Article

https://doi.org/10.1038/s41467-023-44599-9

# A dynamic knowledge graph approach to distributed self-driving laboratories

Received: 14 Ju	ly 2023
-----------------	---------

Accepted: 21 December 2023

Published online: 23 January 2024

Check for updates

Jiaru Bai <sup>1</sup>, Sebastian Mosbach <sup>12</sup>, Connor J. Taylor<sup>3,4,8</sup>, Dogancan Karan<sup>2</sup>, Kok Foong Lee<sup>5</sup>, Simon D. Rihm <sup>1,2</sup>, Jethro Akroyd <sup>1,2</sup>, Alexei A. Lapkin <sup>1,2,4</sup> & Markus Kraft <sup>1,2,6,7</sup>

The ability to integrate resources and share knowledge across organisations empowers scientists to expedite the scientific discovery process. This is especially crucial in addressing emerging global challenges that require global solutions. In this work, we develop an architecture for distributed self-driving laboratories within The World Avatar project, which seeks to create an allencompassing digital twin based on a dynamic knowledge graph. We employ ontologies to capture data and material flows in design-make-test-analyse cycles, utilising autonomous agents as executable knowledge components to carry out the experimentation workflow. Data provenance is recorded to ensure its findability, accessibility, interoperability, and reusability. We demonstrate the practical application of our framework by linking two robots in Cambridge and Singapore for a collaborative closed-loop optimisation for a pharmaceutically-relevant aldol condensation reaction in real-time. The knowledge graph autonomously evolves toward the scientist's research goals, with the two robots effectively generating a Pareto front for cost-yield optimisation in three days.

The concept of laboratory automation, recently reinterpreted as selfdriving laboratories (SDLs)<sup>1,2</sup>, has been in existence since the 1960s, when ref. 3 introduced the first automated chemistry hardware. Since then, SDLs have gained widespread adoption in chemistry<sup>4–7</sup>, materials science<sup>8,9</sup>, biotechnology<sup>10,11</sup> and robotics<sup>12</sup>, resulting in accelerated scientific discovery and societal development. However, the implementation of SDLs can be challenging and typically requires a highly specialised team of researchers with expertise in chemistry, engineering, and computer science. Consequently, studies are often conducted by large research groups within a single organisation. Even in cases where collaborations occur between research groups, the SDL is usually centralised within the same laboratory. In response to the pressing global challenges of today, there is a growing consensus within the scientific community that a paradigm shift towards a globally collaborative research network is necessary<sup>13–15</sup>. This shift requires decentralising SDLs to integrate different research groups to contribute their expertise towards solving emerging problems<sup>16</sup>. Such decentralisation holds great potential in supporting various tasks ranging from automating the characterisation of epistemic uncertainty in experimental research<sup>17</sup> to advancing human exploration in deep space<sup>18</sup>. Achieving this vision is not an easy task and entails three major challenges. The first challenge is efficiently orchestrating heterogeneous resources<sup>19</sup>, which includes hardware from different vendors and diverse computing environments.

<sup>1</sup>Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, UK. <sup>2</sup>Cambridge Centre for Advanced Research and Education in Singapore (CARES), 1 Create Way, CREATE Tower, #05-05, Singapore 138602, Singapore. <sup>3</sup>Astex Pharmaceuticals, 436 Cambridge Science Park Milton Road, Cambridge CB4 0QA, UK. <sup>4</sup>Innovation Centre in Digital Molecular Technologies, Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. <sup>5</sup>CMCL Innovations, Sheraton House, Cambridge CB3 0AX, UK. <sup>6</sup>School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, 637459 Singapore, Singapore. <sup>7</sup>The Alan Turing Institute, London NW1 2DB, UK. <sup>8</sup>Present address: Faculty of Engineering, University of Nottingham, University Park, Nottingham NG7 2RD, UK. The second challenge is sharing data across organisations<sup>20</sup>, which requires standardising language in which the research is communicated<sup>21</sup>. During this process, the source and metadata of the research need to be tracked to facilitate reproducibility, which leads to the third challenge of data provenance recording following FAIR principles – Findable, Accessible, Interoperable and Reusable<sup>22</sup>.

Many attempts have been made to tackle these challenges with different focuses. For resource orchestration, middleware such as ChemOS<sup>23</sup>, ESCALATE<sup>24</sup>, and HELAO<sup>25</sup> exist to glue different components within an SDL and abstract the hardware resources. For data sharing,  $\chi$ DL<sup>26,27</sup> and AnIML<sup>28</sup> are examples of standard protocols developed for synthesis and analysis respectively. In the realm of data provenance, Mitchell et al.<sup>29</sup> proposed a data pipeline to support the modelling of the COVID pandemic, whereas ref. 30 devised a knowledge graph to record experiment provenance in materials research. Although these studies provide insights into building a collaborative research environment, they are developed in isolation with customised





data interfaces. Enhancing interoperability both within and between these systems is essential to establish a truly connected research network.

As discussed in our previous work<sup>31,32</sup>, semantic web technologies such as knowledge graphs<sup>33</sup> offer a viable path forward. Ontologies abstract both resources and data using the same notion. allowing for a common language between participants when allocating tasks and sharing results. The World Avatar<sup>34,35</sup> is such a knowledge graph that aims to encompass all aspects of scientific research laboratories as shown in Fig. 1a in their entirety: The experiment itself, including its physical setup and underlying chemistry; moving handlers that can be of human or robotic nature; and the laboratory providing necessary infrastructure and resources<sup>36</sup>. The World Avatar goes beyond static knowledge representation by encoding software agents as executable knowledge components, enabling dynamicity and continuous incorporation of new concepts and data while preserving connections to existing information. As the knowledge graph expands, this characteristic allows for capturing data provenance from experimental processes as knowledge statements, effectively acting as a living copy of the real world. This dynamic knowledge graph streamlines the immediate dissemination of data between SDLs, offering a promising holistic solution to the aforementioned challenges<sup>32,37</sup> and the pursuit of the Nobel Turing Challenge<sup>36,38</sup>.

In this work, we demonstrate a proof-of-concept for a distributed network of SDLs enabled by a dynamic knowledge graph. This signifies the first step towards digital research scientists (as shown in Fig. 1a) collaborating autonomously. To illustrate the effectiveness of this approach, as shown in Fig. 1b, we present a demonstration using two robots in Cambridge and Singapore collaborating on a multi-objective closed-loop optimisation problem in response to a goal request from scientists.

#### Results

#### Architecture of distributed SDLs

Closed-loop optimisation in SDLs is a dynamic process that revolves around design-make-test-analyse (DMTA) cycles<sup>39,40</sup>. Compared to machine learning systems and scientific workflows that only capture data flows, SDLs offer an integrated approach by orchestrating both computational and physical resources. This involves the integration of data and material flows, as well as the interface that bridges the gap between the virtual and physical worlds. To this end, we propose a conceptual architecture of distributed SDLs that effectively incorporates all three flows, as illustrated in Fig. 2a.

