133] was made available to showcase the possibility of retrieving chemical informatics data from PubChem API, interacting with a Postgres database for submitting experiment jobs to a laboratory and querying the hosted results.

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.

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.

It is worth noting the efforts to improve interoperability. Most of the schemas classify items using annotations based on ontological taxonomies. There are also works that claim to have developed ontologies, but that are not however represented in a formal ontology language such as Web Ontology Language (OWL) – their data is still file-based. In the context of this perspective, we consider these outputs to be taxonomies that formalise the hierarchical relationships, distinguishing them from the chemical ontologies that are introduced in the next section. The difficulty of achieving general interoperability remains an issue to be addressed.

Semantic representation

Since the landmark publication by Berners-Lee et al.[24], 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[134]. 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[135–137] is one of such early attempts by Murray-Rust and co-workers, contemporaneously to Berners-Lee's proposal of the semantic web.[24] In their work, CML was employed to host the data, prior to 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.[138]

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[25] are "too complex for the chemical community to take on board, and provides little effective added value"[139] 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),[140] *i.e.*, CHMO,[124] RXNO,[141] and MOP.[118] 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[26]) and cheminformatics (CHEMINF[142]) 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[143] and PubChemRDF,[144] the semantic version of the current largest open-source chemical information repository – PubChem.[105] 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[145] 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.[119] 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.[146],[147] 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.[148] 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.[115] 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[149] into RDF.

The same issue was acknowledged by the Allotrope Foundation[119] 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.[150] 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.

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.[151] 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.[29] wrapped different algorithms as agents to suggest the next experiments for DFT calculations on stable materials discovery. Gomes et al.[28] standardised various tasks as agents (bots) in a platform for crystal-structure phase mapping. Caramelli et al.[30] 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,[28,29] or posting plain-text on a human messaging platform (Twitter).[30] 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. A relevant first step towards addressing this issue is demonstrated by DLHub,[152] which allows users to publish, share, and cite ML models for applications in science.

Following the introduction of ontological data representations, a natural question is to ask whether the use of agents and ontologies can be combined to harness the strengths of both approaches. The challenge of how best to do this has been an open research question since the 2000s.[31] In theory,[24] 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[153] (FIPA) proposed a set of specifications focusing on communication and interoperability between agents. Specifically, FIPA Ontology Service Specification elaborated the idea of having an ontology agent to support the message interpretation between agents in detail. However, it never made it to the standard stage. In the following years, JADE,[154] 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. An attempt to do this is described in the next section.

Dynamic knowledge-graph-based approach

In this section, we explore how a combination of semantic web technologies and multi-agent systems – a dynamic knowledge-graph-based approach – might be applied to realise a complete digital and self-driving laboratory, *i.e.*, a chemical digital twin. We review an attempt to develop such an approach in the 'World Avatar' project. We subsequently outline a conceptual example of automated closed-loop optimisation powered by a dynamic knowledge graph, and assess its potential in achieving full automation.

Before diving into further details, we also provide a glossary of terms that are heavily used in this section. We acknowledge that the terms may have different meanings in other 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,[134] 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[155] 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.

Current state

The 'World Avatar' (http://theworldavatar.com/) project aims to develop an all-encompassing framework[156] that is capable of describing any aspect of the world. The 'World Avatar' uses a dynamic knowledge graph, based on an ontological representation of physical entities and interoperable agents. The agents are able to update the knowledge graph with new data, analyse data, make decision and control entities in the real world. This approach has been suggested to offer a suitable design for a universal 'digital twin'[157].

Starting from an industrial perspective, the J-Park Simulator – a precursor of 'The World Avatar'– developed a framework that was applied to describe waste energy[158] and optimise the operation[159] of an eco-industrial park on Jurong Island, Singapore.[160]

The 'World Avatar' has also been applied to describe a number of different types of chemical data, and provides ontologies for quantum chemistry (OntoCompChem[161]), chemical reaction kinetics (OntoKin[162]), chemical species (OntoSpecies[163]) and combustion experiments (OntoChemExp[164]). OntoSpecies links other ontologies to provide unambiguous identification of the chemicals, enabling translation of chemical names when integrating chemical data gathered from different sources.[164] The ontologies are connected to many of those described in previous sections. For instance, the development of OntoCompChem is partly based on the CompChem terms as described in the CML and the Gainesville Core (GNVC) ontology.[165] The relationship between these ontologies and other data representations used by the community is shown in Fig. 2.

To facilitate the automated data utilisation within the knowledge graph, an agent ontology (OntoAgent[166]) was developed as the design pattern of interoperable agents. 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.[166] 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.

Tools and resources All outputs from the 'World Avatar' project are available in the public domain. Various agents were developed and released on Github to provide service in the chemistry domain, *e.g.*, automated DFT calculations to address inconsistent thermodynamic data,[167] automated mechanism calibration to improve the alignment between kinetic models and experimental data,[164] and a question answering system enabling intuitive human data interaction – natural language queries of chemical data covering data from different sources.[168] Work is in progress to integrate services provided by agents into the natural language processing system so that on-demand computations can be invoked when a question could not be answered with the current knowledge. Users are welcome to check for more functionalities over time: https://kg.cmclinnovations.com/explore/marie.

Knowledge graph value proposition A core strength of the knowledge graph approach is interoperability. The knowledge graph provides a mechanism to combine data, descriptions of software, and hardware interfaces in a standardised way, facilitating automation and allowing communication between agents acting on data from different domains.[164,167]

Another key feature is the open-world assumption, enabling 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 au-

tonomously 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.

Automated closed-loop optimisation

The characteristics of dynamic knowledge graphs open up the possibility of a new and powerful approach to closed-loop optimisation. In this section, we explore how to apply a dynamic knowledge graph to do this in the context of a case study that was previously automated using a platform-based approach.[42] The case study considers flow chemistry. However, given suitable ontologies and agents, the underlying principles are expected to generalise to any practices in chemistry where a 'design-make-test-analyse' loop is involved.

