Digitization and validation of a chemical synthesis literature database in the ChemPU

Simon Rohrbach†, Mindaugas Šiaučiulis†, Greig Chisholm†, Petrisor-Alin Pirvan, Michael Saleeb,

S. Hessam M. Mehr, Ekaterina Trushina, Artem I. Leonov, Graham Keenan, Aamir Khan,

Alexander Hammer, Leroy Cronin*

School of Chemistry, the University of Glasgow, University Avenue, Glasgow G12 8QQ, UK.

*Corresponding author. Email: lee.cronin@glasgow.ac.uk

†These authors contributed equally to this work.

Abstract

Despite huge potential, automation of synthetic chemistry has only made incremental progress over the past few decades. We present an automatically executable chemical reaction database of 100 molecules representative of the range of reactions found in contemporary organic synthesis. These reactions include transition metal–catalyzed coupling reactions, heterocycle formations, functional group interconversions, and multicomponent reactions. The chemical reaction codes or χDLs for the reactions have been stored in a database for version control, validation, collaboration, and data mining. Of these syntheses, more than 50 entries from the database have been downloaded and robotically run in seven modular chemputers with yields and purities comparable to those achieved by an expert chemist. We also demonstrate the automatic purification of a range of compounds using a chromatography module seamlessly coupled to the platform and programmed with the same language.

To replicate a known chemical reaction the protocol must be obtained from either the literature or a database so that it can be run manually in the laboratory (1). However, not all literature or database entries can be easily reproduced (2). This is a barrier not only to the synthesis of new molecules but also to accumulation of high-quality data for machine learning (3–6) and is exacerbated by the fact that there is also no open standard for coding the procedures or a way to widely report and correct failed experiments (7, 8). An approach that unambiguously captures and codes a chemical synthesis protocol for use by an automated system (9–16) with capacity to be versioned similar to software and record failed experiments would transform the field. Organic synthesis currently requires intensive, highly skilled labor (17) and a typical synthesis can require multiple complex unit operations that are difficult to explicitly encode. This is because the tacit knowledge required is often context-dependent, resulting in ambiguities in the published literature that limit reproducibility, automation, or data mining (18). These limits have been overcome in some specific areas such as oligopeptide (19), oligosaccharide (20), and oligonucleotide (21) chemistry, and in recent years much progress has been made in automating chemical reactions more broadly (22–30).

However, most automated synthetic chemistry platforms remain task-specific or represent islands of automation (10) in an otherwise manual workflow, but even these have bespoke instruction sets with no simple semantic link among them or to the literature. To fully exploit the potential of automation in chemical synthesis and ensure reproducibility of procedures, progress is needed on two fronts (31, 32). First, a truly universal automation platform is required that can perform all unit operations (14, 16); second, a standardized and precise syntax to describe these chemical processes is essential to reliably capture all the critical details of a given chemical process (15). Such a code must also be independent of the type of hardware employed for automation and thus be compiled to work flawlessly on any compatible hardware system.

We present the design, construction, and validation of a workflow that allows us to capture the chemical synthesis literature from manual operation to a fully described and universal Chemical Description Language (χDL) (15, 33) to be run automatically in the Chemical Processing Unit or ChemPU. The process of running the χDL on the ChemPU we call chemputation (similar to computation) and is the reliable conversion of code and reagents into products. We not only show that the χDL can be compiled to run on many different ChemPU configurations but also demonstrate the capacity of the χDL language to encode a wide range of synthetic procedures, which are representative of the organic chemistry toolbox. Overall, 103 different reactions of highly varied chemistry have been translated from the literature to reliable χDL codes, and 53 of these programs have been validated on the hardware with yields and purities comparable to that in the literature. This increased synthesis throughput would not have been possible with earlier versions of the ChemPU (aka chemputer), which could not use χDL (14–16). It also signifies a massive step up in the number of validated χDL procedures compared with the seminal paper on χDL (15) and is testimony to the increased reliability of the hardware employed in this paper. We designed and built a χDL database (34) for our current 103 entries and anticipate this will rapidly expand; the database will be available for anyone to run and validate on suitable hardware. Not only could these χDL entries be implemented on other automated synthesis platforms and material generated on demand, but statistics could be gathered and new versions suggested if required. In addition to directly repeating the validated procedures, the substrate scope for each χDL can be gradually expanded by changing the substrates and adjusting key parameters—such as temperature or time—of the reaction while keeping the rest of the process unchanged. Because we have selected reactions based on popularity, the resulting set of validated χDLs covers a substantial range of common reactions and constitutes an entry point to automation of the entire organic synthesis toolbox. Further, through performing 53 procedures of highly diverse chemistry, the hardware and software of the ChemPU has been pushed to the limit and a path to full universality demonstrated. To do this, key advances have been made by incorporating a χDL-enabled flash column chromatography system in the hardware library. This means that the ChemPU can perform not only the reaction, work up, and concentration, but also the chromatographic separation of the product to directly deliver the purified compound on demand. To achieve this, we show that the platform can react in a dynamic manner, responding to detection of the product to collect the appropriate fractions.

The workflow starting from a literature procedure to a validated entry in our χDL database is illustrated in Fig. 1. By contrast to earlier work on χDL, the focus was not on an exact translation of the original procedure text to χDL but rather on the implementation of a chemical process providing the target molecule. Following this approach allowed us not only to reproduce the literature but also to improve the processes in several instances. Chemical reactions can be captured in χDL, which represents synthetic steps as sequences of physical processes such as Add, Dissolve, Evaporate, and more. There are currently 44 steps within the χDL framework with each step having a fully customizable set of parameters. All often-used tasks in organic syntheses have a boilerplate χDL step

to represent them, such as EvacuateAndRefill to establish an inert atmosphere or Separate to perform a liquid-liquid separation and extraction. The χDL steps help enforce precise descriptions of the process and eliminate any ambiguity such as the number of cycles of evacuation and inert gas refill or process-critical addition speeds. To achieve this, we used our web-based Chemistry Development Environment (ChemIDE) (33) which aids the quick generation of χDL procedures by providing a text-to-χDL translation tool. This works by using a template library of all available χDL steps and an editor in which individual χDL steps are represented as graphical elements, which can be edited and arranged as needed (33). ChemIDE was used in the generation of all χDL procedures detailed in this work.