The proposed architecture presents a framework to enable scientists to set research goals and resource restrictions for a particular chemical reaction and have them trigger a closed-loop process in cyberspace. The process is initiated by the monitoring component, which parses the research goals and requests the iterations needed to achieve the objectives. The iterating component collects prior information about the design space and passes it on to the component that designs the next experiment. The algorithm employed, as well as the availability of prior data, determines the combination of design variables to be proposed within the search space provided by the scientist. Subsequently, the proposed physical experimentation is scheduled for execution in one of the available laboratories, similar to the scheduling of high-performance computing jobs<sup>41</sup>. The suggested conditions are translated to the machine-actionable recipe that enables the control of hardware for reaction and characterisation. In the physical world, this is reflected in the material flow between the two pieces of equipment. The data processing component is then responsible for computing the objectives by analysing the complete job information and raw data. If the resources are still available, a comparison of these objectives with the research goals determines whether the system should proceed to the next iteration.



Fig. 2 | An illustration of a distributed self-driving laboratories (SDLs) architecture. a Conceptual framework of components used to build a network of distributed SDLs for closed-loop optimisation. The framework encompasses a holistic integration of data, software, hardware, and workflow, taking into account the flow of information within cyberspace and materials within physical space. Initiated by the scientist specifications, these flows autonomously evolve across cyber and physical spaces until they accomplish the research goals or exhaust allocated resources. The illustration of the design of experiments was created by ref. 81.

**b** Dynamic knowledge graph approach that is structured into three layers. The first layer represents the real world, where hardware is located and reactions take place. The second layer consists of a dynamic knowledge graph in cyberspace, hosting information such as the digital twin of the hardware and chemical data. The third layer comprises active agents that continually monitor the status of the knowledge graph, dynamically restructuring it, and actuating changes in the real world. The illustration of docker was created using flickr.com.

This architecture liberates the scientists from routine work, however, it also poses challenges in the implementation in terms of ensuring robustness, scalability, maintainability, safety, and ethics. Ideally, the system should enable seamless integration of new devices, resources, and algorithms without disrupting the system's overall functioning. It is also critical to allow for dynamic adaption to changes in research goals and resource restrictions.

We believe dynamic knowledge graph technology can help with realising this architecture<sup>32</sup>. Specifically, as illustrated in Fig. 2b, this technology abstracts the software components as agents that receive inputs and produce outputs. The flow of data between these components is represented as messages exchanged among these agents. Physical entities can be virtualised as digital twins in cyberspace, enabling real-time control and eliminating geospatial boundaries when multiple labs are involved. This reformulation of the closed-loop optimisation problem as information travelling through the knowledge graph and reflecting their changes in the real world offers a powerful framework for achieving true distributed SDLs. In this way, we can think of an occurrence of physical experimentation as a sequence of actions that dynamically generates information about a reaction experiment as it progresses in time, analogous to computational workflows<sup>42</sup>.

This work is part of a series of papers introducing a holistic approach to lab automation by including all aspects of research laboratories (see Fig. 1a) in an all-encompassing digital twin<sup>36</sup>. By employing dynamic knowledge graphs that integrate knowledge models from different domains, we can address the challenges related to interoperability and adaptability commonly encountered in platform-based approaches<sup>32</sup>. The goal-driven architecture facilitates reasoning across the knowledge base, allowing high-level, abstract goals to be decomposed into specific sub-goals and more tangible tasks. Within this framework, humans play a dual role, functioning both as goal setters and operators (when necessary) for executing and intervening in experiments. When acting as operators, humans can be represented in the knowledge graph similarly to robots, and they receive instructions in a human-readable format. This facilitates the realisation of a hybrid and evolving digital laboratory, bridging potential "interim technology gaps"<sup>43</sup>. The operations described in this work are carried out through robotic handling, with humans primarily involved in the preparation of initial materials and the maintenance of the equipment.

#### Chemical ontologies and digital twins

The realisation of SDLs requires a connection between abstract chemistry knowledge and concrete hardware for execution<sup>21</sup>. This calls for a set of connected ontologies, as identified in our previous analysis on the gaps in current semantic representations for chemical digitalisation<sup>32</sup>. Figure 3 presents a selection of concepts and relationships as an effort to address these gaps. These concepts span various levels of abstraction involved in scientific research, ranging from the high-level research goals, through the conceptual level of chemical reactions and the mathematical level of design of experiments, down to the physical execution of reaction experiments and the laboratory digital twin. We describe below ontologies' cross-domain characteristics, for technical details on each ontology please see Supplementary Information section A.1.

For closed-loop optimisation in SDLs, we draw parallels between the pursuit of optimal objectives and the reasoning cycles involved in pursuing a goal<sup>44,45</sup>. The multi-objective problem can be formulated as a GoalSet which comprises individual Goals. Each goal is associated with specified dimensional quantities that can be achieved by a Plan, which consists of multiple Steps to be carried out by corresponding agents. From the implementation perspective, this is akin to a specialised research sub-domain within the scientific workflow community that focuses on the management of iterative workflows abstracted as directed cyclic graphs<sup>46</sup>. In this regard, we adopt the derived information framework<sup>42</sup>, a knowledge-graph-native approach, to manage the iterative workflow.

In developing chemical ontologies for SDLs, we draw upon the lessons learnt in creating ontologies for chemical plants. One prominent example is the OntoCAPE material and chemical process system<sup>47</sup> ontology, which describes materials from three aspects: the ChemicalSpecies that reflects the intrinsic characteristics, Material as part of the phase system which describes macroscopic thermodynamic behaviour, and MaterialAmount that refers to a concrete occurrence of an amount of matter in the physical world. Building on this foundation, we introduce OntoReaction, an ontology that captures knowledge in wet-lab reaction experiments, and OntoDoE, an ontology for the design of experiments (DoE) in optimisation campaigns. As an effort to align with existing data, OntoReaction draws inspiration from established schemas used in chemical reaction databases like ORD<sup>48</sup> and UDM<sup>49</sup>. ReactionExperiment is a concrete realisation of a ChemicalReaction that is sampled at a set of ReactionConditions and measures certain PerformanceIndicators. When grouped together, they can form HistoricalData that are utilised by a  ${\tt DesignOfExperiment}$  study to propose new experiments.

In the development of our hardware ontologies, we have expanded upon concepts from the Smart Applications REFerence (SAREF) ontology<sup>50</sup>, which is widely adopted in the field of the Internet of Things. We introduce OntoLab to represent the digital twin of a laboratory, comprising a group of LabEquipment and ChemicalContainers that contain ChemicalAmount. Furthermore, we create OntoVapourtec and OntoHPLC as ontologies for the equipment involved in this work, linking them to the concrete realisation aspect of OntoCAPE. We establish the link between abstract chemical knowledge and hardware by translating ReactionCondition to ParameterSetting, which can be combined to form EquipmentSettings for configuration.

#### **Contextualised reaction informatics**

By utilising ontologies as blueprints, we can instantiate reaction information while preserving connections to contextual recordings. The reaction we choose for demonstration is an aldol condensation reaction between benzaldehyde **1** (bold numbers for reference) and acetone **2**, catalysed by sodium hydroxide **3** to yield the target product benzylideneacetone **4**<sup>51</sup>, which is pharmaceutically relevant and can be used to treat idiopathic vomiting as an NK-1 receptor inhibitor<sup>52</sup>. Additionally, reported side products include dibenzylideneacetone **5** and further condensation products from acetone polymerisation. The choice of this well-studied reaction is deliberate, aimed at explaining the contribution of our work to developing distributed SDLs to a broad audience, and an application to more interesting chemistry will be presented in a subsequent paper.