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 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,

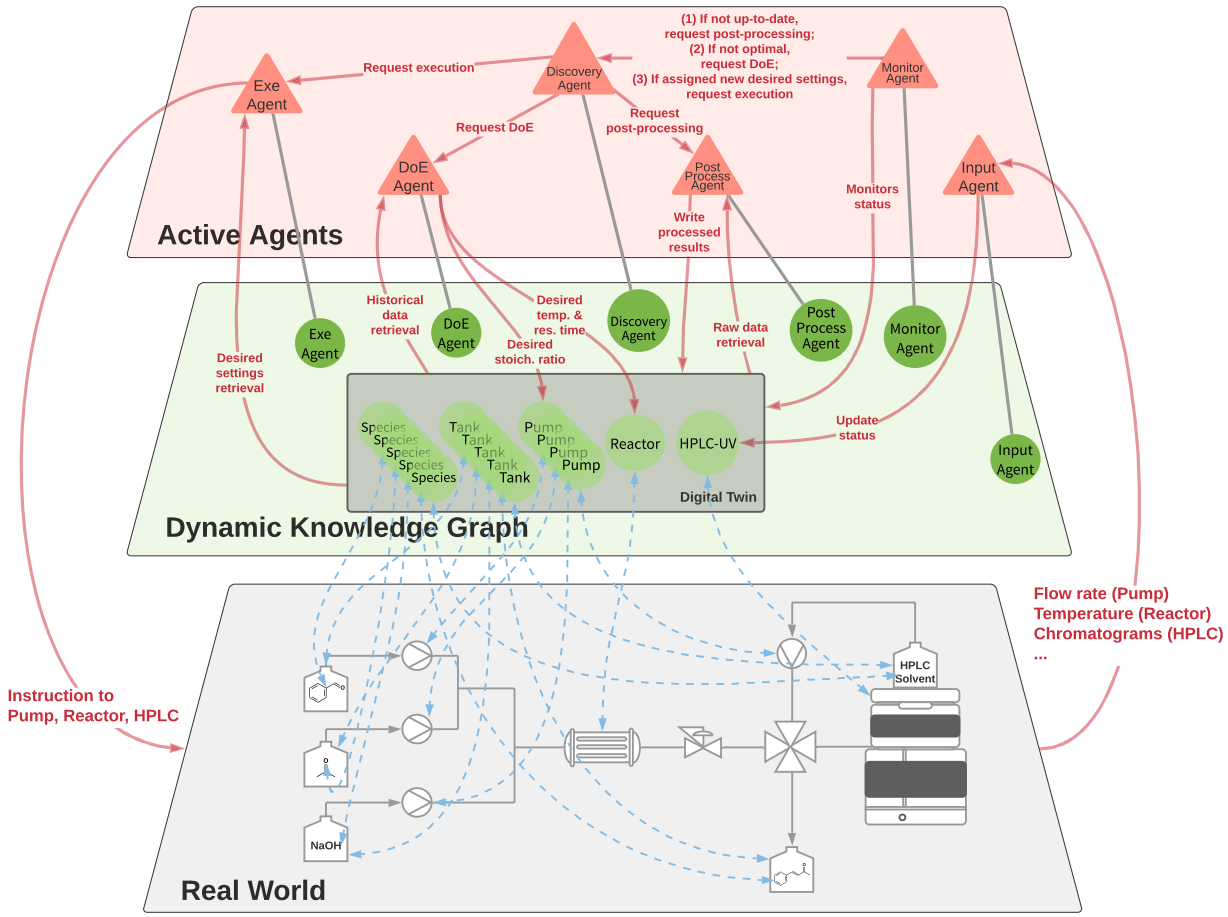


Figure 3: Dynamic knowledge-graph-based approach towards automated closed-loop optimisation. The real world layer demonstrates the existing physical entities, adapting from the experimentation setup of Jeraal et al.[42]. The dynamic knowledge graph layer hosts all the data generated during the experimentation 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.

distributed, multi-domain, interoperable, accessible, and most importantly evolving in time. As all the digital replicas of the hardware are expressed in the same way, new equipment can be immediately accessed by any existing software once it is instantiated in the knowledge graph. The same applies when adding new ML algorithms wrapped following OntoAgent specifications – standardised interactions with data and HPC services can be established in no time.[167] This enables the rapid integration of the most advanced algorithms and equipment. 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 this 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.

Towards a digital laboratory and beyond

Beyond closed-loop optimisation, various researchers have pictured the future towards the next-generation of autonomous laboratories and a global collaborative network.[1,8,11,15,36,40,66,92,146,147] 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.[8,36] 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.[147] 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 knowledge graph, question answering systems can be of help.[168] Researchers can thus interact with data intuitively from anywhere at any time, aligning with FAIR principles.[169] The semantic-rich nature incorporates prior knowledge into the data, presenting the potential to explore informed ML applications.[170]

Orchestration of physical and computational experiment This challenge lies in the emerging trend of physically synthesising the compounds identified by computational high-throughput screening.[8,65,92,171,172] 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.[146],[147] and dynamic knowledge-graph, information are promoted to be accessible to all stakeholders within a laboratory environment, flattening the structural design. For instance, active agents in the 'World Avatar' share the same world-view. The communication between them only serves as a pointer to the correct re-

sources (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.

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,[134] and particular application experience in the chemical engineering community (OntoCAPE[173]), 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.[174] 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.[175] In our view, the development of a digitalised and automated laboratory would enhance the capability of human researchers, enabling them to focus on creative activities, 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.[106,176] 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,[177] as has already been demonstrated by the application of the 'World Avatar' in smart city planning,[178] and the UK Digital Twin[157] (https://kg.cmclinnovations.com/explore/digital-twin). 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.

Conclusions and outlook

This contribution was motivated by the absence of standardised data representations and communication protocols which precludes further development towards the vision of a global collaborative research network.

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. This results in scalability issues when integrating new hardware and software, and interoperability issues when collaborating among different platforms – better

data representation and exchange are desired.

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

With our past experience in closed-loop optimisation and knowledge-graph development, we conjecture that a dynamic knowledge-graph-based approach would enable rapid integration of data and AI-based agents for chemical discovery and development. By integrating physical entities into the cyber space, it promotes better utilisation of the plethora of computational power in our efforts towards a sustainable future.[179]

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.

Following the knowledge graph approach, hopefully in the not too distant future, we will see the realisation of a global collaborative research network. 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 Jacob W. Martin for his advice on information management. The authors thank Dr Andrew C. Breeson for his help with proofreading. The authors thank Yiqun Bian and Guanhua Li for their helpful recommendations and feedback on colour scheme, which helped to improve the overall aesthetic expression of the TOC graphic. 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.

Supporting Information Available

The following files are available free of charge.