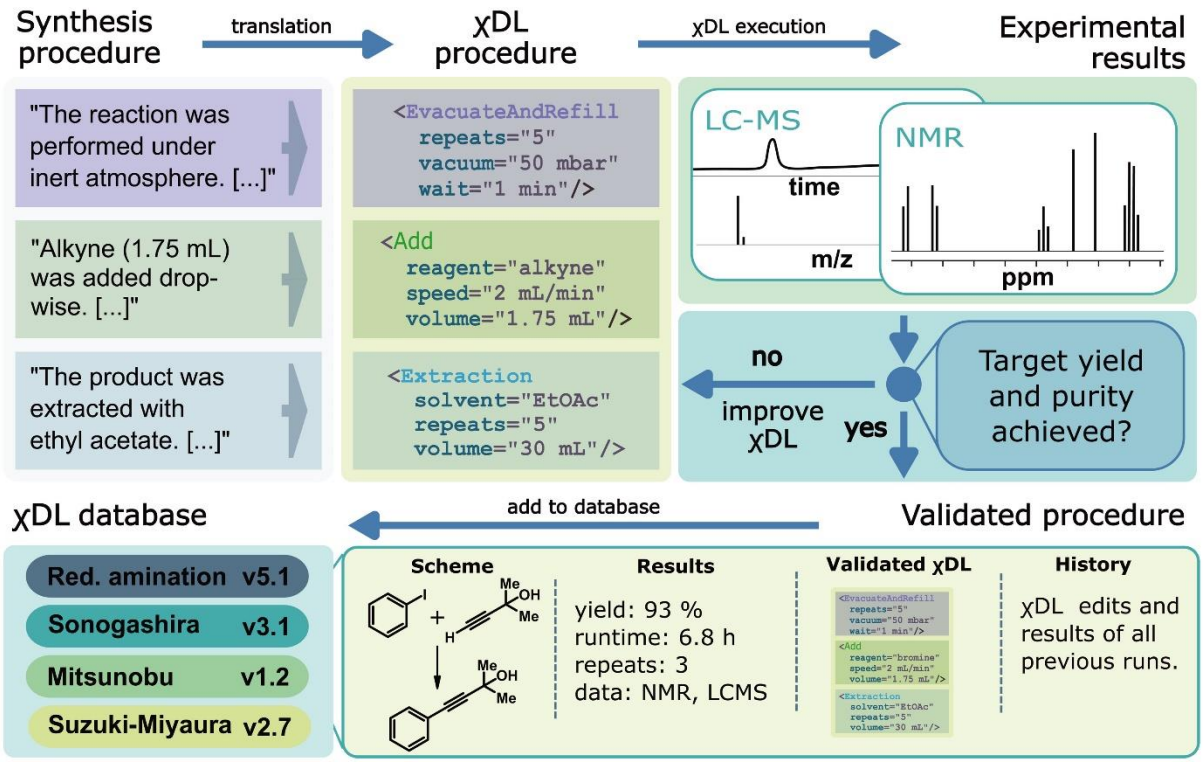

Figure 1 Schematic of χDL protocol optimization.

Chemical reactions can be captured with χDL, which represents synthetic steps as sequences of physical processes such as Add, Dissolve, and Evaporate. The initially established χDL protocol is then executed on a ChemPU and the purity and yield of the product are determined. The χDL protocol can be improved until the process meets the expectation of product purity and yield. At that stage the protocol can be added to the database as validated, backed up by the full characterization of the target product and the process development history.

Expression of a chemical procedure in χDL does not immediately solve the problem of missing information or ambiguity present in the original prose instructions but it does provide an unambiguous path to close it. To do this, some process development and iteration may still be required to maximize yields and purity. After appropriate analysis [Nuclear magnetic resonance spectroscopy (NMR), liquid chromatography mass spectrometry, or gas chromatography mass spectrometry] of the target compound from the ChemPU execution of the χDL code, an assessment of the quality and purity of the product is made. If necessary the χDL is improved to increase the yield and purity and then executed again. The key advantage of χDL is that once a successful process has been encoded, all subsequent users who execute the code on compatible hardware can expect identical results, with no further requirements for process development. All critical knowledge needed to execute the process on qualified hardware both tangible and intangible is now captured

in the χDL. At this stage the protocol can be added to the database as a validated process, backed up by the full characterization of the target product and process development history. Inclusion of process development history is a distinguishing feature of the χDL database; by showing the results of less successful experiments and contrasting them with the final successful run, critical aspects of the process are highlighted and can be quantified.

The χDL database persistently stores information for χDL procedures, experimental results, and relevant analyses. It is a locally hosted PostgreSQL database server containing all validated χDL scripts as described above, which can be accessed through ChemIDE (the web-based χDL development environment) or by using a Python 3–based API for automated database querying. Moreover, for end user experience, ChemIDE is equipped to display characterization parameters of each experiment such as product scale, yield, status (translated, validated, failed) and process duration. Users can submit, search, download, and reproduce trusted syntheses. The database contains final validated synthesis scripts as well as previous developmental versions, which may work to a varying degree, affording the desired products in lower yields, insufficient purities, or leading to process failures (for example, causing blockages or formation of emulsions during liquid-liquid separations) as a result of insufficient or incorrect description of the necessary process parameters for automation. Comparing failed or lower-yield experiments to successful attempts of a given specific reaction or reaction class can unveil critical aspects of the process. Further, the database also contains χDL entries that have been translated but not yet executed on a suitable automation platform. Users interested in unvalidated χDL files can access these and have the option to validate them. The χDL procedures reported here have been validated on a ChemPU, a chemistry automation platform that emulates the manual operations of a bench chemist. Although operationally simple and intuitive, the rigorous implementation means that the platform operates as a finite state machine (Fig. 2). It can be in one of a finite number of states and transitions from one state to the next based on well-defined operations. These operations are defined by the program—the χDL synthesis protocol—as well as the sensor feedback [e.g., temperature, conductivity, pressure, or ultraviolet (UV) absorbance]. The direct mapping of the χDL synthesis instructions to state transitions or “unit operations” highlights the rigorous abstraction of synthesis processes in χDL. Moreover, the clear definition of state transitions as defined in the χDL procedure is critical to ensuring the reproducibility of the χDL synthesis, including on different layouts of the ChemPU and potentially entirely different qualified hardware setups.