Figure 4 provides an illustrative representation of the chosen reaction in the knowledge graph as viewed through various roles within a laboratory, each with its unique perspective on the same chemical. Taking the starting material benzaldehyde as an example, it demonstrates how a knowledge graph can enhance the daily work of different roles. A chemist, more interested in conceptual description, might look at benzaldehyde as a reactant and search for relevant species information. A data scientist might examine its concentration to determine the appropriate usage of other chemicals when designing conditions for a particular reaction experiment. Meanwhile, the 3D digital twin built on top of the knowledge graph offers a lab manager a centralised hub for real-time monitoring of lab status<sup>53</sup>, ensuring the availability of an internal standard that can be mixed with the physical existence of benzaldehyde to enable characterisation during the actual execution of the experiment. In practice, the same individual might play several roles, and the emphasis here is on the cross-domain interoperability facilitated by the amalgamation of different aspects into a unified knowledge graph. This integration ensures the relevance of information to a diverse range of users while maintaining human oversight. Consequently, this approach may present opportunities for the enhancement of various digital applications, such as the utilisation of virtual reality for laboratory training<sup>54</sup>.

The integration of chemical knowledge from PubChem, represented by OntoSpecies for unique species identification<sup>55</sup>, serves as a critical link between these facets of chemicals. It enables the identification of potential input chemicals based on the reactant and solvent during DoE and allows for the selection of appropriate sources of starting materials from multiple chemical containers (see Supplementary Information section A.2). Another aspect enabled by this disambiguation of species relates to the representation of chemical impurities. In this case study, all starting materials were procured and used as received, with purities exceeding 99% for liquid chemicals and 97% for NaOH pellets (see Supplementary Table S4). The impurities are categorised as unknown components, and their presence is indicated using the data property



Fig. 3 | A selection of concepts and relationships capturing different aspects in self-driving laboratories (SDLs). The concepts are categorised based on their level of abstraction, spanning from high-level research goals to conceptual descriptions of chemical reactions and the mathematical expression of design of experiments, as well as the physical execution of reaction experiments and the laboratory digital

twin. These concepts are interlinked with the OntoCAPE Material System, representing an effort to enhance interoperability with the community initiatives. Their namespaces correspond to the colour coding. For complete knowledge representation and namespace definitions see Supplementary Information section A.1.1.

OntoLab:containsUnidentifiedComponent for OntoLab:-ChemicalAmount, a concept used for representing the concrete appearances of chemicals in the physical world. In terms of the collected reaction products, this representation is employed to signify the existence of (at least one) OntoHPLC:ChromatogramPoint in the OntoHPLC:HPLCReport that is designated OntoHPLC:unidentified. A more comprehensive representation of impurities can be achieved in conjunction with concentrationrelated concepts, such as OntoCAPE:Molarity, which we shall incorporate in future work. For concrete examples of ontology instantiation see Supplementary Information section A.1.

#### Goal-driven knowledge dynamics

Figure 5 presents a high-level overview of the goal-driven evolution of the knowledge graph during closed-loop optimisation. The dynamicity of the knowledge graph is enabled by the presence of software agents that realise each component of the distributed architecture and facilitate the flow of information within the graph. The process begins with the goal derivation stage where the scientist initiates a goal request. The Reaction Optimisation Goal (ROG) Agent translates this request into a machine-readable statement that captures the scientist's intention. To accommodate all objectives, a goal set is formulated that considers each objective as a reward function for the agents'





availability. **d** The knowledge graph representation puts chemical informatics into context, allowing for queries and answers across these varied layers of abstraction (views). The colour coding corresponds to the ontological expression.



**Fig. 5** | **Autonomous workflow triggered in response to goal requests from scientists as information travels within the knowledge graph. a** The Reaction Optimisation Goal (ROG) Agent translates the specification of a scientist into a machine-readable statement and instantiates it into the knowledge graph. b The Reaction Optimisation Goal Iteration (ROGI) Agent initiates the design-make-test-

analyse cycle, during which other agents query information from the digital twin and actuate the hardware. **c** The progress of goal pursuit is assessed after each iteration, determining whether to proceed to the next cycle. Steps (**b**) and (**c**) are iterated until either goals are achieved or resources are depleted.

operations. For each participating laboratory, a Goal Iteration Derivation instance is created using the derived information framework<sup>42</sup> and requested for execution by the Reaction Optimisation Goal Iteration (ROGI) Agent.

The goal iteration stage plays a central role in the evolution of the dynamic knowledge graph. It involves the ROGI Agent initiating the flow of information among the participating agents towards achieving the goals. This process begins with the ROGI Agent creating tasks for the corresponding agents according to the DMTA cycle, including the DoE Agent, Schedule Agent, and Post-Processing Agent. The DoE Agent perceives the knowledge graph to retrieve prior data and chemical stock available for experiments and then proposes a new experiment. The Schedule Agent evaluates the hardware available in the specified laboratory according to the proposed conditions and subsequently selects the most appropriate hardware to execute the experiment. This is accomplished by generating tasks for the agents responsible for managing the selected digital twin. These agents actuate the equipment to perform reaction and characterisation in the physical world. When the HPLC report is generated, the Post-Processing Agent analyses the chromatogram data to calculate the objectives.

During the third stage, the ROG Agent utilises the obtained results to determine whether the next iteration should be pursued. To do so, it checks if the Pareto front of the multi-objective fulfils the pre-defined goals and if the resources are still available. The reaction experiment performed in the current iteration then becomes historical data, serving as input for the succeeding round of the Goal Iteration Derivation across all participating SDLs. Afterwards, a new request will be made to the ROGI Agent to start a new iteration, forming a selfevolving feedback loop. To ensure correct data dependencies and the order of task execution, we employed the derived information framework<sup>42</sup> to manage the iterative workflow. We implemented each software agent using the derivation agent template provided by the framework. Once deployed, these agents autonomously update the knowledge graph to actively reflect and influence the state of the world.

This approach enables flexibility and extensibility in the system. As the digital twin of each lab is represented as a node in the knowledge graph, new hardware can be added or removed during the optimisation campaign by simply modifying the list of participating laboratories. The experimental allowance can also be updated when more chemicals become available. The system also supports data sharing across organisations at the very moment the data are generated. Details on the internal logic and technical aspects of the agents in the knowledge graph implementation are available in the Supplementary Information section A.2.