- suppinfo.pdf: This file lists the detailed findings from the selected state-of-the-art studies in chemical automation. To the best of our knowledge, we identified the functional component realisation in a platform-based approach in Table S1. Besides, in Table S2, we categorised the data flow and communication protocols between the functional components following the method described in the main text.

References

(1) Wilbraham, L.; Mehr, S. H. M.; Cronin, L. Digitizing Chemistry Using the Chemical Processing Unit: From Synthesis to Discovery. *Acc. Chem. Res.* **2021**, *54*, 253–262.

(2) Hammer, A. J. S.; Leonov, A. I.; Bell, N. L.; Cronin, L. Chemputation and the Standardization of Chemical Informatics. *JACS Au* **2021**, *1*, 1572–1587.

(3) Tao, F.; Qi, Q. Make More Digital Twins. *Nature* **2019**, *573*, 490–491.

(4) Inderwildi, O.; Zhang, C.; Wang, X.; Kraft, M. The Impact of Intelligent Cyber-Physical Systems on the Decarbonization of Energy. *Energy Environ. Sci.* **2020**, *13*, 744–771.

(5) Cao, L.; Russo, D.; Felton, K.; Salley, D.; Sharma, A.; Keenan, G.; Mauer, W.; Gao, H.; Cronin, L.; Lapkin, A. A. Optimization of Formulations Using Robotic Experiments Driven by Machine Learning DoE. *Cell Rep. Phys. Sci.* **2021**, *2*, 100295.

(6) Merrifield, R. B.; Stewart, J. M.; Jernberg, N. Instrument for Automated Synthesis of Peptides. *Anal. Chem.* **1966**, *38*, 1905–1914.

(7) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem., Int. Ed.* **2020**, *59*, 22858–22893.

(8) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem., Int. Ed.* **2020**, *59*, 23414–23436.

(9) Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discovery* **2018**, *17*, 97–113.

(10) Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; Amador-Bedolla, C.; Brabec, C. J.; Maruyama, B.; Persson, K. A.; Aspuru-Guzik, A. Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation. *Nat. Rev. Mater.* **2018**, *3*, 5–20.

(11) Stach, E. et al. Autonomous Experimentation Systems for Materials Development: A Community Perspective. *Matter* **2021**, *4*, 2702–2726.

(12) Peplow, M. Organic Synthesis: The Robo-Chemist. *Nature* **2014**, *512*, 20.

(13) Dimitrov, T.; Kreisbeck, C.; Becker, J. S.; Aspuru-Guzik, A.; Saikin, S. K. Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24825–24836.

(14) Aspuru-Guzik, A.; Persson, K. Materials Acceleration Platform: Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods and Artificial Intelligence. Mission Innovation: Innovation Challenge 6. 2018; `http://nrs.harvard.edu/urn-3:HUL.InstRepos:35164974`, Accessed 12 November 2021.

(15) Flores-Leonar, M. M.; Mejía-Mendoza, L. M.; Aguilar-Granda, A.; Sanchez-Lengeling, B.; Tribukait, H.; Amador-Bedolla, C.; Aspuru-Guzik, A. Materials Acceleration Platforms: On the Way to Autonomous Experimentation. *Curr. Opin. Green Sustain. Chem.* **2020**, 100370.

(16) MacLeod, B. P. et al. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials. *Sci. Adv.* **2020**, *6*, eaaz8867.

(17) Li, J.; Li, J.; Liu, R.; Tu, Y.; Li, Y.; Cheng, J.; He, T.; Zhu, X. Autonomous Discovery of Optically Active Chiral Inorganic Perovskite Nanocrystals through an Intelligent Cloud Lab. *Nat. Commun.* **2020**, *11*, 1–10.

(18) McNally, A.; Haffemayer, B.; Collins, B. S. L.; Gaunt, M. J. Palladium-Catalysed C–H Activation of Aliphatic Amines to Give Strained Nitrogen Heterocycles. *Nature* **2014**, *510*, 129–133.

(19) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. *Science* **2019**, *363*, eaav2211.

(20) Coley, C. W. et al. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365*, eaax1566.

(21) Fitzpatrick, D. E.; Maujean, T.; Evans, A. C.; Ley, S. V. Across-the-World Automated Optimization and Continuous-Flow Synthesis of Pharmaceutical Agents Operating through a Cloud-Based Server. *Angew. Chem., Int. Ed.* **2018**, *57*, 15128–15132.

(22) Bédard, A.-C.; Adamo, A.; Aroh, K. C.; Russell, M. G.; Bedermann, A. A.; Torosian, J.; Yue, B.; Jensen, K. F.; Jamison, T. F. Reconfigurable System for Automated Optimization of Diverse Chemical Reactions. *Science* **2018**, *361*, 1220–1225.

(23) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**, *583*, 237–241.

(24) Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Sci. Am.* **2001**, *284*, 34–43.

(25) W3C, Semantic Web. 2015; `https://www.w3.org/standards/semanticweb/`, Accessed 1 June 2021.

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

(27) Hastings, J.; Glauer, M.; Memariani, A.; Neuhaus, F.; Mossakowski, T. Learning Chemistry: Exploring the Suitability of Machine Learning for the Task of Structure-Based Chemical Ontology Classification. *J. Cheminf.* **2021**, *13*, 1–20.

(28) Gomes, C. P.; Bai, J.; Xue, Y.; Björck, J.; Rappazzo, B.; Ament, S.; Bernstein, R.; Kong, S.; Suram, S. K.; van Dover, R. B.; Gregoire, J. M. CRYSTAL: A Multi-Agent

AI System for Automated Mapping of Materials' Crystal Structures. *MRS Commun.* **2019**, *9*, 600–608.

(29) Montoya, J. H.; Winther, K. T.; Flores, R. A.; Bligaard, T.; Hummelshøj, J. S.; Aykol, M. Autonomous Intelligent Agents for Accelerated Materials Discovery. *Chem. Sci.* **2020**, *11*, 8517–8532.

(30) Caramelli, D.; Salley, D.; Henson, A.; Camarasa, G. A.; Sharabi, S.; Keenan, G.; Cronin, L. Networking Chemical Robots for Reaction Multitasking. *Nat. Commun.* **2018**, *9*, 1–10.

(31) Hendler, J. Agents and the Semantic Web. *IEEE Intell. Syst.* **2001**, *16*, 30–37.

(32) Cao, L.; Russo, D.; Lapkin, A. A. Automated Robotic Platforms in Design and Development of Formulations. *AIChE J.* **2021**, e17248.