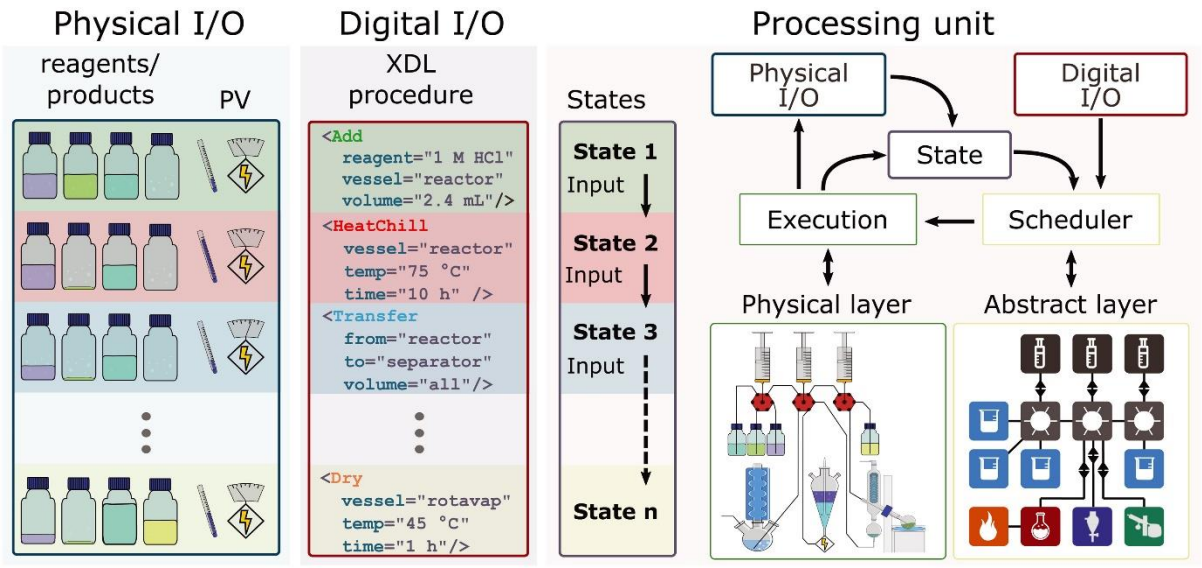

Figure 2 The ChemPU can be regarded as a chemical state machine.

The ChemPU processes the physical input (i.e., reagents, solvents) based on the digital input (i.e., the synthesis script). Each unit operation defines a route for the system to progress from the current state to the next. The state of the system is always well defined, and every detail of the synthesis is known and can be documented. PV, process variables.

The ChemPU state machine consists of three logical parts: the physical input or output (I/O), digital I/O, and the processing unit. The processing unit can transition through several states based on either the initial conditions of the ChemPU or a combination of the physical and digital I/O, that is, the current conditions as defined by the sensors, the process variables, and the χDL step being executed. The execution of the χDL step according to the scheduler gives rise to a new state to be acted upon in later steps and results in a physical change to the physical I/O, e.g., a change of location of reagents, a change in temperature, the phase boundary in a liquid-liquid separation, or the peak elution during chromatography. The scheduler resorts to a graph representation of the hardware (abstract layer) to interpret the χDL script and orchestrate the hardware for concerted tasks (e.g., moving liquid through the liquid-handling backbone). The abstract layer defines the locations and connections of the hardware devices as nodes and contains specific information on each node such as the IP address and temperature limits of the device in question. The graph file together with the χDL file can be compiled into an execution file (executable χDL or xdlexe), which is platform-specific. The strict separation of the description of the chemical process into the χDL file and the hardware platform description into the graph file ensures that the χDL file remains platform-independent. It also allows for flexibility in how the platform is designed and what its exact physical layout is. This means that each χDL can be versioned and compiled to run on any suitable platform and that the ChemPU system is highly modular, flexible, and extensible (Fig. 3).

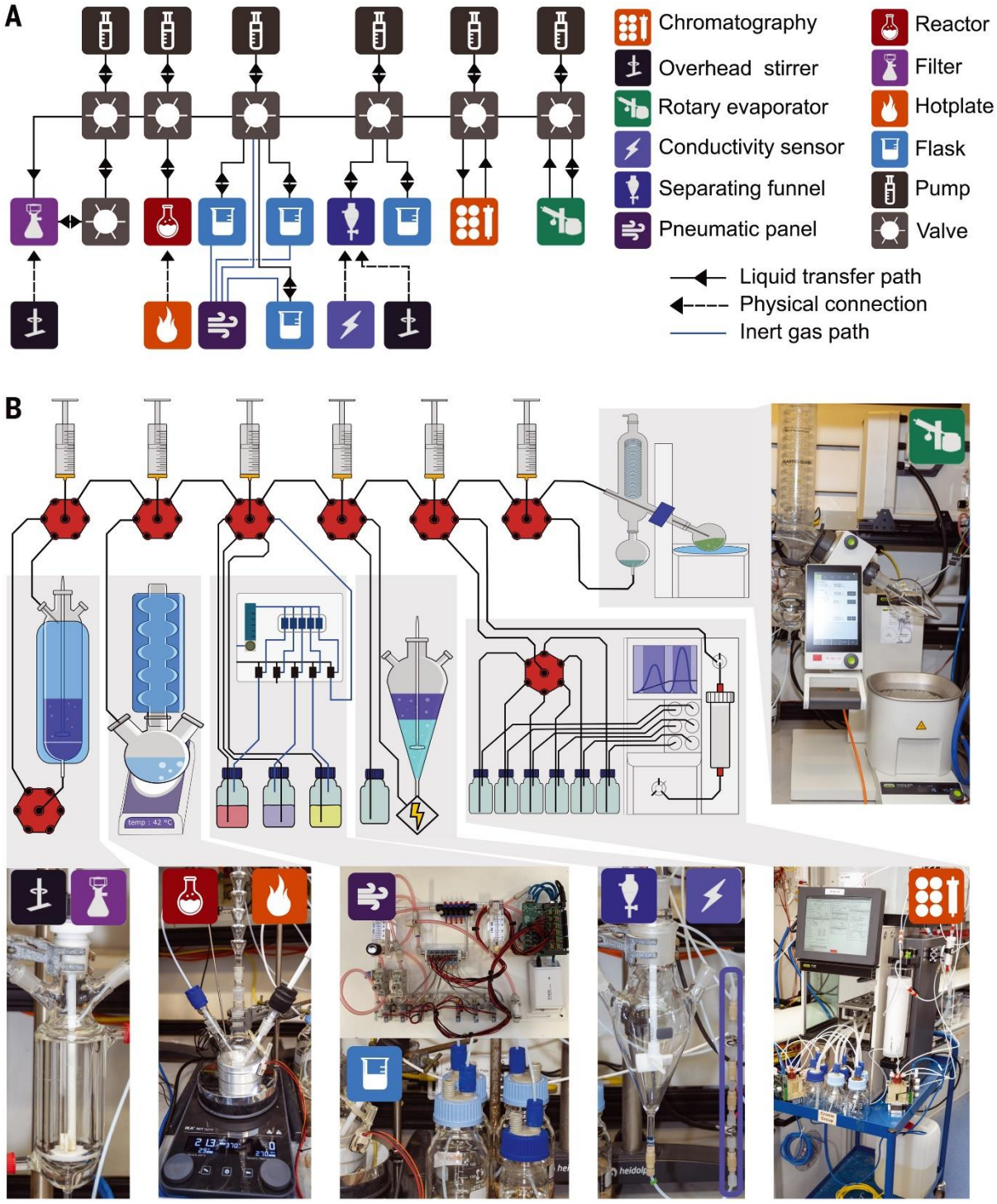

Figure 3 The physical layout of the ChemPU and the available hardware library can be represented as a graph.

