

Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: D. Garay-Ruiz, S. Pablo-García, H. Hao and M. S. Martín-González, *Digital Discovery*, 2025, DOI: 10.1039/D6DD00019C.



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

Cite this: DOI: 00.0000/xxxxxxxxxx

Database Utility for Cyclovoltammetry Knowledge (DUCK): unified platform for electrochemical data

Diego Garay-Ruiz,^a Sergio Pablo-García,^b Han Hao^b and Marisol Martín-González^{a,*}Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Cyclic voltammetry (CV) is a valuable tool for electrochemistry, providing qualitative and quantitative information about redox processes taking place in solution. Despite its ubiquity, the lack of standardized reporting and sharing protocols significantly hinders data reusability and reproducibility. Here, we present DUCK (Database Utility for Cyclovoltammetry Knowledge), a comprehensive platform for organizing, analyzing and exploring CV data through ontology-based knowledge graphs. The system comprises three components: (1) CVOnto, a formal ontology for electrochemical CV measurement data, (2) DUCK-KG, an automated tool for the generation of knowledge graphs from experimental metadata, and (3) DUCK-Viz, an interactive web interface for visualization and analysis of CV data. We demonstrate applicability across two distinct domains: 130 CV experiments from electrodeposition studies in traditional laboratories, and 79 automated measurements from metal-ligand complexes in self-driving laboratories, performing Bayesian optimization workflows. The application to these two paradigms demonstrates DUCK's potential as a unified platform implementing FAIR data principles for the electrochemistry community.

1 Introduction

One of the most important characterization techniques in electrochemistry is, undoubtedly, cyclic voltammetry (CV), which provides insights into key aspects such as reaction kinetics, electron transfer processes and their coupled chemical steps, among other relevant features^{1–4}. CV is based on the application of a sweeping potential (V), at a given scan rate, on a three-electrode cell setup containing a given set of electrolytes. The intensity (I) response of the setup is measured across the potential scan, registering *peaks* as the different species in the solution get reduced or oxidized at the working electrode at a given potential. Figure 1 showcases a simple example of the I vs E plots obtained after the measurement.

The analysis of a single CV characterization does already provide a large amount of important qualitative (plot shape) and quantitative (potential and intensity of the peak, inter-peak separation) information about the target system. Moreover, further details can be obtained by modifying the conditions of the electrochemical setup, such as reactant concentrations or scan rates, to produce additional measurements.

Therefore, CV has become a routine characterization technique for almost any electrochemical application. As an example of this, we may highlight electrodeposition processes, in which differ-

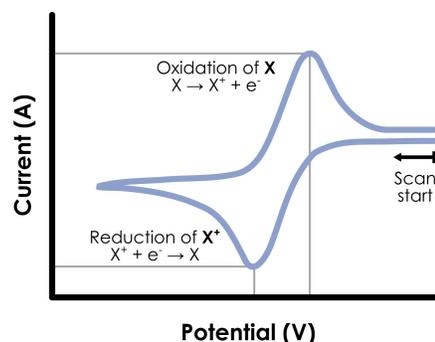


Fig. 1 Example of a reversible CV intensity vs potential plot, identifying the oxidation (right) and reduction (left) peaks for a simple $X^+ + e^- \rightarrow X$ redox process.

ent kinds of materials are applied as coatings over a given surface through the application of external current⁵. The setup of this widely-used, cost-effective approach for the development of layered materials and devices is founded on prior CV measurements^{6,7} that enable the identification of the ideal conditions, like the target potential for the process.

As a consequence, electrochemical laboratories produce large amounts of CV data, which unfortunately is often undervalued and not properly shared. Despite the ubiquity of CVs in electrochemistry, there does not seem to be a universally adopted standard to store, share and reuse this information in a reliable and trustworthy way, especially considering the complex nature

^a Instituto de Micro y Nanotecnología, IMN-CNM, CSIC (CEI UAM+CSIC), Isaac Newton, 8, 28760, Tres Cantos, Madrid, Spain

^b Acceleration Consortium, 700 University Ave, M7A 2S4, Toronto, Canada



of electrochemistry experiments. The organization and accessibility of experimental data is increasingly critical for scientific processes, particularly as laboratories transition towards automated, data-intensive operations. Therefore, there is a growing need for suitable strategies to apply the FAIR (Findable, Accessible, Interoperable and Reusable) principles⁸ to cyclic voltammetric data. Among previous efforts on the topic, we can remark the initiative from the NFDI4Chem consortium⁹, developing open-source tools and schemes for the production of FAIR Data from Electronic Lab Notebooks (ELNs) such as Chemotion¹⁰. This procedure extracts and standardizes the information in the ELN, generating a hierarchy of files that can then be hosted in general-purpose repositories like RADAR¹¹ or Zenodo.

The development of FAIR data management schemes for CV data becomes even more important when taking into account the growing development of self-driving laboratories (SDLs)^{12–16}. The automation of electrochemistry in SDLs directly implies the production of large amounts of high-throughput data, including, of course, the performance of many CV characterizations¹⁷. In this context, standardization protocols could be coupled to the output of these SDLs, immediately producing well-organized databases to enhance the reutilization of the data and the overall reproducibility of the experiments. Moreover, such a protocol would simplify the connection of automated and traditional electrochemical experiments, which should not be disregarded. Despite the enormous potential and increasing relevance of SDLs for current chemistry, "traditional" laboratories (TLs) are still a fundamental part of science and will coexist with SDLs for the foreseeable future. Thus, making the results from both paradigms interoperable through a common schema would contribute to bridge the gap between them.

As in most fields of science, Artificial Intelligence (AI) tools have had a big impact in electrochemistry. Problems that have been tackled by means of AI range from circuit modelling via statistical machine learning¹⁸, or, in the context of CVs, the automated identification of electron transfer mechanisms from CV curves^{19,20}. In light of this and regarding the huge impact of AI in modern science, the relevance of well-organized electrochemical databases to streamline the utilization of datasets is evident.

Looking at the bigger picture, it is clear that CV is only an individual part of electrochemical setups, where many more techniques should be taken into account to tackle a fully digitalized and FAIR approach to electrochemistry. Nonetheless, their importance and relative simplicity provide an ideal starting point for the development of this kind of standardization schemes, as also proposed by Herrmann et al.⁹, paving the way for further extensions until the completion of the puzzle.

One of the leading approaches for the management of understandable and transferrable data is the Semantic Data framework, based on the definition of *ontologies*, where a taxonomy for a given field of interest is built defining the key elements of the field (classes) and their relationships (properties)^{21–23}. Then, data can be transformed to a *knowledge graph* (KG), a machine-readable structure that emphasizes the interconnection between related elements and enables the exploration of data by means of SPARQL (SPARQL Protocol and RDF Query Language) queries.

The relevance of ontologies in science has been growing along the last years, with multiple efforts arising in diverse fields such as biosciences^{24,25}, chemistry^{26–29} or Materials Science^{30–34}. KGs have also been combined with Large Language Models (LLMs) for literature mining, facilitating the ETL (extraction, transformation and load) pipeline required to process information and building foundations for AI applications, as applied to aspects such as catalysis³⁵ or electrochemistry³⁶. KG-based Retrieval Augmented Generation (RAG)³⁷ has also been introduced as a way to improve LLM responses by addressing hallucination issues. In line with this, ontologies and knowledge graphs are an ideal companion to self-driving laboratories, ensuring the FAIRness of the acquired data right from the source^{38,39}.