(33) Godfrey, A. G.; Masquelin, T.; Hemmerle, H. A Remote-Controlled Adaptive Medchem Lab: An Innovative Approach to Enable Drug Discovery in the 21st Century. *Drug Discovery Today* **2013**, *18*, 795–802.

(34) Ley, S. V.; Fitzpatrick, D. E.; Ingham, R. J.; Myers, R. M. Organic Synthesis: March of the Machines. *Angew. Chem., Int. Ed.* **2015**, *54*, 3449–3464.

(35) Fitzpatrick, D. E.; Ley, S. V. Engineering Chemistry for the Future of Chemical Synthesis. *Tetrahedron* **2018**, *74*, 3087–3100.

(36) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. *Trends Chem.* **2019**, *1*, 282–291.

(37) Mateos, C.; Nieves-Remacha, M. J.; Rincón, J. A. Automated Platforms for Reaction Self-Optimization in Flow. *React. Chem. Eng.* **2019**, *4*, 1536–1544.

(38) Knight, N. J.; Kanza, S.; Cruickshank, D.; Brocklesby, W. S.; Frey, J. G. Talk2Lab: The Smart Lab of the Future. *IEEE Internet Things J.* **2020**, *7*, 8631–8640.

(39) Fitzpatrick, D. E.; Battilocchio, C.; Ley, S. V. A Novel Internet-Based Reaction Monitoring, Control and Autonomous Self-Optimization Platform for Chemical Synthesis. *Org. Process Res. Dev.* **2016**, *20*, 386–394.

(40) Ingham, R. J.; Battilocchio, C.; Fitzpatrick, D. E.; Sliwinski, E.; Hawkins, J. M.; Ley, S. V. A Systems Approach Towards an Intelligent and Self-Controlling Platform for Integrated Continuous Reaction Sequences. *Angew. Chem., Int. Ed.* **2015**, *127*, 146–150.

(41) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. *PLoS One* **2020**, *15*, e0229862.

(42) Jeraal, M. I.; Sung, S.; Lapkin, A. A. A Machine Learning-Enabled Autonomous Flow Chemistry Platform for Process Optimization of Multiple Reaction Metrics. *Chem. Methods* **2021**, *1*, 71–77.

(43) Mo, Y.; Rughoobur, G.; Nambiar, A. M. K.; Zhang, K.; Jensen, K. F. A Multifunctional Microfluidic Platform for High-Throughput Experimentation of Electroorganic Chemistry. *Angew. Chem., Int. Ed.* **2020**, *59*, 20890–20894.

(44) Chatterjee, S.; Guidi, M.; Seeberger, P. H.; Gilmore, K. Automated Radial Synthesis of Organic Molecules. *Nature* **2020**, *579*, 379–384.

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

(46) Atinary Technologies Inc. & Atinary Technologies Sàrl, Antinary – Enabling Self-Driving Laboratories. `https://atinary.com/`, Accessed 13 November 2021.

(47) Li, J.; Tu, Y.; Liu, R.; Lu, Y.; Zhu, X. Toward "On-Demand" Materials Synthesis and Scientific Discovery through Intelligent Robots. *Adv. Sci.* **2020**, *7*, 1901957.

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

(49) Li, Z.; Najeeb, M. A.; Alves, L.; Sherman, A. Z.; Shekar, V.; Cruz Parrilla, P.; Pendleton, I. M.; Wang, W.; Nega, P. W.; Zeller, M.; Schrier, J.; Norquist, A. J.; Chan, E. M. Robot-Accelerated Perovskite Investigation and Discovery. *Chem. Mater.* **2020**, *32*, 5650–5663.

(50) Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A. Machine Learning Meets Continuous Flow Chemistry: Automated Optimization Towards the Pareto Front of Multiple Objectives. *Chem. Eng. J.* **2018**, *352*, 277–282.

(51) Air Force Research Laboratory, ARES OS™. 2021; `https://github.com/AFRL-ARES/ARES_OS`, Accessed 13 November 2021.

(52) Nikolaev, P.; Hooper, D.; Perea-Lopez, N.; Terrones, M.; Maruyama, B. Discovery of Wall-Selective Carbon Nanotube Growth Conditions via Automated Experimentation. *ACS Nano* **2014**, *8*, 10214–10222.

(53) Nikolaev, P.; Hooper, D.; Webber, F.; Rao, R.; Decker, K.; Krein, M.; Poleski, J.; Barto, R.; Maruyama, B. Autonomy in Materials Research: A Case Study in Carbon Nanotube Growth. *npj Comput. Mater.* **2016**, *2*, 1–6.

(54) Deneault, J. R.; Chang, J.; Myung, J.; Hooper, D.; Armstrong, A.; Pitt, M.; Maruyama, B. Toward Autonomous Additive Manufacturing: Bayesian Optimization on a 3D Printer. *MRS Bull.* **2021**, *46*, 566—-575.

(55) Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Häse, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A.; Hein, J. Data-Science Driven Autonomous Process Optimization. *Commun. Chem.* **2021**, *4*, 1–12.

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

(57) Soedarmadji, E.; Stein, H. S.; Suram, S. K.; Guevarra, D.; Gregoire, J. M. Tracking Materials Science Data Lineage to Manage Millions of Materials Experiments and Analyses. *npj Comput. Mater.* **2019**, *5*, 1–9.

(58) Statt, M. J.; Rohr, B. A.; Brown, K.; Guevarra, D.; Hummelshøj, J. S.; Hung, L.; Anapolsky, A.; Gregoire, J. M.; Suram, S. K. ESAMP: Event-Sourced Architecture for Materials Provenance Management and Application to Accelerated Materials Discovery. **2021**, ChemRxiv Preprint.

(59) Garud, S. S.; Karimi, I. A.; Kraft, M. Design of Computer Experiments: A Review. *Comput. Chem. Eng.* **2017**, *106*, 71–95.

(60) Clayton, A. D.; Manson, J. A.; Taylor, C. J.; Chamberlain, T. W.; Taylor, B. A.; Clemens, G.; Bourne, R. A. Algorithms for the Self-Optimisation of Chemical Reactions. *React. Chem. Eng.* **2019**, *4*, 1545–1554.

(61) Winicov, H.; Schainbaum, J.; Buckley, J.; Longino, G.; Hill, J.; Berkoff, C. Chemical Process Optimization by Computer—A Self-Directed Chemical Synthesis System. *Anal. Chim. Acta* **1978**, *103*, 469–476.