*(**A**) The exact state of the ChemPU for each individual synthesis is represented as a graph. The nodes indicate the modules and the edges define the tubing connections for liquid transfers and physical connections. (**B**) A schematic representation of the ChemPU with images of the individual modules. The ChemPU emulates the manual batch chemistry workflow and uses much of the typical laboratory hardware. The latest key addition to the ChemPU hardware library, as described in this paper, is a flash chromatography system which allows for fully automated purification of the reaction products. The liquid-handling backbone consists of an array of pumps and valves which transfer reagents, solvents, and solutions of intermediates between the different units of the system. Reactions are either carried out in a round bottom flask reactor or a filter. The work-up is performed in a separator that is agitated with an overhead stirrer. The phase boundary is detected with a conductivity sensor. Solutions are concentrated in a rotary evaporator. Column chromatography is performed on a flash column chromatography machine.*

By mirroring the unit operations of batch synthetic chemistry, the ChemPU represents a universal, programmable hardware platform for execution of synthetic chemistry as previously demonstrated (14–16). The platform can be readily expanded as a result of its modular nature, with individual modules being connected through the liquid-handling backbone, analogous to the bus of a conventional computer. Connection to the liquid-handling backbone (consisting of pumps and valves) is through a single piece of flexible tubing, which allows modules to be easily removed for maintenance or rearranged to optimize operations (e.g., by segregating aqueous and water-sensitive parts of the process). The liquid-handling backbone consists of a series of syringe pumps and valves. A typical backbone consists of six of each; however, the backbone is readily contracted or expanded to accommodate the requirements of the desired chemical process. The valves have six positions and seven ports each. Each valve in the liquid-handling backbone is connected to a pump, its nearest neighboring valves, and a waste container, and can connect to three to four different reagents, solvents, or hardware modules. The connectivity of the modules to the backbone is represented in an abstract manner by a graph as described above. Cleaning of the backbone is carried out through an automated cleaning routine that can be defined by the user to account for different types of contamination present after different procedures. In addition to the liquid-handling backbone, the ChemPU systems used to execute the syntheses reported here incorporated a reaction module consisting of a standard hotplate controlled through an Ethernet-to-serial convertor, a separator for liquid-liquid extractions equipped with an overhead stirrer for agitation, as well as a conductivity sensor for phase boundary detection; it also includes a jacketed filter for precipitation and recrystallization of products, a number of reagent flasks, a rotary evaporator, and an optional chromatography system.

Validation of literature procedures on the ChemPU

With the abstraction of chemputation, the χDL language, and the ChemPU platform, we set out to translate and automate the typical reactions from the organic chemistry toolbox. Organic chemistry encompasses an immense diversity of transformations. Despite a large degree of variety, most reactions can be classified succinctly with fewer than ten categories. Several studies have analyzed the reaction frequencies in different fields, e.g., medicinal chemistry, process chemistry, and total synthesis (35–38). There are some notable differences in the distribution of reaction classes used in synthesis, depending on the primary goal; for example, medicinal chemistry researchers may prefer transition metal–catalyzed C-C bond-forming reactions which allow for convenient generation of large numbers of related compounds for biological assays, whereas modern total synthesis relies more heavily on elaborate ring-forming reactions for assembly of complex molecular scaffolds in the fewest number of steps possible (39).

Additionally, although protecting-group chemistry is the cornerstone of some synthesis fields such as peptide synthesis (40)or carbohydrate chemistry (41), researchers working on total syntheses often prefer more elegant protecting-group–free approaches (42). Despite the minor variations, these categories embody the varied toolbox of modern organic chemistry. To represent these categories with the examples from all types of reactions we chose to translate the χDLs of these procedures and validate them with the ChemPU (Fig. 4). The carbon-carbon bond-forming reaction class was further separated into transition metal–catalyzed and transition metal–free reactions. Furthermore, a separate multicomponent reaction class was introduced as these reactions generally accomplish

multiple chemical transformations in one synthetic operation. The initial reactions were chosen from the most cited papers in the journal Organic Syntheses (43). This journal is notable in the organic chemistry field in that it publishes practical methods for either synthesis of notable compounds or execution of important synthetic methods, and the submitted procedures have been repeated at least once by expert chemists independent of those who submitted the original synthesis. Although the procedures from this journal generally have a high level of detail there was still a need for some process development, highlighting the difficulty of capturing all necessary information in an unstructured prose text format as opposed to χDL. Selecting these highly cited papers from Organic Syntheses covered the top reaction classes but provided an uneven distribution. Hence, further examples were manually selected from notable literature sources to achieve a more balanced representation of the organic chemistry toolbox with our dataset.