In this work, we propose an ontology-based scheme to handle CV data, in connection with data sourced from both traditional and self-driving laboratories, remarking the potential of semantic data schemes for cross-institutional data integration. This approach is fully in line with current European projects, such as the NFDI4Chem consortium, whose current roadmap⁴⁰, explicitly proposes the establishment of KGs for cross-institutional chemical data management as a core guideline. Apart from the data structure itself, we provide a complete toolkit tackling the generation of the knowledge graph from the corresponding data and metadata, as well as a visualization tool to leverage the data organization scheme. With these tools, DUCK conforms not only a relevant data standard in alignment with current research trends, but also a practical tool of immediate use for electrochemical laboratories.

2 Structure of the program

The Database Utility for Cyclic Voltammetry Knowledge (DUCK), outlined in Figure 2, is a platform that manages the whole data pipeline from the collection of experimental CV results to their visualization and exploration.

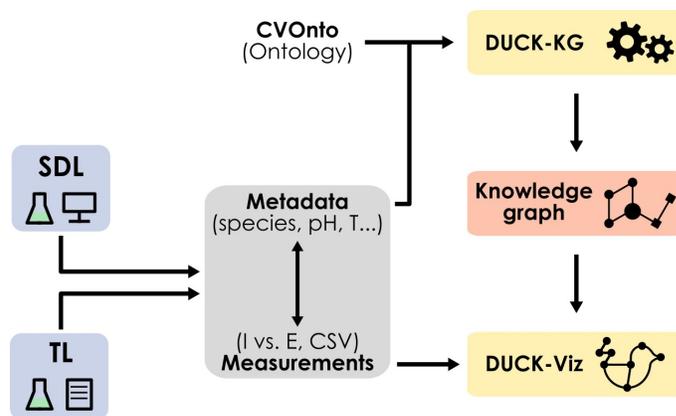


Fig. 2 Schematic depiction of the proposed data pipeline for DUCK.

The structure of DUCK is based on three main components:

- CVOnto: ontology, based on the Elementary Multiperspective Materials Ontology (EMMO)^{41,42} to structure the data.
- DUCK-KG: automated script to build a knowledge graph, based on CVOnto, from the metadata associated with each CV measurement.



- DUCK-Viz: browser-based application for the visualization and exploration of the data contained in the knowledge graph.

As depicted in Figure 2, our platform has been designed to manage data from both traditional laboratories (TLs) and self-driving laboratories (SDLs). In this way, it becomes possible to seamlessly integrate information coming from any of these two paradigms, highlighting their complementarity. On one hand, the connection with SDLs enables the direct treatment of the very large amounts of data produced by automated high-throughput setups. On the other hand, traditional setups are still an essential source of electrochemical information, both in terms of performing novel experiments and reutilizing existing data by adapting it to a common, interoperable standard.

Because of this duality, a major goal of this scheme is to remain simple enough to minimize the possible user frictions when providing metadata. Given that in many cases this information will not be fully digitalized but instead only available in laboratory notebooks, we decided to base *CVOnto* on a small subset of data to simplify the adoption of the protocol for TLs. In this way, DUCK does only require very simple metadata, provided as a spreadsheet with individual records for each CV measurement. This is a main difference with the ELN-based strategy mentioned previously⁹, where laboratories are already expected to work with a digital lab notebook. As the learning curve for the use of ELNs can be quite steep for non-experienced users, a simpler approach such as ours can be helpful to increase the engagement of the general community.

2.1 Ontology development: CVOnto

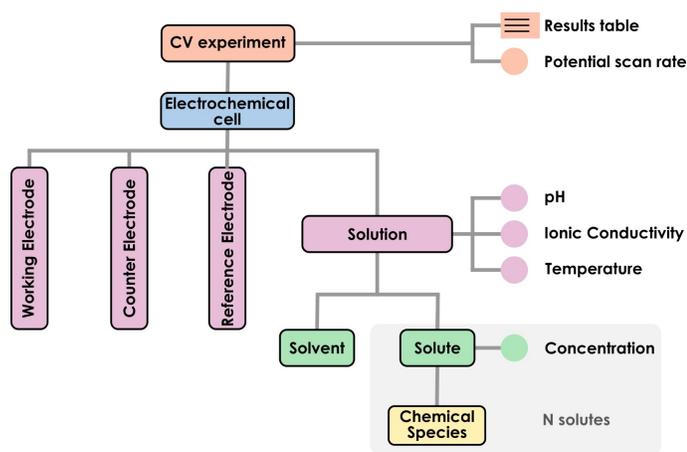


Fig. 3 Schematic depiction of the proposed data structure for CV information. Rounded rectangles correspond to main classes, and circles to data properties. Colors correspond to hierarchy levels: CV measurement (orange), electrochemical cell (blue), electrodes and solution in the cell (pink), components of the solution (green) and chemical species (yellow).

We considered ECHO (Electrochemistry Domain Ontology), the electrochemical extension⁴³ of the Elementary Multiperspective Materials Ontology (EMMO)^{41,42} as the main source of terms for our description of CV data. EMMO has emerged as the standard foundation for ontologies in Materials Science, with

domain-specific conventions providing robust community conventions: in our case, electrochemistry. Following the strategy used by the Battery Testing Ontology (BTO)³⁴, we circumvented the non-readability of EMMO terms, whose identifiers (Universal Resource Identifier, URI) do not include the name of the term, by redefining them with new URIs made equivalent to the original EMMO terms, strongly increasing the readability of future SPARQL queries.

The basic structure of our proposal contains five main levels, depicted in different colors in Figure 3. First, a specific CV experiment is defined and associated to a table containing the actual current vs. voltage data from the measurement, as well as to its voltage scan rate. For simplicity, this raw data is stored in a SQLite database mapping experiment IDs to the CSV-formatted table, which will be referenced from the KG through the corresponding URI. Then, the second level corresponds to the electrochemical cell set up for the experiment, which is related to the four elements in the third level: three electrodes (working, counter and reference) and the electrolyte solution. Regarding the electrolyte solution, we may define multiple data properties (pH, conductivity and temperature) associated to the measurement, together with the individual chemical components of this solution (fourth level). Here we find the solvent, and the *N* possible solutes dissolved in it. For each solute, we must define a *concentration*, and the corresponding chemical compound, as the final level of the data structure. The compound could then be linked to any other chemical ontology if a more expressive description of its behavior and properties was desired. In this context, any connections or extensions over the framework would require direct modifications on the CVOnto ontology. However, the flexible nature of semantic graphs streamlines the application of this kind of updates, enabling an iterative development. From this structure and considering a set of multiple CV experiments, such as an historic of measurements collected in a given laboratory, we can build a knowledge graph, for which a simplified depiction is presented in Figure 4.