(62) Lindsey, J. S. A Retrospective on the Automation of Laboratory Synthetic Chemistry. *Chemom. Intell. Lab. Syst.* **1992**, *17*, 15–45.

(63) McNally, A.; Prier, C. K.; MacMillan, D. W. Discovery of an α-Amino C–H Arylation Reaction Using the Strategy of Accelerated Serendipity. *Science* **2011**, *334*, 1114–1117.

(64) Hoogenboom, R.; Fijten, M. W. M.; Brändli, C.; Schroer, J.; Schubert, U. S. Automated Parallel Temperature Optimization and Determination of Activation Energy for the Living Cationic Polymerization of 2-Ethyl-2-Oxazoline. *Macromol. Rapid Commun.* **2003**, *24*, 98–103.

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

(66) O'Neill, S. AI-Driven Robotic Laboratories Show Promise. *Engineering* **2021**, In Press.

(67) Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020; pp 38–45.

(68) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

(69) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways Using Transformer-Based Models and a Hyper-Graph Exploration Strategy. *Chem. Sci.* **2020**, *11*, 3316–3325.

(70) Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. Auto-

mated Extraction of Chemical Synthesis Actions from Experimental Procedures. *Nat. Commun.* **2020**, *11*, 1–11.

(71) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions. *Sci. Adv.* **2021**, *7*, eabe4166.

(72) Gilmore, K.; Kopetzki, D.; Lee, J. W.; Horváth, Z.; McQuade, D. T.; Seidel-Morgenstern, A.; Seeberger, P. H. Continuous Synthesis of Artemisinin-Derived Medicines. *Chem. Commun.* **2014**, *50*, 12652–12655.

(73) Vieira, T.; Stevens, A. C.; Chtchemelinine, A.; Gao, D.; Badalov, P.; Heumann, L. Development of a Large-Scale Cyanation Process Using Continuous Flow Chemistry En Route to the Synthesis of Remdesivir. *Org. Process Res. Dev.* **2020**, *24*, 2113–2121.

(74) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: Orchestrating Autonomous Experimentation. *Sci. Robot.* **2018**, *3*, eaat5559.

(75) Fitzpatrick, D. E.; O'Brien, M.; Ley, S. V. A Tutored Discourse on Microcontrollers, Single Board Computers and Their Applications to Monitor and Control Chemical Reactions. *React. Chem. Eng.* **2020**, *5*, 201–220.

(76) Quigley, M.; Gerkey, B.; Conley, K.; Faust, J.; Foote, T.; Leibs, J.; Berger, E.; Wheeler, R.; Ng, A. ROS: An Open-Source Robot Operating System. ICRA Workshop on Open Source Software. 2009; Accessed 14 November 2021.

(77) Marquez-Gamez, D.; Maffetton, P. A ROS Based Architecture for an Autonomous Chemistry Laboratory. ROSCon Macau 2019. 2019.

(78) Fakhruldeen, H.; Marquez-Gamez, D.; Cooper, A. I. Development of a ROS Driver

and Support Stack for the KMR iiwa Mobile Manipulator. Annual Conference Towards Autonomous Robotic Systems. 2021; pp 304–314.

(79) Varghese, J. J.; Cao, L.; Robertson, C.; Yang, Y.; Gladden, L. F.; Lapkin, A. A.; Mushrif, S. H. 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.* **2018**, *9*, 485–503.

(80) Thakkar, A.; Johansson, S.; Jorner, K.; Buttar, D.; Reymond, J.-L.; Engkvist, O. Artificial Intelligence and Automation in Computer Aided Synthesis Planning. *React. Chem. Eng.* **2021**, *6*, 27–51.

(81) Tran, K.; Palizhati, A.; Back, S.; Ulissi, Z. W. Dynamic Workflows for Routine Materials Discovery in Surface Science. *J. Chem. Inf. Model.* **2018**, *58*, 2392–2400.

(82) Tran, K.; Ulissi, Z. W. Active Learning Across Intermetallics to Guide Discovery of Electrocatalysts for CO_2 Reduction and H_2 Evolution. *Nat. Catal.* **2018**, *1*, 696–703.

(83) Reuther, A.; Byun, C.; Arcand, W.; Bestor, D.; Bergeron, B.; Hubbell, M.; Jones, M.; Michaleas, P.; Prout, A.; Rosa, A.; Kepner, J. Scalable System Scheduling for HPC and Big Data. *J. Parallel Distrib. Comput.* **2018**, *111*, 76–92.

(84) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Identifying Promising Metal–Organic Frameworks for Heterogeneous Catalysis via High-Throughput Periodic Density Functional Theory. *J. Comput. Chem.* **2019**, *40*, 1305–1318.

(85) Rosen, A.; Iyer, S.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J.; Snurr, R. Q. Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery. *Matter* **2021**, *4*, 1578–1597.

(86) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.

(87) Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafman, M.; Petretto, G.; Rignanese, G.-M.; Hautier, G.; Gunter, D.; Persson, K. A. FireWorks: A Dynamic Workflow System Designed for High-Throughput Applications. *Concurr. Comput. Pract. Exp* **2015**, *27*, 5037–5059.

(88) Mathew, K. et al. Atomate: A High-Level Interface to Generate, Execute, and Analyze Computational Materials Science Workflows. *Comput. Mater. Sci.* **2017**, *139*, 140–152.

(89) Hachmann, J.; Afzal, M. A. F.; Haghighatlari, M.; Pal, Y. Building and Deploying a Cyberinfrastructure for the Data-driven Design of Chemical Systems and the Exploration of Chemical Space. *Mol. Simul.* **2018**, *44*, 921–929.

(90) Haghighatlari, M.; Vishwakarma, G.; Altarawy, D.; Subramanian, R.; Kota, B. U.; Sonpal, A.; Setlur, S.; Hachmann, J. ChemML: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, e1458.

(91) MolSSI QCArchive, The MolSSI Quantum Chemistry Archive. `https://qcarchive.molssi.org/`, Accessed 14 November 2021.

(92) Breen, C. P.; Nambiar, A. M. K.; Jamison, T. F.; Jensen, K. F. Ready, Set, Flow! Automated Continuous Synthesis and Optimization. *Trends Chem.* **2021**, *3*, 373–386.

(93) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73–76.