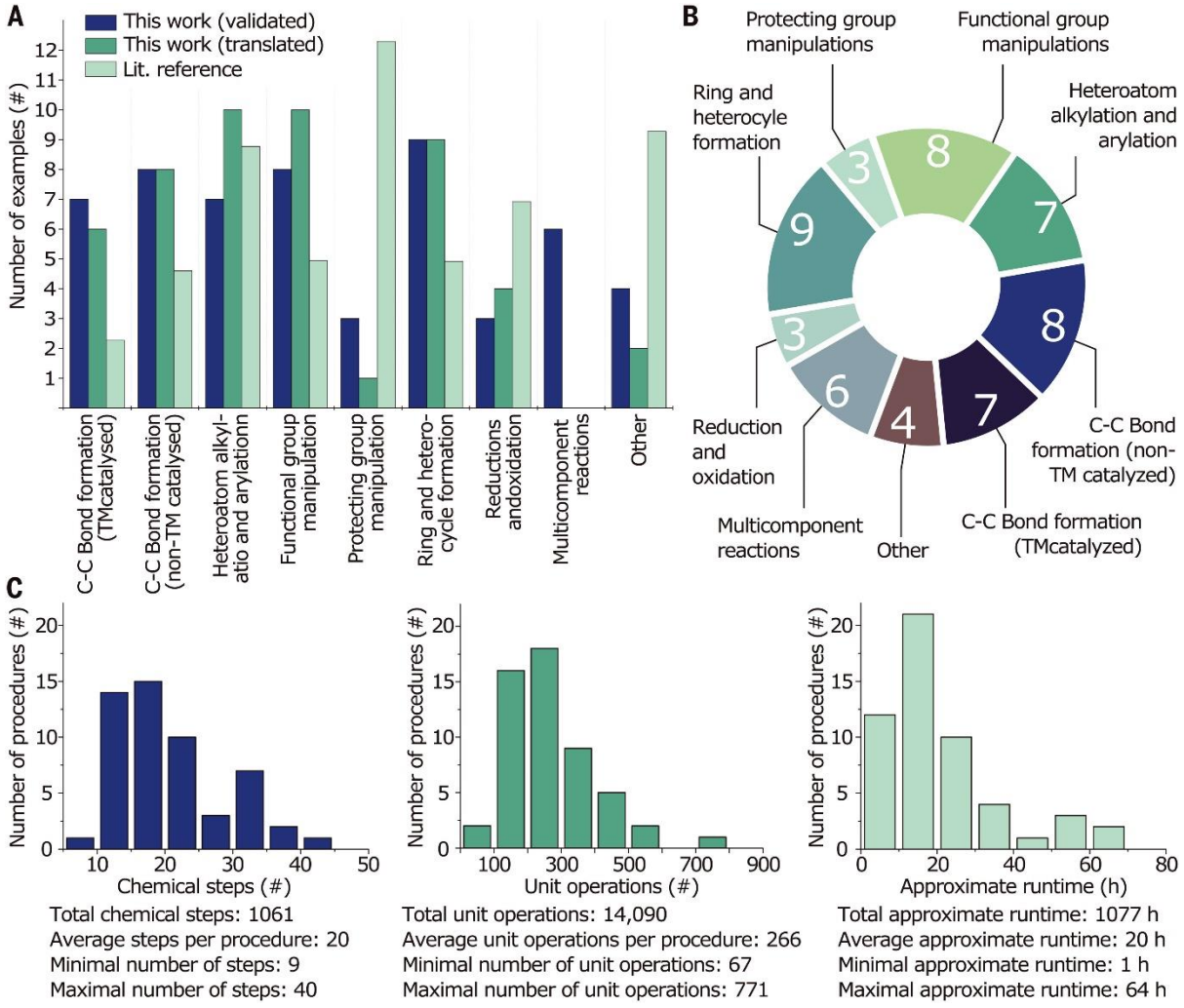

Figure 4 A representative selection of most-used reaction classes have been translated to χDL and validated on a ChemPU.

*(**A**) Number of examples per reaction class. Once a literature procedure has been captured by χDL it is marked as "translated." When a translated procedure is successfully executed on a ChemPU it is moved to the "validated" class of χDL scripts. For reference, the average frequency of reactions over the fields of medicinal chemistry, process chemistry, and total syntheses are shown (35–38). (**B**) The distribution of validated reactions and a specific example for illustration is shown. (**C**) Chemical operations, unit operations, and total runtime per procedure.*

The reactions chosen for each of the categories include well-established classical reactions and important contemporary reactions, as well as some more unconventional synthetic transformations (Fig. 4); for a comprehensive list of all translated reactions, see supplementary materials fig. S143. The selected transition metal–catalyzed carbon-carbon bond forming reactions included commonly used Suzuki, Heck, and Sonogashira couplings, as well as a stereoselective Carroll rearrangement.

The transition metal–free carbon-carbon bond-forming reaction class encompasses such classical reactions as the Wittig reaction, Friedel-Crafts alkylation, and the Aldol and Claisen condensations. Different types of heteroatom alkylations are represented by palladium-catalyzed Buchwald-Hartwig coupling, copper-catalyzed alkylation, SNAr reaction of heteroarenes, and reductive amination reactions. Functional group interconversions include a Mitsunobu reaction, nitrile formation, and esterifications, among others. Manipulations of protecting groups include common boc, benzyl, and tosyl groups. Ring and heterocycle formations include both classical syntheses such as the Fischer indole synthesis and a more exotic formation of a trisubstituted pyrylium salt. Reduction and oxidation reactions span conventional hydride reduction, Jones oxidation, and a palladium-catalyzed hydrogen transfer reaction. Finally, the multicomponent reactions include the well-known Ugi reaction as well as other more unusual cascade reactions and one-pot multistep manifolds. This diverse set of reactions (for a comprehensive list of all validated and translated reactions, see supplementary materials figs. S140 and S143, respectively) covers the standard organic chemistry toolbox. Crucially, automating further reactions simply requires translation of the original synthesis procedure to χDL.

The average procedure consists of 20 discrete, high-level instructions such as Add, Separate, and Evaporate with some procedures having up to 40 such instructions (Fig. 4C). Unpacking these high-level χDL steps into the corresponding unit operation—e.g., StartStir, WaitForTemp, ApplyVacuum—gives an average of 266 operations that have been executed. The successful executions of all χDL scripts took >1000 hours of chemputation across seven different systems. This figure only includes the operations from the final iteration of each χDL protocol and includes the reaction time but does not account for asynchronous steps, i.e., steps in which two processes are running in parallel on the same ChemPU hardware, such as a cleaning step for a rotary evaporator running at the same time as a reaction. The yields of the reactions performed on the ChemPU were in general comparable to that of the literature yields after a period of process development. This could be required to fill the gaps in the original protocol and is common to all synthetic development whether manual or automated, or to adapt elements of the protocol not amenable to automation, such as unexpected formation of precipitates that lead to blocked lines. A selection of reactions is shown in Fig. 5 to illustrate the performance of the platforms and give specific examples to show the breadth of chemistry that has been performed.