This strongly simplified representation aims to highlight the connections between experiments that arise upon the generation of the graph. In this case, CVs (in orange) become connected at the fourth level of the hierarchy (in green), due to their solutions having common solvents and/or solutes. The existence of these paths between related experiments is the foundation of SPARQL queries, a fundamental tool for the further utilization of KGs for filtering and processing tasks. The power of using KGs, in comparison with alternative approaches for FAIR CV data focused on repositories⁹, lies in bringing together these related experiments in a single entity, enhancing the detection of patterns throughout the dataset. For instance, independent KGs generated from two different laboratories could be easily interconnected, exploring them altogether to find new connections.

2.2 Automated KG generation: DUCK-KG

The DUCK-KG component is a Python tool that generates knowledge graphs based on the CVOnto ontology from user-supplied metadata. This metadata is passed as a single CSV file contain-



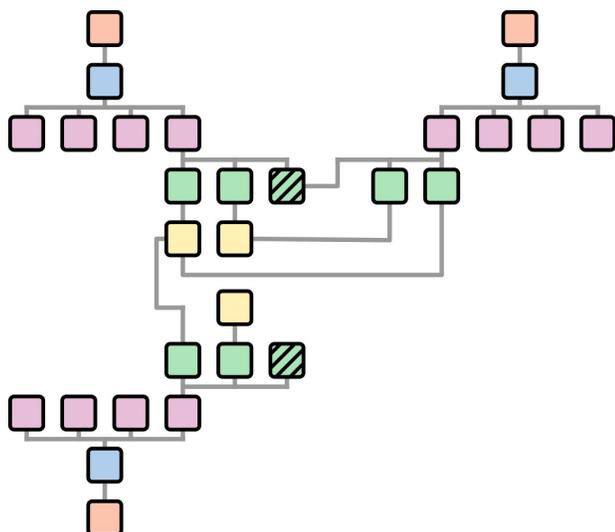


Fig. 4 Schematic depiction of a knowledge graph based on CVOnto. Colors correspond to the levels outlined in Figure 3: CV measurement (orange), electrochemical cell (blue), electrodes and solution in the cell (pink), components of the solution (green) and chemical compounds (yellow). Hatched green squares correspond to solvents, empty green squares to solutes.

ing one record for each CV experiment, containing: name, electrolytes and their concentrations, solvent, electrode composition (working, counter and reference), temperature, pH, solution conductivity, and the timestamp of the measurement. Optionally, another file might be passed, specifying the ions coming from electrolyte dissociation (e.g., $\text{AgNO}_3 \rightarrow \text{Ag}^+ + \text{NO}_3^-$). In general, we recommend to follow standard naming conventions (e.g. IUPAC) — nonetheless, DUCK does not enforce any specific criterion, ensuring total flexibility. Details on the format of these files are available in the Supporting Information.

Apart from the metadata, the actual CV measurements (I vs. E curves) should also be regarded. For simplicity, this is done through a minimal SQLite database, which contains a single table where the CV experiment IDs are mapped to the raw content of a CSV file with columns for current and voltage.

Once the KG is constructed through this component, it is possible to apply SPARQL queries to the dataset, enabling users to easily extract relevant information. While it is possible to directly write custom SPARQL code against the KG, DUCK also provides tools to simplify the generation of common queries via Python (see Supporting Information). This semantic querying showcases the main advantage of DUCK against "traditional" data sharing based on unstandardized data deposition on general-purpose repositories, leveraging data structure to enhance the discoverability of the information in the dataset.

Furthermore, additional properties beyond the basic metadata scheme, such as externally defined peak positions or further details about the cell setup (interelectrode distance, electrode conditioning, inert atmosphere, electrode geometry, number of cycles per voltammogram...) can be incorporated to the KG to enrich the final description.

2.3 Visualization of CV data: DUCK-Viz

Apart from the mere standardization of data, DUCK leverages the produced knowledge graph to enable the user to interactively explore all the experiments in the collection, via the browser-based dashboards generated by the DUCK-Viz component.

Figure 5 showcases the main features of the component. On the top side of the dashboard, the full metadata table can be directly consulted, while at the bottom an interactive plot containing all CV measurements is presented. Over this plot, it is possible to zoom, pan, or hide/show requested traces. The side panel enables the most important functionality of DUCK-Viz, which is the possibility of *filtering* the results by the presence of specific compounds, the author of the measurement, the solvent, or combinations of all these factors. When these filters are toggled, the table and the plots are automatically updated, only showcasing the entries fulfilling the requested conditions. Moreover, regarding the plots, it is also possible to *group* the selected voltammograms by their composition, obtaining individual panels for all the measurements sharing the same species and concentrations, to explore variations associated to temperature, pH or even detect outliers, where very similar conditions lead to very different I vs. E profiles. It is worth mentioning that, while in principle voltammograms are displayed just as stored, DUCK-Viz can scale the *potential* of each measurement to a common scale (Standard Hydrogen Electrode — SHE), with the potential of the reference being included in the KG upon generation if the name of the reference electrode material matches an internal database. This ensures the comparability of CV profiles coming from different sources with inconsistent reference electrodes. Going beyond these exploratory features, DUCK-Viz does also enable a more thorough interpretation of the data in the KG. The *analysis* mode allows to choose between the individual CV measurements in a given selection, automatically locating the peaks across the profiles. We employ the algorithm implemented in the SciPy package⁴⁴, with two main control parameters: the minimum inter-peak distance, which can be directly modified by the user in the DUCK-Viz interface, and the relative *prominence* of the peak. For the latter, we apply an adaptive criterion based on the largest absolute intensities in each CV on the positive and negative directions. Through this approach, we obtained a reasonable peak identification for all the measurements in the TL dataset (Section 3.1). However, limitations might arise for situations such as overlapping peaks, high background current, or irreversible electron transfer processes. To address this, apart from planning improvements for the algorithm, it is also possible to detect peaks with external procedures or manual annotation, and then include this information in the KG via DUCK-KG to be visualized in the dashboard. As we will discuss in the Applications section, this flexibility is essential to properly couple DUCK with automated closed-loop optimization workflows requiring performance indicators, such as Bayesian optimization campaigns maximizing peak potential.

DUCK can also be employed to explore the electrochemical reactions happening across a voltammogram. To obtain the necessary information, an additional component (DUCK-MP) connects to the Materials Project API to query the phase stability