(94) Skilton, R. A. et al. Remote-Controlled Experiments with Cloud Chemistry. *Nat. Chem.* **2015**, *7*, 1–5.

(95) Herres-Pawlis, S.; Koepler, O.; Steinbeck, C. NFDI4Chem: Shaping a Digital and Cultural Change in Chemistry. *Angew. Chem., Int. Ed.* **2019**, *58*, 10766–10768.

(96) Zhou, Q.; Tang, P.; Liu, S.; Pan, J.; Yan, Q.; Zhang, S.-C. Learning Atoms for Materials Discovery. *Proc. Natl. Acad. Sci.* **2018**, *115*, E6411–E6417.

(97) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(98) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*, 1–34.

(99) Daylight, SMARTS - A Language for Describing Molecular Patterns. 2014; `https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`, Accessed 27 May 2021.

(100) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.

(101) Grethe, G.; Blanke, G.; Kraut, H.; Goodman, J. M. International Chemical Identifier for Reactions (RInChI). *J. Cheminf.* **2018**, *10*, 1–9.

(102) Daylight, SMIRKS - A Reaction Transform Language. 2014; `https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html`, Accessed 27 May 2021.

(103) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 1–14.

(104) Landrum, G., et al. RDKit: Open-Source Cheminformatics. Accessed 27 May 2021.

(105) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109.

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

(107) Lowe, D. Chemical Reactions from US Patents (1976-Sep2016). **2017**,

(108) NextMove Software, Pistachio. `https://www.nextmovesoftware.com/pistachio.html`, Accessed 15 July 2021.

(109) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. *J. Med. Chem.* **2016**, *59*, 4385–4402.

(110) Bradshaw, J.; Kusner, M. J.; Paige, B.; Segler, M. H. S.; Hernández-Lobato, J. M. A Generative Model for Electron Paths. Proceedings of the 7th International Conference on Learning Representations (ICLR 2019). 2019; pp 1–19.

(111) Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017; pp 2604–2613.

(112) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. "Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *Chem. Sci.* **2018**, *9*, 6091–6098.

(113) Open Reaction Database Project Authors, Welcome to the Open Reaction Database! 2021; `https://docs.open-reaction-database.org/en/latest/`, Accessed 27 May 2021.

(114) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143*, 18820–18826.

(115) Kearnes, S. The Open Reaction Database. NIH Virtual Workshop on Reaction Informatics, May 18-20. 2021; `https://cactus.nci.nih.gov/presentations/NIHReactInf_2021-05/Kearnes_Open_Reaction_Database-NIH_Reaction_Informatics.pptx`, Accessed 13 November 2021.

(116) Pistoia Alliance, Unified Data Model. 2020; `https://github.com/PistoiaAlliance/UDM`, Accessed 27 May 2021.

(117) Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **2009**, *49*, 2897–2898.

(118) EMBL-EBI, Molecular Process Ontology. 2014; `https://www.ebi.ac.uk/ols/ontologies/mop`, Accessed 14 June 2021.

(119) Millecam, T.; Jarrett, A. J.; Young, N.; Vanderwall, D. E.; Della Corte, D. Coming of Age of Allotrope: Proceedings from the Fall 2020 Allotrope Connect. *Drug Discovery Today* **2021**, *26*, 1922–1928.

(120) Roth, D. L. SPRESIweb 2.1, a Selective Chemical Synthesis and Reaction Database. *J. Chem. Inf. Model.* **2005**, *45*, 1470–1473.

(121) Blanke, G. The Unified Data Model (UDM). NIH Virtual Workshop on Reaction Informatics, May 18-20. 2021; `https://cactus.nci.nih.gov/presentations/NIHReactInf_2021-05/UDM_at_NIH_Reaction_conference_May_2021_-_Gerd_Blanke.pdf`, Accessed 13 November 2021.

(122) Tremouilhac, P.; Lin, C.-L.; Huang, P.-C.; Huang, Y.-C.; Nguyen, A.; Jung, N.; Bach, F.; Ulrich, R.; Neumair, B.; Streit, A.; Bräse, S. The Repository Chemotion:

Infrastructure for Sustainable Research in Chemistry. *Angew. Chem., Int. Ed.* **2020**, *59*, 22771–22778.

(123) Lampen, P.; Lambert, J.; Lancashire, R. J.; McDonald, R. S.; McIntyre, P. S.; Rutledge, D. N.; Fröhlich, T.; Davies, A. N. An Extension to the JCAMP-DX Standard File Format, JCAMP-DX V. 5.01. *Pure Appl. Chem.* **1999**, *71*, 1549–1556.

(124) EMBL-EBI, Chemical Methods Ontology. 2019; `https://www.ebi.ac.uk/ols/ontologies/chmo`, Accessed 14 June 2021.

(125) Jung, N. Documentation and Publication of Reactions with Chemotion ELN and Repository. NIH Virtual Workshop on Reaction Informatics, May 18-20. 2021; `https://cactus.nci.nih.gov/presentations/NIHReactInf_2021-05/Nicole_Jung_Chemotion_NIH_2021.pdf`, Accessed 13 November 2021.

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

(127) AnIML Working Group, AnIML: Overview. `https://www.animl.org/overview`, Accessed 31 July 2021.

(128) Rühl, M. A.; Schäfer, R.; Kramer, G. W. Spectro ML-A Markup Language for Molecular Spectrometry Data. *JALA: J. Assoc. Lab. Autom.* **2001**, *6*, 76–82.

(129) SiLA, SiLA Rapid Integration | Standardization in Lab Automation. 2021; `https://sila-standard.com/`, Accessed 27 May 2021.

(130) Schäfer, B. Data Exchange in the Laboratory of the Future – A Glimpse at AnIML and SiLA. 2018; `https://analyticalscience.wiley.com/do/10.1002/gitlab.17270/full/`, Accessed 15 July 2021.

(131) Mehr, S. H. M.; Craven, M.; Leonov, A. I.; Keenan, G.; Cronin, L. A Universal System for Digitization and Automatic Execution of the Chemical Synthesis Literature. *Science* **2020**, *370*, 101–108.

(132) Noack, M. M.; Sethian, J. A. Autonomous Discovery in Science and Engineering. 2021; `https://www.osti.gov/biblio/1818491`, Accessed 14 November 2021.

(133) ESCALATE, Interacting with the ESCALATE REST API. `https://github.com/darkreactions/ESCALATE/blob/master/demonstrations/REST_API_DEMO.ipynb`, Accessed 15 November 2021.