#### Collaborative closed-loop optimisation

To demonstrate the scalability and modularity, the knowledge graph approach was applied to a real-time collaborative closed-loop optimisation distributed over two SDLs in Cambridge and Singapore. The objectives selected are run material cost and yield that were sampled for a search space of molar equivalents (relative to benzaldehyde **1**) of acetone **2**, NaOH **3**, residence time and reaction temperature. The research goals and restrictions were populated in the knowledge graph via a web front end. As no prior experimental data was provided, the agents start experiments with random conditions and gradually update their beliefs using TSEMO algorithm<sup>56</sup>. Before running the optimisation, two labs were verified to produce consistent results for





**Fig. 6** | **Objectives and design variables of experiments conducted in the closed-loop optimisation campaign in distributed self-driving laboratories (SDLs).** Each dot refers to a single run. The animation of the optimisation progress is available in Supplementary Movie 1. Interactive versions of 3D plots are available in Supplementary Movies 2 and 3 for cost and yield objectives, respectively. Source data are provided as a Source Data file. **a** Pareto front plot of the yield and cost

two control conditions, in line with the practice of Shields et al.<sup>57</sup>. For experimental details see Supplementary Information section A.3.

Figure 6a presents the cost-yield objectives consisting of 65 data points collected during the self-optimisation. Throughout the operation, two SDLs share the results with each other when proposing new experimental conditions. The real-time collaboration demonstrated faster advances in the Pareto front with the highest yield of 93%. The chemicals used in this study were obtained from different vendors compared to ref. 51, the cost is therefore not directly comparable due to different prices. Although not considered in the optimisation, the environment factor and space-time yield were found to be highly



objectives for the aldol condensation reaction collaboratively optimised by two distributed SDLs. **b** Three-dimensional plot of the four sampled design variables colour coded for run material cost during the closed-loop optimisation. The size of the dots denotes the molar equivalents of 3 in each run. **c** Three-dimensional plot of the four sampled design variables colour coded for yield during the closed-loop optimisation. The size of the dots denotes the molar equivalents of 3 in each run.

correlated to the yield objective. The best values obtained are 26.17 and 258.175 g L<sup>-1</sup> h<sup>-1</sup> when scaled to the same benzaldehyde injection volume (5 mL), both outperformed the previous study<sup>S1</sup>.

Figure 6b, c illustrate the influence of the continuous variables on the cost and yield objectives, with their interactive versions as Supplementary Movie 2 and 3, respectively. The cost is calculated to count for the molar amount of input chemicals sourced from the pumps for the reaction. Therefore, it increases linearly with the molar equivalents of the starting materials. Similarly as identified by ref. 51, reaction temperature has a positive correlation with the yield of reaction, whereas the residence time shows a poor correlation. Upon examination of the molar equivalent of acetone **2**, it can be observed that its further increase after 30 results in a reduction in yield. This decrease can be attributed to the formation of more side product **5** and other further condensation products of acetone and benzaldehyde.

Notably, the Singapore setup encountered an HPLC failure after running for approximately 10 h. This caused peak shifting of the internal standard which resulted in a wrongly identified peak that gives more than 3500% yield. This point is considered abnormal by the agents and therefore not utilised in the following DoE. An email notification was sent to the developer for maintenance which took the hardware out of the campaign. The asynchronous and distributed design enabled the Cambridge side to further advance the Pareto front for the cost-yield trade-offs. It is also notable that the product peak was missed for one run at the Cambridge side due to a small shift of the peak which gives a yield of 0%. This point was taken into consideration in the DoE, but fortunately, it did not affect the final Pareto front as the corrected yield is still Pareto-dominated. The optimisation campaign was stopped since no more significant improvement was observed in terms of hypervolume, and also due to requests for repurposing the equipment for other projects. The complete provenance records (knowledge graph triples) are provided as Supplementary Data, along with an interactive animation of the optimisation progress extracted from them as Supplementary Movie 1.

## Discussion

In this contribution, we presented a dynamic knowledge graph approach to realise a conceptual architecture for distributed SDLs. We developed ontologies to represent various aspects of chemical knowledge and hardware digital twins involved in a closed-loop optimisation campaign. By employing autonomous agents as executable knowledge components to update and restructure the knowledge graph, we have enabled collaborative management of data and material flow across SDLs. Our approach allows scientists to initiate the autonomous workflow by setting up a goal request, which triggers the flow of information through the knowledge graph as the experimentation workflow progresses.

As a proof-of-concept demonstration, we applied the system to an aldol condensation reaction using two setups across different parts of the globe. Despite the differences in configurations, the reaction data produced by both machines were interoperable owing to the layered knowledge abstraction. Throughout the experiment, the system recorded all data provenance as the knowledge graph evolved autonomously, providing opportunities for informed machine learning<sup>58</sup>. Our collaborative approach resulted in faster data generation and advanced the Pareto front while exhibiting resilience to hardware failure.

The implementation of this work has provided valuable insights and identified areas for future improvement in the realm of dynamic knowledge graph systems. In terms of orchestration, it is crucial for the system to be robust to network disruption since it is distributed over the internet. We have implemented measures to ensure that agents deployed in the lab can handle internet cut-offs and resume operations once back online. To minimise downtime during reconnection, future developments could provide on-demand, localised deployment of critical parts of the knowledge graph to sustain uninterrupted operation.

For efficient optimisation and data quality, it is critical to have control conditions in place when adding new setups to the network, and only those generated results within the tolerance should be approved. Complex reactions with high-dimensional domains may not be sufficiently evaluated using only two control conditions. This highlights the persisting challenges in maintaining data quality and opens avenues for incorporating strategic cross-workflow validation experiments. Further development could also be made to federate the SDLs, where each lab hosts its data and digital twins locally and only exposes its capabilities in the central registry (a "yellow page") without revealing confidential information. An authentication and authorisation mechanism should be added to control access to the equipment and grant permission for federated learning.

When reflecting on the vision of distributed SDLs, our approach exhibits both commonalities and distinctions when compared to contemporary designs. Table 1 summarises the key design features, to the best of our knowledge, as they relate to the three major challenges, with the first challenge further divided into the abstraction of resources and workflow coordination.

In terms of resource abstraction, all approaches (including the one presented in this work) employ a modular design that considers hardware limitations in granularity. This modularity is key for a seamless integration of new resources into a plug-and-play system. However, the way resources are exposed to the coordinator varies and this significantly impacts the orchestration of workflows across laboratories. This applies to both workflow template encoding and its actual execution. The dynamic knowledge graph approach uses agents acting as lab resource wrappers with knowledge graph access. Agents can register for jobs and proactively execute tasks assigned to the digital twin of the resources they manage. This approach is preferable compared to the practices in the remote procedure call paradigm, where lab resources are made accessible as web servers. Based on our experience, it can raise concerns among IT staff when exposing resources across university or company firewalls. Similar to agents, our approach encodes the workflow in the knowledge graph with each step overseen by an agent. Compared to encoding workflows as a sequence of function calls in scripting languages (such as Python), where execution may struggle with asynchronous workflows evolving during optimisation, our approach allows for real-time workflow assembly and modification. For a detailed technical discussion, interested readers can refer to the derived information framework<sup>42</sup>.