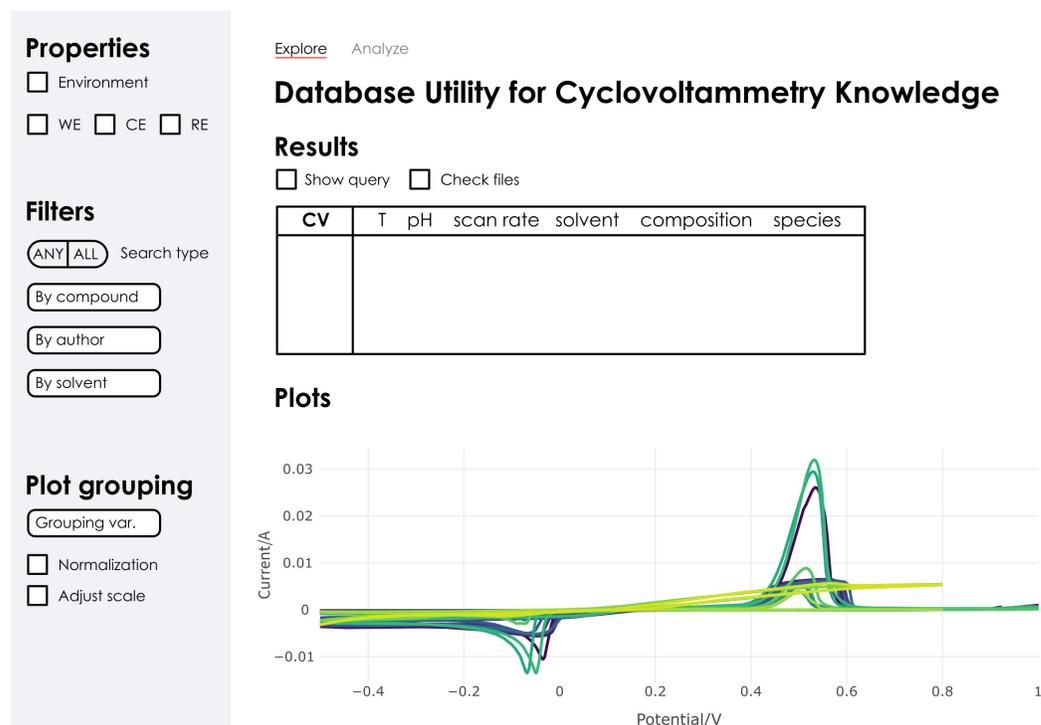


Fig. 5 Main interface of DUCK-Viz exploration dashboard.

information (stoichiometries and energies) for all the unique element combinations across the KG, which is then stored as a JSON file. From this phase information, DUCK-Viz can retrieve the stable phases for a given voltammogram, using element composition, concentrations and the pH value of the measurement to determine the corresponding regions across the Pourbaix diagram. Then, this information is added to the voltammogram in the analysis mode, highlighting regions with different phases in different colors to determine the reactions that happen along each of the peaks. The complete Pourbaix diagram around the target pH can be plotted below the CV, completing the description of the system (see Figure 8 in the *Applications* section and Figure S1 in the Supporting Information). However, we emphasize that this Pourbaix-based phase identification should be interpreted as a mostly *qualitative* guide to the possible reactions happening in solution; these diagrams represent thermodynamic equilibrium conditions, which are not often achieved under common CV scan rates (10 - 500 mV/s) due to kinetic phenomena such as overpotential or mass transport limitations. In this way, the potentials observed in the cell may differ from the standard ones. Thus, we recommend users to combine the interpretation of DUCK analysis results with complementary techniques (e.g. spectroelectrochemistry) for definitive mechanistic interpretation.

2.4 Implementation and future design

The knowledge graphs generated and employed by DUCK are serialized in the RDF/Turtle (RDF — Resource Description Language format). The current approach regards a fully local setup, directly employing Python tools for the establishment of the triple store used for KG querying. Nonetheless, in the near future we

aim to directly integrate DUCK with graph database servers (e.g. Blazegraph, Neo4J, etc), setting up a community resource to host CV data from multiple sources and developing an API for programmatic data access. A main goal for this future community-wide database is to be robust data source for the development of ML/AI models based on cyclic voltammetry. While as aforementioned previous studies have already employed CVs for ML-based studies^{4,19,20}, ontology-structured data will enable more systematic and reproducible ML workflows to be developed, so complex problems like electrodeposition mechanisms can be addressed. For instance, although out of the scope of this communication, DUCK-generated KGs might be employed to apply RAG methods³⁷ for higher-quality LLM interaction.

3 Applications

As a proof of concept for our methodology, we constructed two datasets from different sources and employed them for the generation of knowledge graphs based on CVOnto.

- TL dataset: collection of **130** CV experiments from the FINDER group, corresponding to the characterization of systems used as target for *electrodeposition* experiments.
- SDL dataset: collection of **79** experiments performed in-situ in a SDL setup, considering metal ions and ligands in solution following previous works¹⁷.

We then created KGs for each of the individual datasets, via DUCK-KG, and explored different aspects of the corresponding graphs and their associated CV plots.



3.1 TL dataset

This dataset focuses on electrochemical measurements for solutions employed for electrodeposition processes, targeting materials like bismuth telluride^{45–50}, zinc oxide^{51,52}, copper-nickel alloys⁵³, PEDOT⁵⁴ and copper and silver selenides^{55,56}. Here, CVs provide essential insights on the behavior of the solution, determining the peak potentials at which deposition should take place — we refer the reader to the publications cited in the previous paragraph for further details about the specific setups and applications of each system. Thus, the dataset comprises multiple metals (Ag, Cu, Ni, Zn, Fe) and semimetals (Bi, Se, Te), at varying concentrations and pH values.

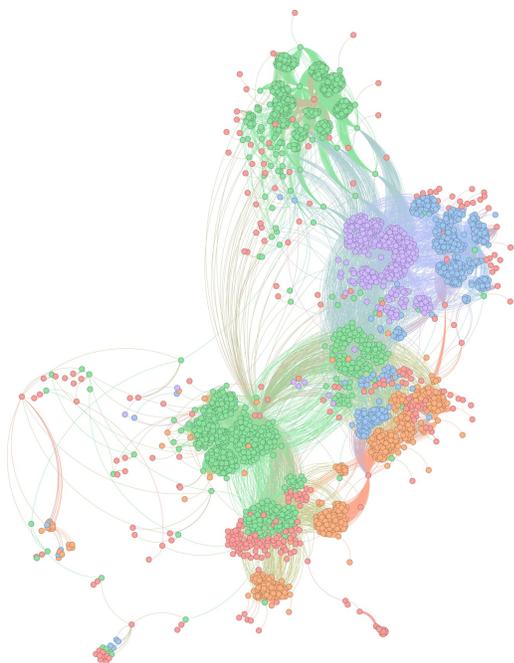


Fig. 6 Representation of the complete KG for the TL dataset, colored by the source ontology of each node: green for CVOnto (including the *individuals* for the dataset), blue for EMMO, orange for echo, red for CHEM (Chemical Substance Domain Ontology), and purple for other terms (including data properties and general definitions outside the main namespaces).