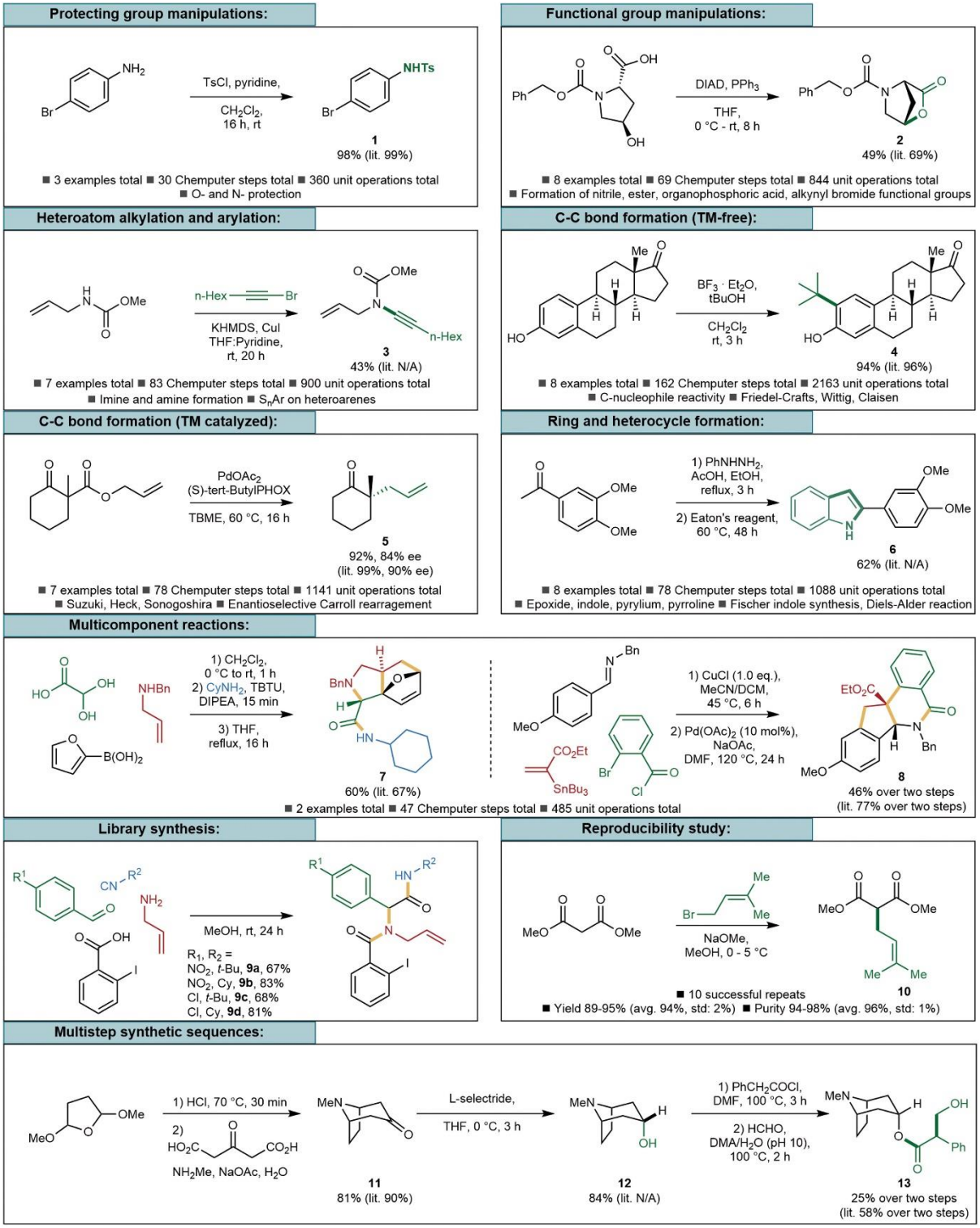

Figure 5 ***Representative examples of χDL procedures validated on the ChemPU.***

Transformations from all the main reaction classes afforded products in yields comparable to those reported in the literature for manual synthesis. Additionally, multicomponent reactions were performed and a small library of compounds was prepared by varying the starting materials of one of the multicomponent reactions. Finally, the reproducibility of a validated χDL procedure has been examined and a χDL procedure of a multistep synthetic sequence has been validated. The counts for total steps and total unit operations include all examples in a given category.

Automation of a diverse set of reactions on the ChemPU

The system is tolerant of moisture-sensitive or highly reactive reagents such as potassium bis(trimethylsilyl)amide (KHMDS) used in a copper-mediated alkynylation of a carbamate to afford **3**, boron trifluoride used in a Friedel-Crafts alkylation of a steroid estrone to afford derivative **4**, or Eaton's reagent (10% phosphorus pentoxide solution in methanesulfonic acid) used in a Fischer indole synthesis of **6**. Additionally, reactions requiring inert atmosphere were successfully executed on the platform including a palladium-catalyzed enantioselective Carroll rearrangement to give **5**. Procedures of up to 90 mmol scale were efficiently executed on our ChemPU platform. Conveniently, once a χDL script is produced, a particular reaction can be scaled up or down within the constraints of the available vessel sizes and the chemical process, such as safety considerations or heat and mass transfer (see supplementary materials section 3, fig. S141, for the full distribution of reaction scales). The χDL procedures for the generation of more complex products arising from multicomponent and cascade reactions were also successfully executed on the platform. For example, a Petasis/Diels-Alder cyclization cascade has been used for rapid generation of a scaffold containing multiple stereogenic centers **7**, with potential for further derivatization in a library synthesis. Similarly, a copper(I)-catalyzed three-component coupling/palladium(0)-catalyzed annulation cascade was also successfully applied, affording product **8** which contains the indenoisoquinoline scaffold.

Expanding the substrate scope

The substrate scope of validated χDL procedures can be expanded by generating a compound library with the ChemPU. One particularly attractive prospect is the use of validated χDL procedures for the construction of large libraries of compounds for biological screening. Such libraries could conveniently be accessed simply by changing the starting materials without major modifications to the synthesis scripts; i.e., once a process has been established it can be applied to many different substrates as a general procedure by only varying key parameters such as the substrates, reaction solvent, and reaction time. To showcase such an approach, a small library of α-acylamino amides **9a to 9d** was synthesized through a multicomponent Ugi reaction. To do this, we conducted simultaneous execution of multiple or "multithreaded" reactions in parallel on the ChemPU by using reactant combinations from two different isocyanide and two aldehyde starting materials affording four structurally related α-acylamino amide products. Further expansion of the set of reactants used would rapidly expand the number of products generated and allow for swift generation of larger libraries.

Reproducibility of the ChemPU synthesis

To examine the consistency and reliability of executing the curated χDL procedures, we set out to repeat the same reaction protocol multiple times on the ChemPU platform. An alkylation of malonate ester (affording **10**) was chosen as a suitable reaction for the reproducibility study, as accurate temperature control and rate of addition are key to the success of the process. After the initial process development, a validated χDL procedure script was obtained and the reaction

protocol was successfully replicated 10 times in 12 attempted runs. The two failures were caused by incorrect phase boundary determination during liquid-liquid separations; product could have been recovered through manually restarting the system, but that was not done here. Crucially, execution of the curated χDL procedure reliably afforded the product in consistent yields (avg. 94%, min 89%, std: 2%) and purities (avg. 96%, min 94%, std: 1%). Together with the ability to generate libraries of compounds, the ChemPU can be used to automate the highly repetitive work of generating multiple batches of the same material or repeating the same reaction with different substrates once the initial protocol has been set up.

Multistep synthesis

The versatility of the platform is further demonstrated by the ability to execute multistep synthetic sequences. Atropine **13**, an anticholinergic medication used in treatment of nerve agent poisoning, was synthesized in four steps from simple commercially available starting materials. Synthetic protocols for individual steps from multiple sources—as well as a reduction protocol that was previously reported for related substrates but not for the synthesis of **12**—were successfully converted to χDL procedures. The ability to efficiently execute multistep reaction protocols combined with the reliability offered by reproducible execution of a well-defined synthesis script reaffirms the universality of the platform toward the breadth of synthetic organic chemistry.

Fully automated purification on the ChemPU

Chromatographic separation of the product compound from a reaction is the go-to method of purification for small- and medium-scale organic syntheses. Many commercially available chromatography systems exist for assisting lab-based chemists in chromatographic separations. However, these systems still require a substantial amount of user interaction. For example, the crude material must be manually loaded onto the column and the product fractions must be manually identified, washed out of the fraction vials, and combined. Further, these commercial systems require user interactions at several different stages, thus tying the chemist to the lab even if it is only for a trivial task such as loading the sample onto the column. To integrate the Buchi Pure C-815 chromatography system with the ChemPU, two auxiliary hardware units were built: a column carousel that allows preinstallation of different columns on the system and an extension to the fraction tray. The latter allows for recovery of the product fraction by the ChemPU. The first operation that is challenging to automate is the sample loading onto the column. The laboratory-based chemist usually chooses between dry-loading and liquid injection of the sample. We aimed to implement the liquid injection method which ties in nicely with the liquid-handling backbone of the ChemPU; further, the liquid injection sample loading method entailed little process development, requiring only the identification of a suitable solvent mixture and volume to dissolve the crude material. The second challenge to full automation of normal-phase chromatography is to reliably select the product peak. Usually, chemists need to analyze individual fractions by thin-layer chromatography, mass spectrometry, or NMR after chromatographic separation. For the ChemPU integration of the module several alternative options were considered. We found that considering the UV/visible response or the signal from the elastic light scattering detector of the eluting fractions and choosing the peak with the largest area under the curve for a specified signal trace gave the best

trade-off between reliability and flexibility; for a given well-performing reaction the product peak can correctly be identified independent of the exact retention time. Moreover, this method does not rely on more elaborate product identification such as mass spectrometry or NMR.