(134) Hitzler, P. A Review of The Semantic Web Field. *Commun. ACM* **2021**, *64*, 76–83.

(135) Gkoutos, G. V.; Murray-Rust, P.; Rzepa, H. S.; Wright, M. Chemical Markup, XML, and the World-Wide Web. 3. Toward a Signed Semantic Chemical Web of Trust. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1124–1130.

(136) Murray-Rust, P.; Rzepa, H. S.; Tyrrell, S. M.; Zhang, Y. Representation and Use of Chemistry in the Global Electronic Age. *Org. Biomol. Chem.* **2004**, *2*, 3192–3203.

(137) Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. Enhancement of the Chemical Semantic Web through the Use of InChI Identifiers. *Org. Biomol. Chem.* **2005**, *3*, 1832–1834.

(138) Murray-Rust, P. Chemistry for Everyone. *Nature* **2008**, *451*, 648–651.

(139) Murray-Rust, P. CML - Frequently Asked Questions. `http://www.xml-cml.org/documentation/FAQ.html#chemistry`, Accessed 31 July 2021.

(140) Batchelor, C.; Corbett, P. Semantic Enrichment of Journal Articles Using Chemical Named Entity Recognition. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. 2007; pp 45–48, Accessed 31 July 2021.

(141) EMBL-EBI, Name Reaction Ontology. 2021; `https://www.ebi.ac.uk/ols/ontologies/rxno`, Accessed 14 June 2021.

(142) Hastings, J.; Chepelev, L.; Willighagen, E.; Adams, N.; Steinbeck, C.; Dumontier, M. The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web. *PloS One* **2011**, *6*, e25513.

(143) Willighagen, E. L.; Waagmeester, A.; Spjuth, O.; Ansell, P.; Williams, A. J.; Tkachenko, V.; Hastings, J.; Chen, B.; Wild, D. J. The ChEMBL Database as Linked Open Data. *J. Cheminf.* **2013**, *5*, 1–12.

(144) Fu, G.; Batchelor, C.; Dumontier, M.; Hastings, J.; Willighagen, E.; Bolton, E. PubChemRDF: Towards the Semantic Annotation of PubChem Compound and Substance Databases. *J. Cheminf.* **2015**, *7*, 1–15.

(145) Galgonek, J.; Vondrášek, J. IDSM ChemWebRDF: SPARQLing Small-Molecule Datasets. *J. Cheminf.* **2021**, *13*, 1–19.

(146) Roberts, J. M.; Bean, M. F.; Cole, S. R.; Young, W. K.; Weston, H. E. Informatics in the Analytical Laboratory: Vision for a New Decade. *Am. Pharm. Rev.* **2010**, *13*, 60.

(147) Roberts, J. M.; Bean, M. F.; Cole, S. R.; Young, W. K.; Weston, H. E. The Adaptable Laboratory: A Holistic Informatics Architecture. *Am. Pharm. Rev.* **2011**, *14*, 12.

(148) Bard, J. B. L.; Rhee, S. Y. Ontologies in Biology: Design, Applications and Future Challenges. *Nat. Rev. Genet.* **2004**, *5*, 213–222.

(149) Menon, A.; Krdzavac, N. B.; Kraft, M. From Database to Knowledge Graph—Using Data in Chemistry. *Curr. Opin. Chem. Eng.* **2019**, *26*, 33–37.

(150) Godfrey, A. G.; Michael, S. G.; Sittampalam, G. S.; Zahoránszky-Köhalmi, G. A Perspective on Innovating the Chemistry Lab Bench. *Front. Rob. AI* **2020**, *7*, 24.

(151) Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Prentice Hall, 2010.

(152) Chard, R.; Li, Z.; Chard, K.; Ward, L.; Babuji, Y.; Woodard, A.; Tuecke, S.; Blaiszik, B.; Franklin, M. J.; Foster, I. DLHub: Model and Data Serving for Science. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 2019; pp 283–292.

(153) The Foundation for Intelligent Physical Agents, Welcome to the Foundation for Intelligent Physical Agents. 2020; `http://www.fipa.org/`, Accessed 27 May 2021.

(154) JADE, Java Agent DEvelopment Framework: Jade Site. 2021; `https://jade.tilab.com/`, Accessed 27 May 2021.

(155) Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. Eur. J.* **2017**, *23*, 6118–6128.

(156) Eibeck, A.; Lim, M. Q.; Kraft, M. J-Park Simulator: An Ontology-Based Platform for Cross-domain Scenarios in Process Industry. *Comput. Chem. Eng.* **2019**, *131*, 106586.

(157) Akroyd, J.; Mosbach, S.; Bhave, A.; Kraft, M. Universal Digital Twin – A Dynamic Knowledge Graph. *Data-Centric Engineering* **2021**, *2*, e14.

(158) Zhang, C.; Romagnoli, A.; Zhou, L.; Kraft, M. Knowledge Management of Eco-industrial Park for Efficient Energy Utilization Through Ontology-Based Approach. *Appl. Energy* **2017**, *204*, 1412–1421.

(159) Zhou, L.; Pan, M.; Sikorski, J. J.; Garud, S.; Aditya, L. K.; Kleinelanghorst, M. J.; Karimi, I. A.; Kraft, M. Towards an Ontological Infrastructure for Chemical Process Simulation and Optimization in the Context of Eco-industrial Parks. *Appl. Energy* **2017**, *204*, 1284–1298.

(160) Pan, M.; Sikorski, J.; Kastner, C. A.; Akroyd, J.; Mosbach, S.; Lau, R.; Kraft, M. Applying Industry 4.0 to the Jurong Island Eco-industrial Park. *Energy Procedia* **2015**, *75*, 1536–1541.

(161) Krdzavac, N.; Mosbach, S.; Nurkowski, D.; Buerger, P.; Akroyd, J.; Martin, J.; Menon, A.; Kraft, M. An Ontology and Semantic Web Service for Quantum Chemistry Calculations. *J. Chem. Inf. Model.* **2019**, *59*, 3154–3165.

(162) Farazi, F.; Akroyd, J.; Mosbach, S.; Buerger, P.; Nurkowski, D.; Salamanca, M.; Kraft, M. OntoKin: An Ontology for Chemical Kinetic Reaction Mechanisms. *J. Chem. Inf. Model.* **2020**, *60*, 108–120.