The corresponding KG (Figure 6) contains 3312 nodes and 10865 edges. While this view clarifies the deep interconnection of the underlying structure, it is of course quite hard to parse. Nonetheless, this structure can be exploited through queries to retrieve specific node types, following the scheme in Figure 3, and thus determine more chemically meaningful connections. To illustrate this, we considered a random selection of 50 of the CVs in the dataset, and extracted only two core elements: the CV itself, and the chemical compounds associated to its corresponding solution (Figure 7)

This representation shows the connection between voltammeteries in the dataset sharing common electrolytes, that get organized in six distinct subgraphs. The largest clusters correspond to Bi/Te solutions (25 CVs, middle left) and to Ag/Se and Cu/Se solutions (13 CVs, middle right), which are brought together due to sharing H_2SeO_3 as one of the electrolytes. Smaller subsets can

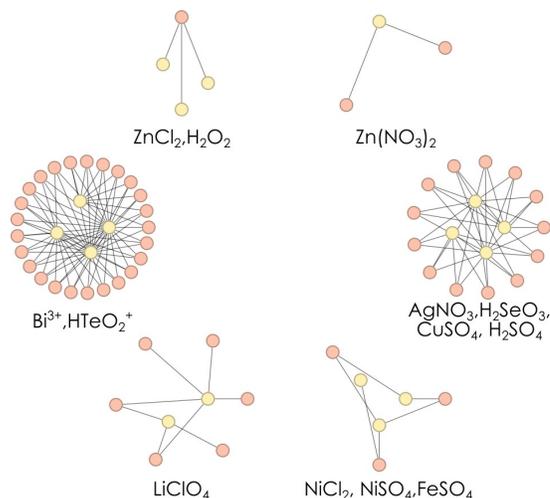


Fig. 7 Example subset of a knowledge graph based on CVOnto, representing CV measurements (orange) and chemical compounds (yellow) for 50 randomly selected voltammograms from the TL dataset.

be identified for LiClO_4 (6 CVs, bottom left), Ni/Fe (4, bottom right), and finally for two distinct Zn salts, chloride and nitrate, with one CV each (top). Thus, we get an immediate glance into the underlying structure of the data, revealing the presence of completely distinct sets of voltammeteries at the chemical level in the TL dataset (as separate clusters), as well as the amount of measurements available for each of them. In this way, we highlight how the complex, highly connected KG structure (Figure 6) can be further simplified to provide more readable subgraphs containing relevant pieces of information. While this might seem trivial, we believe that it confirms the interest of employing graph databases (KGs) to store and process this kind of data. Beyond the mere visualization, these subgraphs aim to illustrate possible query patterns that might reveal important trends throughout large datasets.

Furthermore, taking into account that the measurements of this dataset were made with electrodeposition in mind, the *analysis* mode of DUCK-Viz strongly simplifies the assessment of the phases associated to each peak in the CV profile, streamlining the selection of the most adequate potentials. This is illustrated in Figure 8, which summarizes the main plots generated in the analysis mode for an example CV in the dataset, corresponding to a Bi/Te system⁵⁷.

Following the previously described procedure (Section 2.3), DUCK-Viz automatically identifies two well-defined peaks in the CV, at -0.07 V and 0.55 V. Then, the stable phases at the target pH (1.0) throughout the whole potential range are identified by means of the Pourbaix diagram (lower part of Figure 8), and then compared with the located intensity peaks. From the phase information, DUCK-Viz matches the peak at -0.07 V with the formation of a solid phase of bismuth telluride, with the one at 0.55 V involving two aqueous anions ($[\text{BiO}]^+$ and $[\text{Te}(\text{OH})_3]^+$). Thus, our strategy enables a direct identification of the actual electrodeposition peak at -0.07 V, where a net *reduction* of tellurium happens going from Te(IV) to Te(-II) to form solid Bi_2Te_3 that will then be deposited. While this application might seem trivial, involving



a well-studied electrodepositing system (bismuth telluride) and a very simple CV, it showcases the potential of DUCK as an analysis tool for the exploration of new systems, especially when coupled to the data acquisition phase to directly load new experiments.

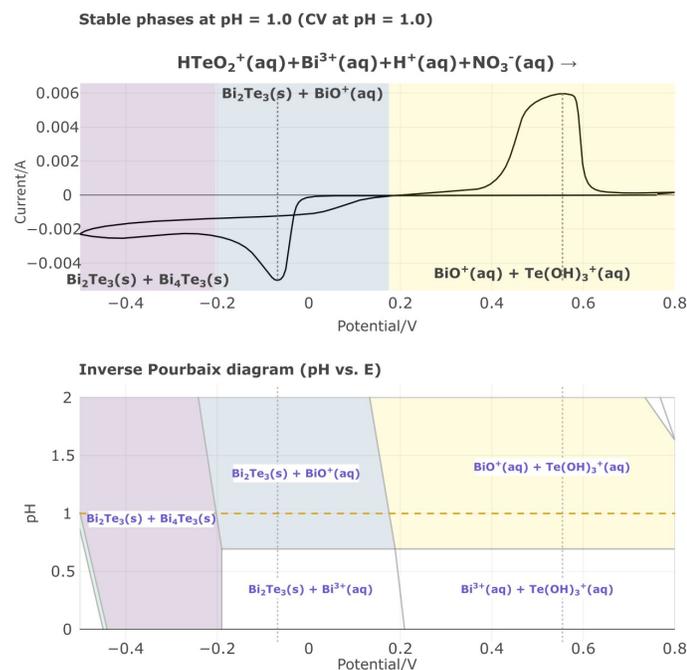


Fig. 8 Results of the DUCK-Viz analysis dashboard, showcasing bismuth(III)/tellurium speciation for a example voltammogram in the TL dataset.

3.2 SDL dataset

There are two main advantages associated to using DUCK as a tool for the management of SDL-produced cyclic voltammeteries. On the one hand, the ontology is employed as a database scheme to properly organize the results generated by the laboratory, providing the data backend. On the other, the visualization frontend of DUCK-Viz enables the easy supervision of the generated results (human-in-the-loop), which is an essential asset for the actual applicability of SDLs. To demonstrate this, we have developed a proof of concept in which we have directly integrated DUCK with the MEDUSA platform¹⁷ for the automated synthesis and electrochemical characterization of metal - ligand complexes. Under this framework, we have developed a simple Bayesian optimization strategy, starting from the basic parameters of the Honegumi framework⁵⁸, to showcase the optimization of the maximum peak potential^{15,59} across a set of different metal centers (V, Ni and Cu), ligands (en — ethylenediamine) and pH conditions (buffers centered at pH 6 — MES — and 7 — HEPES: see Methods section for more details). Following the same core principles as for digitalizing TL results (flexibility and simple adaptation to existing frameworks), we implemented a simple interface between MEDUSA and DUCK, working in batches. Thus, the optimization and laboratory control operations are carried out independently via MEDUSA, with results being loaded into a KG periodically. Then, DUCK-Viz is used to follow optimization results, deciding

whether to continue the campaign or to modify it if any issues are observed.

We carried out three different optimization campaigns, labelled as #1, #2 and #3.

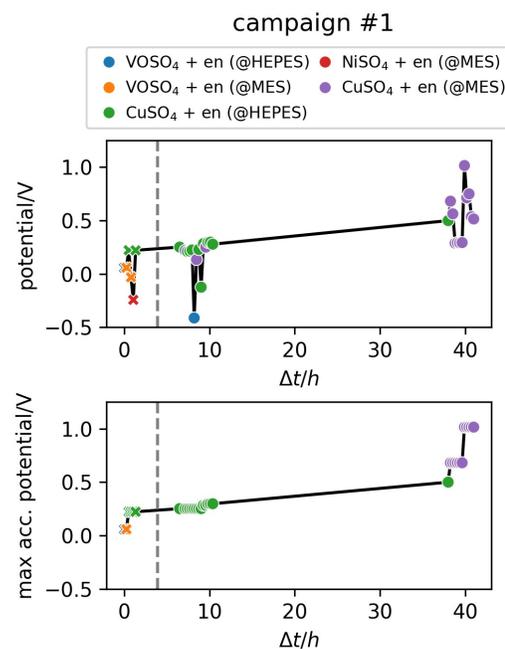


Fig. 9 Optimization campaign #1. Above, largest peak potential for each experiment against time, colored by the nature of the involved compounds. Below, maximum accumulated peak potential across the campaign. Dashed lines separate the initial exploration region (crosses) from the optimized points (circles).