The integration of data serialisation and storage within workflow aims to ease community adoption. As seen in Table 1, practices range from transmitting diverse file formats to enforcing a unified data representation. Starting with ad hoc extraction-transformationloading tools for new devices prototyping is practical and minimally disruptive when upgrading a single lab. However, we find this approach less effective for scaling up to a large network of SDLs<sup>32</sup>. This limitation is the driving force behind the development of the dynamic knowledge graph approach, despite the initial cost required for creating ontologies that capture a collective understanding of the field. Our design delegates the responsibility of digesting and translating ontologies into the requisite language and file formats to autonomous agents. Compared to adopting a central coordinator to handle data transfer and format translation, our approach emphasises information propagation within a unified data layer, obviating the need for peer-to-peer data transfer and alleviating network congestion. Drawing an analogy to self-driving cars, once the "driving rules" (ontologies) are learned, SDLs are granted permission to drive on the "road" (information flow). Compared to traditional relational databases used in other studies, where schema modification can be challenging, the open-world assumption inherent in the dynamic knowledge graph enhances its extensibility. Organising concepts and relationships within a knowledge graph is also more intuitive than

lable 1	Comparison between contemporary d	ssigns and this work towards the realisatio	in of distributed self-driving laboratorio	es (SDLs)
Reference	Resource abstraction	Workflow orchestration	Data serialisation and storage	Experimental provenance
SiLA <sup>28</sup>	Hardware functions are abstracted as SILA Features following a micro-service architecture with their behaviour described as a state machine.	A sequence of function calls using gRPC and HTTP/2 protocols.	AnIML <sup>28</sup> is employed as a file-less medium for bidir- ectional analytical data transmission between laboratory information management systems and chromatography data systems.	Device metadata collected during measurements are stored in XML files.
Chem OS <sup>76</sup>	Both software and hardware are abstracted as SiLA servers. Time-consuming computational jobs are managed by AiiDA <sup>77</sup> on SLURM <sup>41</sup> .	The central coordinator executes a sequence of Python function calls. The coordinator is also responsible for creating job files in the format required by each hardware/ software.	Data are streamed between the central coordinator and each device in diverse file formats, e.g., pickle object, JSON, and CSV. The storage is done in an internal database with a schema consisting of device- agnostic and device-specific tables.	Job execution logs are stored with timestamps in the database table corresponding to each device.
ESCALATE	<sup>24</sup> The Django framework is used to abstract resour- ces as REST API endpoints.	A sequence of steps encoded for "ExperimentTemplate" and accessible via REST API endpoints.	Data are stored in a PostgreSQL database and served via Django REST API <sup>78</sup> endpoints that serialise the data into JSON format for web transfer and inspection.	Metadata associated with the execution of work- flow steps are stored with experiment instances in the relational database.
HALEO <sup>25</sup>	Device and their functions ("actions") are repre- sented as hierarchical and asynchronous FastAPI <sup>78</sup> web servers.	A central coordinator executes a sequence of API calls (wrapped as Python functions).	Data are recorded as "groups" and "datasets" in the HDF5 file format and deposited into institutional repositories.	Metadata of "actions" are recorded in the same HDF5 file as the experimental measurements.
χDL <sup>26,80</sup>	Hardware is categorised/abstracted based on the unit operations in chemical reactions that it can execute.	The sequence of synthesis steps is expressed in XML format, which is later compiled into machine-actionable instruc- tions in a Python script.	The analysis reports are kept in their native file format and bundled with the synthesis description in a PostgreSQL database.	Hardware instructions, the actual performed actions, and other metadata are stored with experiment instances in the relational database.
This work	Hardware is virtualised as a digital twin in a knowledge graph, where its control interface, akin to software resources, is wrapped using the derivation agent template.	DMTA cycles are expressed as directed acyclic graphs of "derivations" in the knowledge graph each referring to a step managed by the respective software agent.	Data are expressed in ontological format (triples) wherever possible. Files (e.g., CSV and XLS) are stored on a file server. Their ontological translation and pointers to the file server location are stored in the triple store.	The inputs/outputs annotation of each step in the workflow is recorded by the derived information framework as triples. The detailed operation timing is not recorded owing to API limitations in obtaining information at this level of granularity.

traditional tabular structures. However, this flexibility may come at the cost of performance issues when handling extensive data volumes, especially when dealing with data on the scale of ORD. To counter this, technologies such as ontology-based data access<sup>59</sup> can create a virtual knowledge graph from relational databases, combining the strengths of both approaches.

Our approach to experimental provenance differs from others due to hardware constraints. It focuses less on exact operation timing, such as robotic arm motions, and more on capturing inputs and outputs within DMTA cycles. This facilitates high-level analysis, enabling answering questions like "which experiments from lab A informed the DoE study for a specific reaction in lab B". This capability has been effectively demonstrated in the interactive Pareto progress animation provided in Supplementary Movie 1. However, for a deeper understanding of epistemic uncertainties associated with operations in complex reactions, it is imperative to expand the ontologies for a more granular abstraction of the experimental procedures. A potential expansion in this regard could involve the ontologisation of  $\chi$ DL.

Looking forward, achieving a globally collaborative research network requires collective efforts. As the knowledge graph aims to reflect a communal understanding of the field, involving different stakeholders early on can accelerate collaboration and increase the chance of success. Specifically, there exists an opportunity for using knowledge graph technology as an integration hub for all aforementioned initiatives. Industrial partners are encouraged to work together and provide a unified API for interacting with their proprietary software and hardware interfaces. This can be facilitated by efforts such as OPC UA<sup>60</sup> and SiLA<sup>28</sup>. Recent studies have shown the successful exchange of HPLC methods between vendors in the Chromatography Data System (CDS), demonstrating the potential for the ontology-based approach<sup>61</sup>. Collaboration between scientists and industry is also important at various stages of research and development<sup>62</sup>.

Overall, we believe the dynamic knowledge graph approach demonstrated in this work provides the first evidence of its potential to establish a network of globally distributed SDLs. Although we focus on flow chemistry in this study, the principles are generic. The same approach can be applied to DMTA cycles for other domains should relevant ontologies and agents be made available, for example, to support research in deep space<sup>18</sup>.

## Methods

#### The World Avatar knowledge graph

This work follows the best practices in the World Avatar project. All ontologies and agents are version-controlled on GitHub. We provide our thought process during the development below. The same principles can be followed for self-optimisation applications in other domains.

Ontology development. Developing ontologies is often an iterative process and it is not a goal in and of itself<sup>63</sup>. As suggested in<sup>32,36,37</sup>, we follow the steps from specifying target deliverables to conceptualising relevant concepts and finally implementing codes for queries. Aimed at capturing data and material flow in distributed SDLs, the relevant concepts range from the reaction experiment to the hardware employed to conduct it. In the World Avatar, ontologies are typically developed to be digested by software agents which mimic the human way of conducting different tasks<sup>64</sup>. Therefore, the development draws inspiration from relevant software tools<sup>51,65,66</sup> and existing reaction database schemas<sup>48,49</sup>. Views of the domain experts<sup>67-69</sup> are also consulted to better align with the communal understanding of the subject. During iterations, competency questions are used to test if the ontologies meet case study requirements. The answers to these questions are provided in the form of SPARQL queries that are executed by the agents during their operations. Another essential aspect to consider is data instantiation, where we adopted pydantic to

simplify the querying and processing of data from the knowledge graph. Overall, the ontology development process starts as easily as drawing concepts and their relationships on a whiteboard and then gradually materialising them in code.