Once the method is developed and coded in χDL it can be executed on the ChemPU or equivalent automation system as shown in Fig. 6. The platform controller starts the chromatography process by defining the run parameters on the commercial chromatography unit (central hub), such as flowrate and detector settings. The actual run preparations, such as baseline corrections and the equilibration of the column, are then executed. Next, the sample of crude material is dissolved, transferred to the chromatography machine, and injected onto the column. The sample injection process also includes a rinsing sequence to minimize loss of material during the sample dissolution and transfer. Once the sample loading is complete the gradient run is commenced. During the gradient run the chromatography machine continuously reads the detector signals and sends them to the ChemPU controller software in real time.

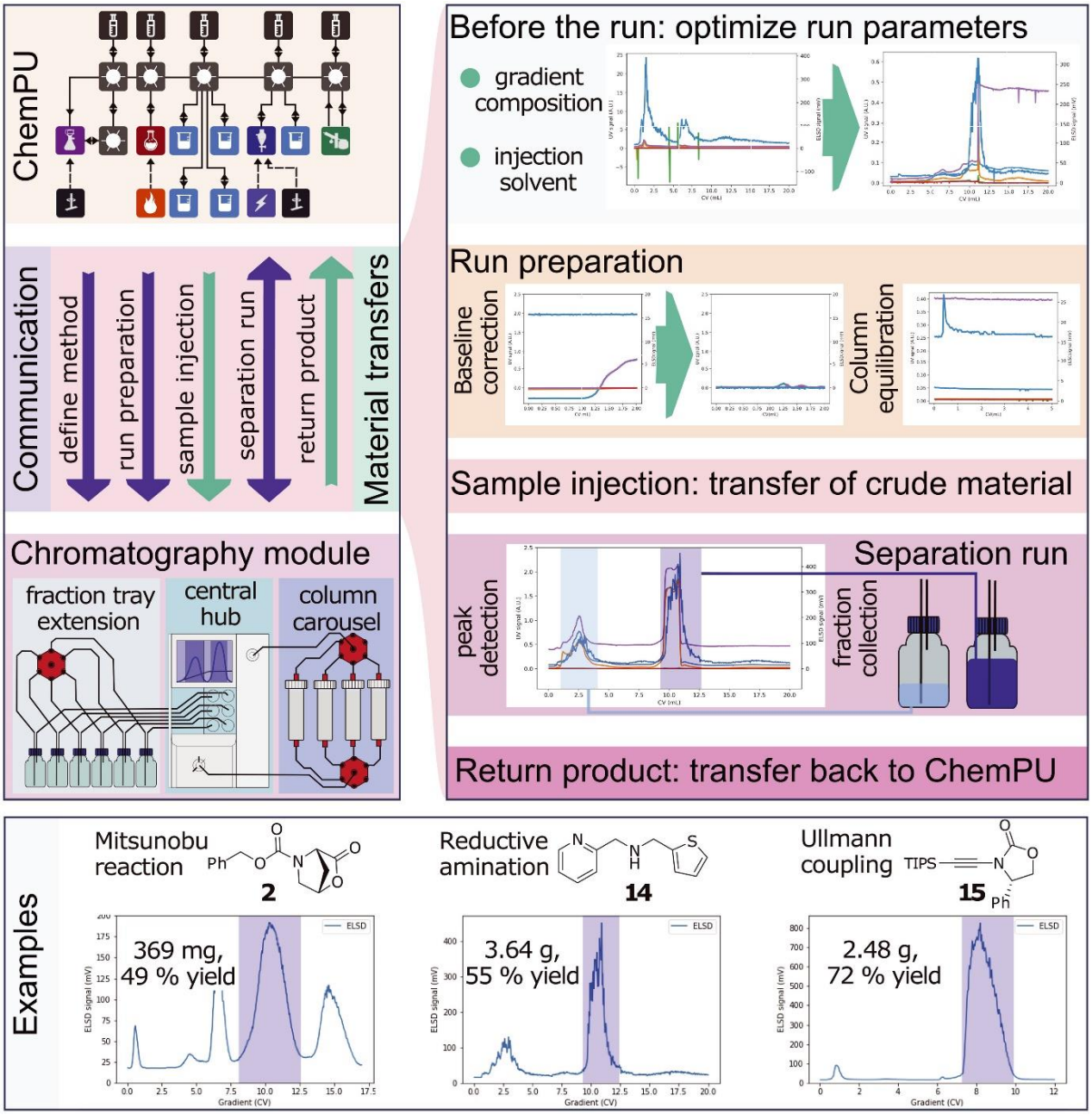

Figure 6 The chromatography module.

This component stands out from the ChemPU hardware library in terms of the complexity of the information and material flow between the module and the controller and the other hardware. The chromatography method consists of two crucial parts: the sample injection protocol and the solvent gradient. Once these parameters have been optimized the separation run can be initiated. The ChemPU defines the run parameters on the commercial chromatography unit. Then the run preparations (baseline correction over the gradient, and column equilibration) are performed. Next the sample is injected onto the column. During the gradient run the chromatography machine sends the detector signals to the ChemPU controller in real time. The controller performs the peak detection and triggers the fraction collection mechanism of the chromatography machine. When the separation run is complete the product peak is identified and transferred to the next module (usually the rotary evaporator).

The ChemPU controller then performs the peak detection and triggers the fraction collection mechanism of the chromatography machine. The controller also keeps track of fraction vial filling levels and various run parameters such as back pressure buildup, solvent vapor levels, and solvent levels of the gradient solvents and the solvent waste drum. If any of these parameters exceed the specified threshold, an appropriate error-handling routine is initiated that pauses the chromatographic separation in a controlled way. When the separation run is complete the product peak is identified and transferred to the next module (usually the rotary evaporator). The crude material is typically transferred from the rotary evaporator to the chromatography module followed by transfer of the purified product back from the chromatography module to the rotary evaporator, and as such the rotary evaporator flask needs to be cleaned in between. Hence an optional cleaning routine for the target vessel of the purified product has been implemented and can be performed during the chromatographic separation. The integrated chromatographic separation was used for three reactions. The process of these chromatographic separations has been captured by χDL, specifying every minute and critical detail in a concise, easy-to-understand way. Hence, reproduction of the chromatographic separations on another ChemPU or equivalent system or even manually with a commercially available chromatography machine is readily possible.