In the first campaign (Figure 9), we showcase a moderate improvement from the initial region, without altering the nature of the mixture (CuSO₄ + en (@HEPES), green points), up until a total time of around 40 hours, when the optimal buffer switches to MES, at a more acidic pH (purple points), with a larger increment of the potential. However, careful inspection of these voltammograms reveals limitations of the peak detection algorithm, which identifies an ill-defined peak for the largest accumulated potential (Supporting Information, Figure S3). Far from a drawback, this reveals the strengths of DUCK as a platform to enhance the setup of automated experiments, enabling the easy identification of mistakes to plan correcting actions. From there, we carried two further campaigns, increasing the number of CV cycles per measurement from 2 (as in campaign #1) to 5. This improves peak shape, ensuring a more consistent detection of the target potential, although it also accelerates electrode degradation, which occurs after a smaller number of optimization steps (and marks the end of the campaign). Results for these campaigns are shown in the Supporting Information (Figure S4). #2 can be considered as a failed campaign, where the potential could not be optimized from the stochastically selected initial exploration. In contrast, #3 shows the desired behavior of a BO procedure, showcasing a clean optimization of the target property, although the overall potential increase is limited by the reduced number of experiments in the campaign. This behavior could be improved by designing



longer campaigns with longer-lasting electrodes: nonetheless, we should recall that the goal of this proof of concept is not to achieve a perfect optimization, but instead to showcase the capabilities of DUCK as an integrated platform to manage and follow this kind of procedures in the context of SDLs. In this way, the tool allows the scientist to monitor the adequacy of the optimization and seamlessly access all produced data. As a final note, DUCK does also enhance data sharing and, consequently, the reuse of the collected results, contributing to the openness and FAIRness of modern SDLs.

Conclusions and outlook

In this work we introduce DUCK, a toolkit to manage CV data that relies on the generation of knowledge graphs to organize CV datasets. On one hand, we introduce CVOnto, an ontology based on the Elementary Multiperspective Materials Ontology (EMMO) which defines the fundamental classes and relationships required to organize this kind of information. Together with the ontology, the DUCK-KG utility enables a straightforward generation of KGs from CSV files specifying the metadata associated with each CV measurement in the dataset. This tool has been tested against two dataset from two different sources: traditional labs and self-driving laboratories. In this way, DUCK provides a simple way to standardize information and unify these two co-existing paradigms.

On the other hand, we also introduce DUCK-Viz, an interactive browser-based visualization tool that enables the direct exploration and analysis of CV data in the form of a knowledge graph. DUCK-Viz provides a simple interface between the user and the KG, lifting barriers to the adoption of the standardization scheme, as well as relevant tools to simplify, for instance, the determination of the electrochemical processes occurring along the full CV profile. For SDLs, this tool offers an easy-to-use frontend to monitor the results produced by the automated setup, as we demonstrate with a set of Bayesian optimization campaigns.

By combining all the aforementioned aspects, DUCK becomes a unified platform to handle the complete data pipeline for the management of CVs under the FAIR principles. Although the current installment is limited to local laboratory-wide databases, future developments aim to deploy a global server collecting data from multiple sources, effectively implementing a community-wide CV database. While the current work is presented as a proof of concept, the flexibility of ontologies ensures the scalability and adaptability of the overall framework to iteratively develop the database schema and capture more complex, nuanced processes. In parallel to this schema extension, we also intend to expand the DUCK ecosystem by building new tools beyond the actual components, incorporating new functionalities of interest for general data analysis in electrochemistry, such as for instance a more specialized component for voltammogram peak characterization.

Methods

SDL Experimental The experimental setup was modified from previously published electrochemistry SDL based on the MEDUSA framework^{17,60}. In the context of this work, a simplified version with one syringe pump and one electrochemistry cell and one po-

tentiostat was used, and all the syntheses and measurements were performed on a single platform automatically without human intervention except for refilling the stock solution and changing the electrodes.

The following connections were made the 12 ports of the syringe pump, namely 1) deionized water, 2) 3 mol/L NaCl, 3) 0.1 mmol/L CuSO₄, 4) 0.1 mol/L NiSO₄, 5) 0.1 mol/L VOSO₄, 6) 0.3 mol/L ethylenediamine (en), 7) air (for pressure balancing), 8) 1 mol/L 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethane-1-sulfonic acid (HEPES) buffer, 9) 1 mol/L 2-morpholin-4-ylethanesulfonic acid (MES) buffer, 10) reaction vial, 11) electrochemistry cell, 12) waste.

Complexation reactions were performed in the following steps in the reaction vessel: 0) clean up the reaction vial and electrochemical cell by washing with 1 mL water for 3 times each, 1) add V_{buffer} mL 1 mol/L buffer solution, 2) add 0.4 mL 3 mol/L NaCl solution, 3) add V_{ligand} mL 0.3 mol/L en ligand solution, 4) add V_{metal} mL 0.1 mol/L metal solution, 5) add $(0.8 - V_{buffer} - V_{ligand} - V_{metal})$ mL water, 6) purge 1.0 mL N₂ through the reaction mixture and then draw and dispense 1 mL of the reaction mixture; this step is repeated 3 times to ensure a good mix. After the complexation is finished, the reaction mixture is transferred to the electrochemistry flow cell for CV measurement, starting at open circuit potential (ocp), in the range of -1.2 to 1.2 V for 2 (first campaign) or 5 (second and third campaign) cycles at 500 mV/s. Considering the short distance between electrodes and the relatively high conductivity of NaCl aqueous electrolyte solution, the iR drop in our measurement is expected to be low and was not considered.

Bayesian optimization campaigns were performed using the *ax*-platform package, employing a Sobol generator model with 6 trials. Peak detection for maximum potential optimization was performed with the code in the following repository: <https://github.com/aurelienblanc2/Potentiostat-datapipeline>. The amount ranges for the buffer, ligand, and metal complex solutions, in mL of the reference solutions detailed above, are 0.05 to 0.20 mL (metal and ligand) and 0.20 to 0.40 mL (buffer).

TL Experimental Measurements from this dataset were collected from experiments in the FINDER group from 2010 to 2025. While no specific details on the individual setups will be provided, some general guidelines are:

- iR drops are not negligible in general, in opposition to the SDL dataset. Consequently, current densities are larger in the TL dataset.
- Scan rates range from 5 to 200 mV/s.
- All measurements start at open circuit potential (ocp) with pristine working electrodes.

Conflicts of interest

There are no conflicts to declare.

Data availability

The DUCK package, including all the modules discussed in the text, is available at GitLab <https://gitlab.com/dgarayr/duck>.



The TL and SDL datasets are available in Zenodo 10.5281/zenodo.18015308.

The peak-picking module employed for the SDL experiments is available at GitHub <https://github.com/aurelienblanc2/Potentiostat-datapipeline>.