Agent development. Following the development of ontologies, agents are defined as executables that process inputs and generate outputs. Their I/O signatures are represented following OntoAgent<sup>70</sup>. At the implementation level, all agents inherit the DerivationAgent template in Python provided by the derived information framework<sup>42</sup>. Specifically, agents utilise the asynchronous communication mode when interacting with the knowledge graph as conducting experiments is inherently a time-consuming process. Each of the agents monitors the jobs assigned to itself and records the progress of execution in the knowledge graph. The derived information framework does most of the work behind the scenes, leaving the developer with the only task of implementing each agent's internal logic. As agents modify the knowledge graph and subsequently actuate the real world autonomously once active, it is important to make sure they behave as expected. In this regard, unit and integration tests are provided to help with responsible development. For instance, the integration tests in folder RxnOptGoalAgent/tests simulate the behaviour of distributed SDLs to verify that the data flows are as expected upon goal request from scientists. Detailed descriptions of tests for each agent can be found in section A.2 of the Supplementary Information.

Distributed deployment. Taking inspiration from remote control practices in lab automation<sup>71-73</sup>, the knowledge graph is designed to span across the internet. It follows deployment practices commonly used by cloud-native applications and is implemented through docker containers. The triplestore and file server containing the knowledge statements are deployed at internet-resolvable locations. Depending on capabilities, agents are located at different host machines. Those who monitor and control the hardware are deployed in the corresponding laboratory for security reasons. They transmit data collected from the hardware to the knowledge graph and in reverse configure and actuate the equipment when a new experiment arises. At start-up, agents register their OntoAgent instances in the knowledge graph, then act autonomously should tasks be assigned to them. Altogether, these agents form a distributed network that facilitates the transfer of information within the knowledge graph and bridges cyberspace and the physical world.

## Flow chemistry platforms

This work connects two similar automated flow chemistry platforms located in Cambridge and Singapore. The method of sourcing input chemicals differs, with a liquid handler employed in Cambridge and reagent bottles utilised in Singapore. We provide below brief descriptions of the experimental setup. All chemicals were used as received.

**Cambridge lab.** On the Cambridge side, the experimental setup consists of two Vapourtec R2 pump modules, one Vapourtec R4 reactor module, one Gilson GX-271 liquid handler, one four-way VICI switching valve (CI4W.06/.5 injector), and Shimadzu CBM-20A HPLC analytical equipment equipped with Eclipse XDB-C18 column (Agilent part number: 993967-902). To initiate the reaction, the liquid handler dispenses a 2 mL solution of 0.5 M benzaldehyde **1** dissolved in acetonitrile (with 0.06 M biphenyl as an internal standard) into the sample loop of pump A. Acetone **2** (50% v/v in acetonitrile) and 0.1 M NaOH **3** in ethanol are similarly loaded into sample loops for pump B and C. After being transferred by the switching valve, the product (benzylideneacetone **4**) is analysed using online HPLC. The HPLC analysis lasts 17 min, with a mobile phase consisting of an 80:20 (v/v) binary mixture of water and

Singapore lab. On the Singapore side, the experimental setup consists of two Vapourtec R2 pump modules, one Vapourtec R4 reactor module, one 6-port 2-position VICI switch valve equipped with 60 nL sampling rotor, and an Agilent 1260 Infinity II system equipped with a G1311B quaternary pump, Eclipse XDB-C18 column (Agilent product number: 961967-302), and G1314F variable wavelength detector (VWD). The input chemical for the reaction is sourced from three reagent bottles that are directly attached to the Vapourtec pumps: pump A contains 0.5 M benzaldehyde 1 in acetonitrile (with 0.05 M naphthalene as an internal standard), pump B contains 6.73 M acetone 2 in acetonitrile (50% v/v in acetonitrile), and pump C contains 0.1 M NaOH 3 in ethanol. The following HPLC quaternary pump method for online HPLC is used: the initial mobile phase was a 5:95 (v/v) binary mixture of acetonitrile and water flowing at 0.2 mL min<sup>-1</sup>. Immediately after sample injection, the flow rate and ratio of acetonitrile to water were steadily changed to  $1 \text{ mL min}^{-1}$  and 95:5 (v/v) during the first 5 min. At a flow rate of 1 mL min<sup>-1</sup>, the binary mixture ratio is then returned to 5:95 (v/v) acetonitrile:water over 1.5 min in a linear gradient. This binary mixture ratio is held constant at 1 mL min<sup>-1</sup> for the next 1.5 min, after which the analysis is complete (after a total of 8 min), and the method returns to a flow rate of 0.2 mL min<sup>-1</sup>. The VWD wavelength was changed over the 8 min analysis time as follows: the absorption wavelength is 248 nm for the initial 6.05 min and then switched to 228 nm until the end of acquisition.

#### **Reporting summary**

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## **Data availability**

Research data generated in this study has been deposited in the University of Cambridge data repository under accession code https://doi.org/10.17863/CAM.97058<sup>74</sup>. Source data are provided with this paper. The tabular format of relevant experimental results that were displayed in Fig. 6 is provided in the Source Data XLSX file. Source data are provided with this paper.

## **Code availability**

All the codes developed are publicly available on The World Avatar GitHub repository https://github.com/cambridge-cares/ TheWorldAvatar or the Zenodo repository at https://doi.org/10.5281/ zenodo.10151236<sup>75</sup>. The docker images of agents are available at GitHub's public registry located at ghcr.io/cambridge-cares/: doe\_agent:1.2.0, vapourtec\_schedule\_agent:1.2.0, vapourtec\_agent:1.2. 0, hplc\_agent:1.2.0, hplc\_postpro\_agent:1.2.0, rxn\_opt\_goal\_iter\_agent: 1.2.0, and rxn\_opt\_goal\_agent:1.0.0. The deployment instructions can be found in folder TheWorldAvatar/Deploy/pips.

## References

- Häse, F., Roch, L. M. & Aspuru-Guzik, A. Next-generation experimentation with self-driving laboratories. *Trends Chem.* 1, 282–291 (2019).
- 2. Abolhasani, M. & Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2**, 483–492 (2023).
- Merrifield, R. B., Stewart, J. M. & Jernberg, N. Instrument for automated synthesis of peptides. *Anal. Chem.* 38, 1905–1914 (1966).
- Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science 365, eaax1566 (2019).
- 5. Steiner, S. et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, eaav2211 (2019).