Outlook

We have shown how the chemical synthesis literature can be easily converted to a universal chemical code that can run on any robot capable of chemputation; the only requirements for this are a batch reactor, a separator, evaporator, and purification system. This means that potentially many different robotic approaches will be able to use identical χDL codes to produce identical results. The use of a χDL Chemify database will not only facilitate reproduction of published procedures but also provide the community with a rich source of validated data amenable to state-of-the-art machine learning for reaction optimization, route planning, increased safety, and reduced environmental impact of synthesis while substantially reducing labor for bench chemists repeating well-known procedures.

Acknowledgements

We thank BUCHI for supplying us with a pure C-815 chromatography system and API to interface it with the ChemPU software package. The authors gratefully acknowledge the assistance of D. Doran and V. Sandoval in the preparation of and interfacing with the χDL database.

Funding: We gratefully acknowledge financial support from the EPSRC (EP/L023652/1, EP/R020914/1, EP/S030603/1, EP/R01308X/1, EP/S017046/1, and EP/S019472/1), the ERC (670467 SMART-POM), the EC (766975 MADONNA), and DARPA (W911NF-18- 2-0036, W911NF-17-1-0316, and HR001119S0003).

Author contributions: L.C. conceived the concept, architecture, and programming approach. S.R., M.Š., A.P., M.S., H.M.M., E.T., A.I.L., and A.H. configured the robots, ran the synthetic protocols, and characterized the products. G.K. and A.K. helped with the development of the database and integration with the ChemIDE. L.C. wrote the paper together with S.R., M.Š., and G.C. with help from all authors.

Competing interests: L.C. is the founder of Chemify Ltd. L.C. is listed as an inventor on the UK patent GB 2209476.7., which describes this system.

Data and materials availability: Supplementary materials include full details to reproduce this work, including instructions for how to build and run the platform. Additional details of the electronic and mechanical components of the platform, videos of the platform working, and the software to produce and run the χDL files, and the raw analytical data for all experiments are available at Zenodo (44).

References

1. W. A. Warr, Mol. Inform. 33, 469–476 (2014).

2. M. Baker, Nature 533, 452–454 (2016).

3. H. Gelernter, J. R. Rose, C. H. Chen, J. Chem. Inf. Comput. Sci. 30, 492–504 (1990).

4. M. H. S. Segler, M. Preuss, M. P. Waller, Nature 555, 604–610 (2018).

5. O. Engkvist et al., Drug Discov. Today 23, 1203–1218 (2018).

6. B. A. Grzybowski, K. J. Bishop, B. Kowalczyk, C. E. Wilmer, Nat. Chem. 1, 31–36 (2009).

7. I. W. Davies, Nature 570, 175–181 (2019).

8. N. Matosin, E. Frank, M. Engel, J. S. Lum, K. A. Newell, Dis. Model. Mech. 7, 171–173 (2014).

9. M. Trobe, M. D. Burke, Angew. Chem. Int. Ed. 57, 4192–4214 (2018).

10. J. Li et al., Science 347, 1221–1226 (2015).

11. C. W. Coley et al., Science 365, eaax1566 (2019).

12. T. Jiang et al., Chem. Sci. 12, 6977–6982 (2021).

13. A. C. Bédard et al., Science 361, 1220–1225 (2018).

14. S. Steiner et al., Science 363, eaav2211 (2019).

15. S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan, L. Cronin, Science 370, 101–108 (2020).

16. D. Angelone et al., Nat. Chem. 13, 63–69 (2021).

17. P. G. Nantermet, Chem 1, 335–336 (2016).

18. Z. Wang, W. Zhao, G. F. Hao, B. A. Song, Drug Discov. Today 25, 2006–2011 (2020).

19. R. B. Merrifield, Science 150, 178–185 (1965).

20. O. J. Plante, E. R. Palmacci, P. H. Seeberger, Science 291, 1523–1527 (2001).

21. G. Alvarado-Urbina et al., Science 214, 270–274 (1981).

22. M. Legrand, P. Bolla, J. Automat. Chem. 7, 31–37 (1985).

23. S. B. Boga et al., React. Chem. Eng. 2, 446–450 (2017).

24. B. Burger et al., Nature 583, 237–241 (2020).

25. A. G. Godfrey, T. Masquelin, H. Hemmerle, Drug Discov. Today 18, 795–802 (2013).

26. B. P. MacLeod et al., Sci. Adv. 6, eaaz8867 (2020).

27. H. Okamoto, K. Deuchi, Lab. Robot. Autom. 12, 2–11 (2000).

28. A. Orita, Y. Yasui, J. Otera, Org. Process Res. Dev. 4, 333–336 (2000).

29. Y. Tanaka, S. Fuse, H. Tanaka, T. Doi, T. Takahashi, Org. Process Res. Dev. 13, 1111–1121 (2009).

30. S. Chatterjee, M. Guidi, P. H. Seeberger, K. Gilmore, Nature 579, 379–384 (2020).

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

32. L. Wilbraham, S. H. M. Mehr, L. Cronin, Acc. Chem. Res. 54, 253–262 (2021).

33. M. Craven, G. Keenan, A. Khan, M. Lee, L. Wilbraham, ChemIDE. https://croningroup.gitlab.io/chemputer/xdlapp/ (2021).

34. Cronin Group, cDL Database: https://croningroup.gitlab.io/ chempu/xdl-database/.

35. S. D. Roughley, A. M. Jordan, J. Med. Chem. 54, 3451–3479 (2011).

36. N. I. Vasilevich, R. V. Kombarov, D. V. Genis, M. A. Kirpichenok, J. Med. Chem. 55, 7003–7009 (2012).

37. N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli, G. A. Landrum, J. Med. Chem. 59, 4385–4402 (2016).

38. J. S. Carey, D. Laffan, C. Thomson, M. T. Williams, Org. Biomol. Chem. 4, 2337–2347 (2006).

39. P. S. Baran, J. Am. Chem. Soc. 140, 4751–4755 (2018).

40. A. Isidro-Llobet, M. Alvarez, F. Albericio, Chem. Rev. 109, 2455–2504 (2009).

41. A. G. Volbeda, G. A. Marel, J. D. C. Codée, in Protecting Groups. S. Vidal, Ed. (Wiley, 2019), pp. 1–27.

42. R. A. Fernandes, P. Kumar, P. Choudhary, Chem. Commun. 56, 8569–8590 (2020).

43. W. R. Roush et al., Organic Syntheses (2021); http://www. orgsyn.org/.

44. S. Rohrbach synthesis literature database in the ChemPU, Zenodo (2022); doi:10.5281/zenodo.6534009.