Acknowledgements

DGR and MMG acknowledge the ERC Advanced Grant POWER-byU (Grant No. 101052603) for funding. SPG and HH acknowledge that this research was partially supported by funding provided to the University of Toronto's Acceleration Consortium through the Canada First Research Excellence Fund (CFREF) (CFREF-2022-00042). We thank Dr. Cristina V. Manzano, Dr. Olga Caballero-Calero, Elena Pérez-Picazo and Nadia Pastor for their contributions to the TL dataset. We also thank Prof. Núria López for valuable initial guidance and discussion.

References

- R. S. Nicholson, *Analytical Chemistry*, 1965, **37**, 1351–1355, DOI: 10.1021/ac60230a016.
- N. Elgrishi, K. J. Rountree, B. D. McCarthy, E. S. Rountree, T. T. Eisenhart and J. L. Dempsey, *Journal of Chemical Education*, 2018, **95**, 197–206, DOI: 10.1021/acs.jchemed.7b00361.
- H. Yamada, K. Yoshii, M. Asahi, M. Chiku and Y. Kitazumi, *Electrochemistry*, 2022, **90**, 102005–102005, DOI: 10.5796/electrochemistry.22-66082.
- D. A. Rosser and K. C. Leonard, *ACS Electrochemistry*, 2025, **1**, 1038–1043, DOI: 10.1021/acselectrochem.5c00012.
- L. P. Bicelli, B. Bozzini, C. Mele and L. D'Urzo, *International Journal of Electrochemical Science*, 2008, **3**, 356–408, DOI: 10.1016/S1452-3981(23)15460-5.
- S. Fletcher, C. Halliday, D. Gates, M. Westcott, T. Lwin and G. Nelson, *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, 1983, **159**, 267–285, DOI: 10.1016/S0022-0728(83)80627-5.
- V. A. Isaev, O. V. Grishenkova and Y. P. Zaykov, *Journal of Solid State Electrochemistry*, 2018, **22**, 2775–2778, DOI: 10.1007/s10008-018-3989-9.
- M. D. Wilkinson *et al.*, *Scientific Data*, 2016, **3**, 160018, DOI: 10.1038/sdata.2016.18.
- D. Herrmann, P. Hodapp, M. Starman, P.-C. Huang, C.-L. Lin, L. B. Q. Le, T. G. Fischer, C. Bizzarri, P. Röse, N. Oppel, J. Klar, P. Tremouilhac, L. Holzhauer, S. Herres-Pawlis, A. Hoffmann, T. Seitz, A. Dorn, K. Zeitler, N. Jung and S. Bräse, *Chemical Science*, 2025, **16**, 4430–4441, DOI: 10.1039/D4SC08620A.
- P. Tremouilhac, A. Nguyen, Y.-C. Huang, S. Kotov, D. S. Lütjohann, F. Hübsch, N. Jung and S. Bräse, *Journal of Cheminformatics*, 2017, **9**, 54, DOI: 10.1186/s13321-017-0240-0.
- NFDI4Chem, *RADAR4Chem*, <https://radar4chem.radar-service.eu/radar/en/home>, 2025, DOI: null.
- F. Häse, L. M. Roch and A. Aspuru-Guzik, *Trends in Chemistry*, 2019, **1**, 282–291, DOI: 10.1016/j.trechm.2019.02.007.
- M. Abolhasani and E. Kumacheva, *Nature Synthesis*, 2023, **2**, 483–492, DOI: 10.1038/s44160-022-00231-0.
- G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid and A. Aspuru-Guzik, *Chemical Reviews*, 2024, **124**, 9633–9732, DOI: 10.1021/acs.chemrev.4c00055.
- R. J. Hickman, M. Sim, S. Pablo-García, G. Tom, I. Woolhouse, H. Hao, Z. Bao, P. Bannigan, C. Allen, M. Aldeghi and A. Aspuru-Guzik, *Digital Discovery*, 2025, **4**, 1006–1029, DOI: 10.1039/D4DD00115J.
- X. Zhang *et al.*, *Nexus*, 2025, **2**, 100083, DOI: 10.1016/j.ynexus.2025.100083.
- S. Pablo-García, Á. García, G. D. Akkoc, M. Sim, Y. Cao, M. Somers, C. Hatrick, N. Yoshikawa, D. Dworschak, H. Hao and A. Aspuru-Guzik, *Device*, 2025, **3**, 100567, DOI: 10.1016/j.device.2024.100567.
- M. A. Sadeghi, R. Zhang and J. Hatrick-Simpers, *Journal of Open Source Software*, 2025, **10**, 6256, DOI: 10.21105/joss.06256.
- B. B. Hoar, W. Zhang, S. Xu, R. Deeba, C. Costentin, Q. Gu and C. Liu, *ACS Measurement Science Au*, 2022, **2**, 595–604, DOI: 10.1021/acsmeasuresciau.2c00045.
- B. B. Hoar, W. Zhang, Y. Chen, J. Sun, H. Sheng, Y. Zhang, Y. Chen, J. Y. Yang, C. Costentin, Q. Gu and C. Liu, *ACS Electrochemistry*, 2025, **1**, 52–62, DOI: 10.1021/acselectrochem.4c00014.
- T. R. Gruber, *Knowledge Acquisition*, 1993, **5**, 199–220, DOI: 10.1006/knac.1993.1008.
- T. Berners-Lee, J. Hendler and O. Lassila, *Scientific American*, 2001, **284**, 34–43, DOI: 10.1038/scientificamerican0501-34.
- C. Bizer, T. Heath and T. Berners-Lee, in *Semantic Services, Interoperability and Web Applications*, IGI Global, 2011, pp. 205–227, DOI: 10.4018/978-1-60960-593-3.ch008.
- J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes and C. Steinbeck, *Nucleic Acids Research*, 2016, **44**, D1214–D1219, DOI: 10.1093/nar/gkv1031.
- S. Carbon *et al.*, *Nucleic Acids Research*, 2021, **49**, D325–D334, DOI: 10.1093/nar/gkaa1113.
- C. Pacht, N. Frank, J. Breitbart and S. Bräse, *ArXiv*, 2020, DOI: 10.48550/arXiv.2002.03842.
- F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca and M. Kraft, *Journal of Chemical Information and Modeling*, 2020, **60**, 108–120, DOI: 10.1021/acs.jcim.9b00960.
- P. Strömert, J. Hunold, A. Castro, S. Neumann and O. Koepfer, *Pure and Applied Chemistry*, 2022, **94**, 605–622, DOI: 10.1515/pac-2021-2007.
- D. Garay-Ruiz and C. Bo, *Journal of Cheminformatics*, 2022, **5**, 1–12, DOI: 10.1186/s13321-022-00610-x.
- P. Del Nostro, G. Goldbeck and D. Toti, *Applied Ontology*, 2022, **17**, 401–421, DOI: 10.3233/AO-220271.
- P. Del Nostro, G. Goldbeck, A. Pozzi and D. Toti, *Applied Ontology*, 2023, **18**, 99–118, DOI: 10.3233/AO-230024.
- A. De Baas, P. D. Nostro, J. Friis, E. Ghedini, G. Goldbeck,