## https://doi.org/10.1038/s41467-023-44599-9

# Article

- Burger, B. et al. A mobile robotic chemist. *Nature* 583, 237–241 (2020).
- Chatterjee, S., Guidi, M., Seeberger, P. H. & Gilmore, K. Automated radial synthesis of organic molecules. *Nature* 579, 379–384 (2020).
- Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* 3, 5–20 (2018).
- 9. Zhu, Q. et al. An all-round Al-chemist with a scientific mind. *Natl. Sci. Rev.* **9**, nwac190 (2022).
- 10. King, R. D. et al. The automation of science. *Science* **324**, 85–89 (2009).
- Elder, S. et al. Cross-platform Bayesian optimization system for autonomous biological assay development. *SLAS Technol.* 26, 579–590 (2021).
- Yang, H. et al. Automatic strain sensor design via active learning and data augmentation for soft machines. *Nat. Mach. Intell.* 4, 84–94 (2022).
- Stach, E. et al. Autonomous experimentation systems for materials development: a community perspective. *Matter* 4, 2702–2726 (2021).
- Delgado-Licona, F. & Abolhasani, M. Research acceleration in selfdriving labs: technological roadmap toward accelerated materials and molecular discovery. *Adv. Intell. Syst.* 5, 2200331 (2022).
- Leins, D. A., Haase, S. B., Eslami, M., Schrier, J. & Freeman, J. T. Collaborative methods to enhance reproducibility and accelerate discovery. *Digital Discov.* 2, 12–27 (2023).
- Seifrid, M., Hattrick-Simpers, J., Aspuru-Guzik, A., Kalil, T. & Cranford, S. Reaching critical MASS: crowdsourcing designs for the next generation of materials acceleration platforms. *Matter* 5, 1972–1976 (2022).
- Ren, Z., Ren, Z., Zhang, Z., Buonassisi, T. & Li, J. Autonomous experiments using active learning and Al. *Nat. Rev. Mater.* 8, 563–564 (2023).
- Sanders, L. M. et al. Biological research and self-driving labs in deep space supported by artificial intelligence. *Nat. Mach. Intell.* 5, 208–219 (2023).
- Pyzer-Knapp, E. O. et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput. Mater.* 8, 84 (2022).
- Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous discovery in the chemical sciences Part II: outlook. *Angew. Chem. Int. Ed.* 59, 23414–23436 (2020).
- Wilbraham, L., Mehr, S. H. M. & Cronin, L. Digitizing chemistry using the chemical processing unit: from synthesis to discovery. Acc. *Chem. Res.* 54, 253–262 (2021).
- 22. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
- Roch, L. M. et al. ChemOS: orchestrating autonomous experimentation. Sci. Robot. 3, eaat5559 (2018).
- Pendleton, I. M. et al. Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): a software pipeline for automated chemical experimentation and data management. *MRS Commun.* 9, 846–859 (2019).
- Rahmanian, F. et al. Enabling modular autonomous feedback-loops in materials science through hierarchical experimental laboratory automation and orchestration. *Adv. Mater. Interfaces* 9, 2101987 (2022).
- 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* **370**, 101–108 (2020).
- Hein, J., Rauschen, R., Guy, M. & Cronin, L. Universal chemical programming language for robotic synthesis reproducibility research square platform LLC. https://doi.org/10.21203/rs.3.rs-2761997/v1 (2023).

- Schäfer, B. Data exchange in the laboratory of the future a glimpse at AnIML and SiLA (2018) https://doi.org/10.1002/gitlab.17270/full/. accessed 30 May 2023.
- 29. Mitchell, S. N. et al. FAIR data pipeline: provenance-driven data management for traceable scientific workflows. *Philos. Trans. R.* Soc. A **380**, 20210300 (2022).
- Statt, M. J. et al. The materials experiment knowledge graph. *Digital Discov.* 2, 909–914 (2023).
- Menon, A., Krdzavac, N. B. & Kraft, M. From database to knowledge graph-using data in chemistry. *Curr. Opin. Chem. Eng.* 26, 33–37 (2019).
- 32. Bai, J. et al. From platform to knowledge graph: evolution of laboratory automation. *JACS Au* **2**, 292–309 (2022).
- 33. Hogan, A. et al. Knowledge graphs. ACM Comput. Surv. 54, 1–37 (2022).
- Akroyd, J., Mosbach, S., Bhave, A. & Kraft, M. Universal Digital Twin - A Dynamic Knowledge Graph. *Data-Centric Eng.* 2, e14 (2021).
- Lim, M. Q., Wang, X., Inderwildi, O. & Kraft, M. The World Avatar A World Model for Facilitating Interoperability. in *Intelligent Decarbonisation: Can Artificial Intelligence and Cyber-Physical Systems Help Achieve Climate Mitigation Targets*?, 39–53 (Springer International Publishing, 2022).
- Rihm, S. D. et al. The Digital Lab Framework as Part of The World Avatar Preprint at https://como.ceb.cam.ac.uk/preprints/ 314/ (2023).
- Kondinski, A., Bai, J., Mosbach, S., Akroyd, J. & Kraft, M. Knowledge engineering in chemistry: from expert systems to agents of creation. Acc. Chem. Res. 56, 128–139 (2023).
- Kitano, H. Nobel turing challenge: creating the engine for scientific discovery. npj Syst. Biol. Appl. 7, 1–12 (2021).
- Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous discovery in the chemical sciences part I: progress. *Angew. Chem. Int. Ed.* 59, 22858–22893 (2020).
- 40. Taylor, C. J. et al. A brief introduction to chemical reaction optimization. *Chem. Rev.* **123**, 3089–3126 (2023).
- Yoo, A. B., Jette, M. A. & Grondona, M. SLURM: Simple Linux Utility for Resource Management. in Job Scheduling Strategies for Parallel Processing, 44–60 (Springer Berlin Heidelberg, 2003).
- Bai, J. et al. A derived information framework for a dynamic knowledge graph and its application to smart cities. *Future Gener. Comput. Syst.* **152**, 112–126 (2024).
- Holland, I. & Davies, J. A. Automation in the life science research laboratory. Front. Bioeng. Biotechnol. 8, 571777 (2020).
- Rao, A. S. & Georgeff, M. P. Modeling rational agents within a BDI-Architecture. in Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning, KR'91, 473-484 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991).
- 45. Stein, H. S. & Gregoire, J. M. Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chem. Sci.* **10**, 9640–9649 (2019).
- 46. Krämer, M., Würz, H. M. & Altenhofen, C. Executing cyclic scientific workflows in the cloud. J. Cloud Comput. **10**, 1–26 (2021).
- Morbach, J., Yang, A. & Marquardt, W. OntoCAPE a large-scale ontology for chemical process engineering. *Eng. Appl. Artif. Intell.* 20, 147–161 (2007).
- 48. Kearnes, S. M. et al. The open reaction database. *J. Am. Chem.* Soc. **143**, 18820–18826 (2021).
- 49. Pistoia Alliance. Unified Data Model (2020) https://github.com/ PistoiaAlliance/UDM. accessed 30 May 2023.
- Daniele, L., Garcia-Castro, R., Lefrançois, M. & Poveda-Villalon, M. SAREF: The Smart Applications REFerence ontology (2020) https:// saref.etsi.org/core/v3.1.1/. Accessed 21 Feb 2023.