(163) Farazi, F.; Krdzavac, N. B.; Akroyd, J.; Mosbach, S.; Menon, A.; Nurkowski, D.; Kraft, M. Linking Reaction Mechanisms and Quantum Chemistry: An Ontological Approach. *Comput. Chem. Eng.* **2020**, *137*, 106813.

(164) Bai, J.; Geeson, R.; Farazi, F.; Mosbach, S.; Akroyd, J.; Bringley, E. J.; Kraft, M. Automated Calibration of a Poly(oxymethylene) Dimethyl Ether Oxidation Mechanism Using the Knowledge Graph Technology. *J. Chem. Inf. Model.* **2021**, *61*, 1701–1717.

(165) Chemical Semantics, GNVC: Gainesville Core Ontology - Standard for Publishing Results of Computational Chemistry. 2015; `http://ontologies.makolab.com/gc/gc07.owl`, Accessed 21 September 2021.

(166) Zhou, X.; Eibeck, A.; Lim, M. Q.; Krdzavac, N. B.; Kraft, M. An Agent Composition Framework for the J-Park Simulator - a Knowledge Graph for the Process Industry. *Comput. Chem. Eng.* **2019**, *130*, 106577.

(167) Mosbach, S.; Menon, A.; Farazi, F.; Krdzavac, N.; Zhou, X.; Akroyd, J.; Kraft, M. Multiscale Cross-Domain Thermochemical Knowledge-Graph. *J. Chem. Inf. Model.* **2020**, *60*, 6155–6166.

(168) Zhou, X.; Nurkowski, D.; Mosbach, S.; Akroyd, J.; Kraft, M. Question Answering System for Chemistry. *J. Chem. Inf. Model.* **2021**, *61*, 3868–3880.

(169) Wilkinson, M. D. et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 1–9.

(170) von Rueden, L.; Mayer, S.; Beckh, K.; Georgiev, B.; Giesselbach, S.; Heese, R.; Kirsch, B.; Walczak, M.; Pfrommer, J.; Pick, A.; Ramamurthy, R.; Garcke, J.; Bauckhage, C.; Schuecker, J. Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Trans. Knowl. Data Eng.* **2021**, In Press.

(171) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191–201.

(172) Coley, C. W. Defining and Exploring Chemical Spaces. *Trends Chem.* **2021**, *3*, 133–145.

(173) Morbach, J.; Yang, A.; Marquardt, W. OntoCAPE - a Large-scale Ontology for Chemical Process Engineering. *Eng. Appl. Artif. Intell.* **2007**, *20*, 147–161.

(174) Noy, N.; Gao, Y.; Jain, A.; Narayanan, A.; Patterson, A.; Taylor, J. Industry-Scale Knowledge Graphs: Lessons and Challenges. *Commun. ACM* **2019**, *62*, 36–43.

(175) Brazil, R. Automation in the Chemistry Lab. 2021; `https://www.chemistryworld.com/careers/automation-in-the-chemistry-lab/4012832.article`, Accessed 31 July 2021.

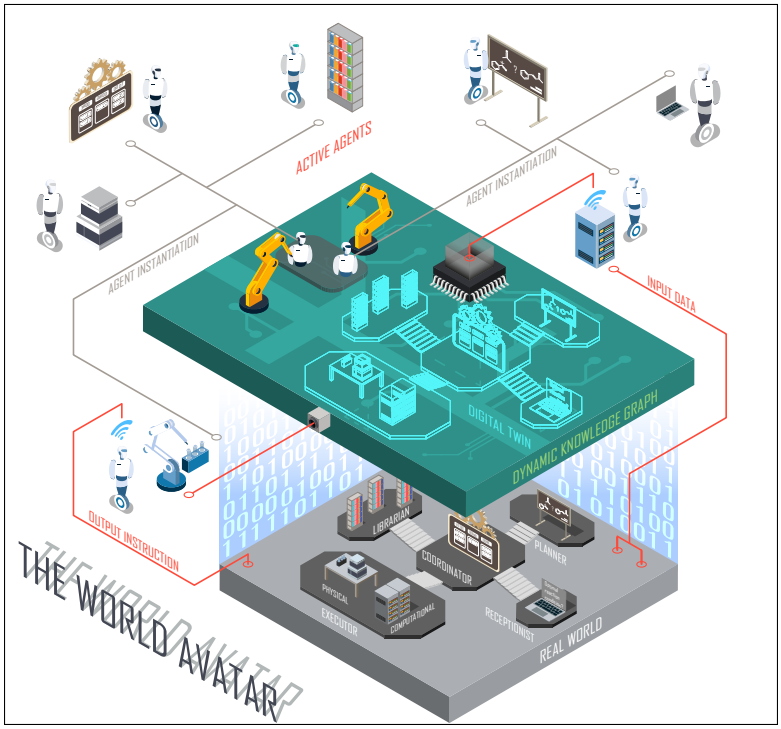
(176) Nicklaus, M. C. NIH Virtual Workshop on Ultra-Large Chemistry Databases, Dec 1-3. 2020; `https://cactus.nci.nih.gov/presentations/NIHBigDB_2020-12/NIHBigDB.html`, Accessed 31 July 2021.

(177) Sabou, M.; Biffl, S.; Einfalt, A.; Krammer, L.; Kastner, W.; Ekaputra, F. J. Semantics for Cyber-Physical Systems: A Cross-Domain Perspective. *Semantic Web* **2020**, *11*, 115–124, Accessed 13 June 2021.

(178) Chadzynski, A.; Krdzavac, N.; Farazi, F.; Lim, M. Q.; Li, S.; Grisiute, A.; Herthogs, P.; von Richthofen, A.; Cairns, S.; Kraft, M. Semantic 3D City Database - An Enabler for a Dynamic Geospatial Knowledge Graph. *Energy and AI* **2021**, *6*, 100106.

(179) Gomes, C. et al. Computational Sustainability: Computing for a Better World and a Sustainable Future. *Commun. ACM* **2019**, *62*, 56–65.

TOC Graphic


ACTIVE AGENTS
AGENT INSTANTIATION
AGENT INSTANTIATION
INPUT DATA
DIGITAL TWIN
DYNAMIC KNOWLEDGE GRAPH
OUTPUT INSTRUCTION
LIBRARIAN
COORDINATOR
PLANNER
PHYSICAL
COMPUTATIONAL
EXECUTOR
RECEPTIONIST
REAL WORLD
THE WORLD AVATAR