- I. M. Paponetti, A. Pozzi, A. Sarkar, L. Yang, F. A. Zaccarini and D. Toti, *IEEE Access*, 2023, **11**, 120372–120401, DOI: 10.1109/ACCESS.2023.3327725.
- 33 A. Kondinski, P. Rutkevych, L. Pascazio, D. N. Tran, F. Farazi, S. Ganguly and M. Kraft, *Digital Discovery*, 2024, **3**, 2070–2084, DOI: 10.1039/D4DD00166D.
- 34 P. Del Nostro, G. Goldeck, F. Kienberger, M. Moertelmaier, A. Pozzi, N. Al-Zubaidi-R-Smith and D. Toti, *Computers in Industry*, **164**, 104203, DOI: 10.1016/j.compind.2024.104203.
- 35 A. S. Behr, D. Chernenko, D. Koßmann, A. Neyyathala, S. Hanf, S. A. Schunk and N. Kockmann, *Catalysis Science & Technology*, 2024, **14**, 5699–5713, DOI: 10.1039/D4CY00369A.
- 36 S. X. Leong, S. Pablo-García, B. Wong and A. Aspuru-Guzik, *Matter*, 2025, 102331, DOI: 10.1016/j.matt.2025.102331.
- 37 X. Zhu, Y. Xie, Y. Liu, Y. Li and W. Hu, *Knowledge Graph-Guided Retrieval Augmented Generation*, 2025, DOI: 10.48550/arXiv.2502.06864.
- 38 J. Bai, L. Cao, S. Mosbach, J. Akroyd, A. A. Lapkin and M. Kraft, *JACS Au*, 2022, **2**, 292–309, DOI: 10.1021/jacsau.1c00438.
- 39 J. Bai, S. Mosbach, C. J. Taylor, D. Karan, K. F. Lee, S. D. Rihm, J. Akroyd, A. A. Lapkin and M. Kraft, *Nature Communications*, 2024, **15**, 462, DOI: 10.1038/s41467-023-44599-9.
- 40 C. Steinbeck *et al.*, *Research Ideas and Outcomes*, 2025, **11**, e177037, DOI: 10.3897/rio.11.e177037.
- 41 E. Ghedini, J. Friis, G. Goldbeck, F. L. Bleken, O. M. Roscioni, S. Clark, S. Stier and I. M. Paponetti, *Emmo-Repo/EMMO: 1.0.0*, <https://zenodo.org/doi/10.5281/zenodo.5730500>, DOI: 10.5281/zenodo.5730500.
- 42 E. Ghedini, G. Goldbeck, J. Friis, A. Hashibon and G. Schmitz, *European Materials Modelling Ontology*, <https://emmo-repo.github.io/versions/1.0.0-beta/emmo.html>, DOI: null.
- 43 S. Clark, X. Raynaud, E. Flores, F. L. Bleken, J. Friis and C. W. Andersen, *Emmo-Repo/Domain-Electrochemistry: 0.27.0-Beta*, <https://zenodo.org/doi/10.5281/zenodo.14875879>, DOI: 10.5281/ZENODO.14875879.
- 44 P. Virtanen *et al.*, *Nature Methods*, 2020, **17**, 261–272, DOI: 10.1038/s41592-019-0686-2.
- 45 M. S. Martín-González, A. L. Prieto, R. Gronsky, T. Sands and A. M. Stacy, *Journal of The Electrochemical Society*, 2002, **149**, C546, DOI: 10.1149/1.1509459.
- 46 C. V. Manzano, A. A. Rojas, M. Decepeida, B. Abad, Y. Feliz, O. Caballero-Calero, D.-A. Borca-Tasciuc and M. Martín-Gonzalez, *Journal of Solid State Electrochemistry*, 2013, **17**, 2071–2078, DOI: 10.1007/s10008-013-2066-7.
- 47 B. Abad, M. Rull-Bravo, S. L. Hodson, X. Xu and M. Martín-Gonzalez, *Electrochimica Acta*, 2015, **169**, 37–45, DOI: 10.1016/j.electacta.2015.04.063.
- 48 O. Caballero-Calero, M. Mohner, M. Casas, B. Abad, M. Rull, D. A. Borca-Tasciuc and M. Martín-González, *Materials Today: Proceedings*, 2015, **2**, 620–628, DOI: 10.1016/j.matpr.2015.05.087.
- 49 C. V. Manzano, B. Abad and M. Martín-González, *Journal of The Electrochemical Society*, 2018, **165**, D768–D773, DOI: 10.1149/2.1131814jes.
- 50 C. V. Manzano, M. N. Polyakov, J. Maiz, M. H. Aguirre, X. Maeder and M. Martín-González, *Science and Technology of Advanced Materials*, 2019, **20**, 1022–1030, DOI: 10.1080/14686996.2019.1671778.
- 51 C. V. Manzano, D. Alegre, O. Caballero-Calero, B. Alén and M. S. Martín-González, *Journal of Applied Physics*, 2011, **110**, 043538, DOI: 10.1063/1.3622627.
- 52 C. Manzano, O. Caballero-Calero, S. Hormeño, M. Penedo, M. Luna and M. S. Martín-González, *The Journal of Physical Chemistry C*, 2013, **117**, 1502–1508, DOI: 10.1021/jp3107099.
- 53 C. V. Manzano, P. Schürch, L. Pethö, G. Bürki, J. Michler and L. Philippe, *Journal of The Electrochemical Society*, 2019, **166**, E310–E316, DOI: 10.1149/2.0961910jes.
- 54 C. V. Manzano, O. Caballero-Calero, A. Serrano, P. M. Resende and M. Martín-González, *Nanomaterials*, 2022, **12**, 4430, DOI: 10.3390/nano12244430.
- 55 C. V. Manzano, C. Llorente Del Olmo, O. Caballero-Calero and M. Martín-González, *Sustainable Energy & Fuels*, 2021, **5**, 4597–4605, DOI: 10.1039/D1SE01061A.
- 56 E. Pérez-Picazo, M. Á. Tenaguillo, K. Lohani, M. Martín-González and C. V. Manzano, (submitted).
- 57 O. Caballero-Calero, P. Díaz-Chao, B. Abad, C. Manzano, M. Ynsa, J. Romero, M. M. Rojo and M. Martín-González, *Electrochimica Acta*, 2014, **123**, 117–126, DOI: 10.1016/j.electacta.2013.12.185.
- 58 S. G. Baird, A. R. Falkowski and T. D. Sparks, *arXiv*, 2025, DOI: 10.48550/ARXIV.2502.06815.
- 59 A. Blanc, *Potentiostat Data Pipeline*, <https://github.com/aurelienblanc2/Potentiostat-datapipeline>, 2025, DOI: null.
- 60 F. Strieth-Kalthoff *et al.*, *Science*, 2024, **384**, eadk9227, DOI: 10.1126/science.adk9227.



Data availability

The DUCK package, including all the modules discussed in the text, is available at GitLab <https://gitlab.com/dgarayr/duck>.

The TL and SDL datasets are available in Zenodo 10.5281/zenodo.18015308.

The peak-picking module employed for the SDL experiments is available at GitHub [https://github.com/aurelienblanc2/](https://github.com/aurelienblanc2/Potentiostat-datapipeline)

Potentiostat-datapipeline.