- Article
- Jeraal, M. I., Sung, S. & Lapkin, A. A. A machine learningenabled autonomous flow chemistry platform for process optimization of multiple reaction metrics. *Chem. Methods* 1, 71–77 (2021).
- Park, J. S., Hu, P., Lin, Y. & Reinsalu, L. A. Composition and Method for Preventing, Reducing, Alleviating or Treating Idiopathic Vomiting (2020) https://testpubchem.ncbi.nlm.nih.gov/patent/US-10548935-B2. US Patent 10,548,935. Accessed 11 Feb 2023.
- Quek, H. Y. et al. BIM-GIS Integration: Knowledge Graphs in a World of Data Silos Preprint at https://como.ceb.cam.ac.uk/preprints/ 311/ (2023).
- 54. Dreyer, J. A. et al. Digitalisering af Forskning og Undervisning på DTU Kemiteknik. *Dan. Kemi* **104**, 6–11 (2023).
- Pascazio, L. et al. Chemical species ontology for data integration and knowledge discovery. J. Chem. Inf. Model. 63, 6569–6586 (2023).
- Bradford, E., Schweidtmann, A. M. & Lapkin, A. Efficient multiobjective optimization employing gaussian processes, spectral sampling and a genetic algorithm. *J. Glob. Optim.* **71**, 407–438 (2018).
- 57. Shields, B. J. et al. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).
- von Rueden, L. et al. Informed machine learning a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Trans. Knowl. Data Eng.* **35**, 614–633 (2021).
- 59. Calvanese, D. et al. Ontop: answering SPARQL queries over relational databases. *Semant. Web* **8**, 471–487 (2016).
- OPC Foundation. Unified Architecture (2023) https:// opcfoundation.org/about/opc-technologies/opc-ua/ accessed 8 March 2023.
- Pistoia Alliance. Update from the Pistoia Alliance's Methods Hub Project (2022) https://www.pistoiaalliance.org/methods/april-2022-methods-database-hplc-uv-methods/ accessed 10 Feb 2023.
- 62. Christensen, M. et al. Automation isn't automatic. *Chem. Sci.* **12**, 15473–15490 (2021).
- 63. Staab, S. & Studer, R. Handbook on Ontologies (Springer Science & Business Media, 2010).
- Kondinski, A. et al. Automated rational design of metal-organic polyhedra. J. Am. Chem. Soc. 144, 11713–11728 (2022).
- 65. Guo, J. et al. Automated chemical reaction extraction from scientific literature. J. Chem. Inf. Model. **62**, 2035–2045 (2021).
- Felton, K. C., Rittig, J. G. & Lapkin, A. A. Summit: benchmarking machine learning methods for reaction optimisation. *Chem. Meth*ods 1, 116–122 (2021).
- Plutschack, M. B., Pieber, B., Gilmore, K. & Seeberger, P. H. The Hitchhiker's guide to flow chemistry. *Chem. Rev.* 117, 11796–11893 (2017).
- 68. Garud, S. S., Karimi, I. A. & Kraft, M. Design of computer experiments: a review. *Comput. Chem. Eng.* **106**, 71–95 (2017).
- 69. Clayton, A. D. et al. Algorithms for the self-optimisation of chemical reactions. *React. Chem. Eng.* **4**, 1545–1554 (2019).
- Zhou, X., Eibeck, A., Lim, M. Q., Krdzavac, N. B. & Kraft, M. An agent composition frramework for the J-Park simulator - a knowledge graph for the process industry. *Comput. Chem. Eng.* **130**, 106577 (2019).
- Fitzpatrick, D. E., Maujean, T., Evans, A. C. & Ley, S. V. Across-theworld automated optimization and continuous-flow synthesis of pharmaceutical agents operating through a Cloud-Based Server. *Angew. Chem. Int. Ed.* 57, 15128–15132 (2018).
- Ramírez, J. et al. A virtual laboratory to support chemical reaction engineering courses using real-life problems and industrial software. *Educ. Chem. Eng.* **33**, 36–44 (2020).
- CMU Cloud Lab (2023) https://cloudlab.cmu.edu/ accessed 12 May 2023.

- Bai, J. et al. Research data supporting "A Dynamic Knowledge Graph Approach to Distributed Self-Driving Laboratories". Apollo -University of Cambridge Repository https://doi.org/10.17863/CAM. 97058 (2023).
- Bai, J. et al. A Dynamic Knowledge Graph Approach to Distributed Self-Driving Laboratories. *Zenodo* https://doi.org/10.5281/zenodo. 10151236 (2023).
- Sim, M. et al. ChemOS 2.0: An Orchestration Architecture for Chemical Self-Driving Laboratories Preprint at https://doi.org/10. 26434/chemrxiv-2023-v2khf (2023).
- 77. Huber, S. P. et al. AiiDA 1.0, a Scalable Computational Infrastructure for Automated Reproducible Workflows and Data Provenance. *Sci. Data* **7**, 300 (2020).
- Django Developers. Web APIs for Django (2023) https://github. com/encode/django-rest-framework accessed 19 October 2023.
- 79. FastAPI Developers. FastAPI Framework (2023) https://github.com/ tiangolo/fastapi accessed 9 October 2023.
- 80. Rohrbach, S. et al. Digitization and validation of a chemical synthesis literature. *Science* **377**, 172–180 (2022).
- 81. Agnihotri, A. & Batra, N. Exploring bayesian optimization. *Distill* (2020).

# 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". This project was cofunded by European Regional Development Fund via the project "Innovation Centre in Digital Molecular Technologies", UKRI via project EP/S024220/1 "EPSRC Centre for Doctoral Training in Automated Chemical Synthesis Enabled by Digital Molecular Technologies". Part of this work was also supported by Towards Turing 2.0 under the EPSRC Grant EP/W037211/1. The authors thank Dr. Andrew C. Breeson for his helpful suggestions on graphical design, J.B. acknowledges financial support provided by CSC Cambridge International Scholarship from Cambridge Trust and China Scholarship Council. C.J.T. is a Sustaining Innovation Postdoctoral Research Associate at Astex Pharmaceuticals and thanks Astex Pharmaceuticals for funding, as well as his Astex colleagues Chris Johnson, Rachel Grainger, Mark Wade, Gianni Chessari, and David Rees for their support. S.D.R. acknowledges financial support from Fitzwilliam College, Cambridge, and the Cambridge Trust. M.K. gratefully acknowledges the support of the Alexander von Humboldt Foundation. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

# **Author contributions**

M.K., A.A.L., J.B. and S.M. conceived the project. J.B., S.M. and M.K. designed the ontological representation and agent workflow. J.B. implemented the ontologies/agents and deployed the knowledge graph under the advisement of S.M. and K.F.L. The chemistry and HPLC method were developed by C.J.T. (Cambridge) and D.K. (Singapore). D.K. and C.J.T. validated the calculation of objective functions. The self-optimisation campaign involving two labs was set up by C.J.T. (hardware and chemicals in Cambridge), D.K. (hardware and chemicals in Singapore) and J.B. (software on both sides and goal request). M.K. and A.A.L. acquired funding and administrated the project. J.B. draughted the body of this manuscript and SI with inputs from S.M., J.A., S.D.R. and C.J.T. All authors provided feedback on the manuscript.

# **Competing interests**

The authors declare no competing interests.

# Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-44599-9.

**Correspondence** and requests for materials should be addressed to Markus Kraft.

**Peer review information** *Nature Communications* thanks Martin Seifrid, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/ licenses/by/4.0/.

© The Author(s) 2024