Good Practice Report

Andrew P. Cornell, Sunghwan Kim, Jordi Cuadros, Ehren C. Bucholtz, Harry E. Pence, Rudy Potenzone and Robert E. Belford*

IUPAC International Chemical Identifier (InChI)-related education and training materials through InChI Open Education Resource (OER)

https://doi.org/10.1515/cti-2023-0009
Received March 9, 2023; accepted December 11, 2023; published online January 3, 2024

Abstract: The IUPAC International Chemical Identifier (InChI) is a structure-based chemical identifier that encodes various aspects of a chemical structure into a hierarchically layered line notation. Because InChI is non-proprietary, open-source, and freely available to everyone, it is adopted in popular chemical information resources and software programs. This paper describes the InChI Open Education Resource (OER) (https://www.inchi-trust.org/oer/), designed to provide educators and other interested parties with resources, training material, and information related to InChI. Currently, the OER contains over 100 materials collected from various sources and provides users with search, filtering, and sorting functionalities to locate specific records. New relevant materials can be suggested by anyone, allowing the scientific community to share and find InChI-related resources. This paper will show how to use the InChI OER tag taxonomy to filter content and demonstrate two resources within the InChI OER; the ChemNames2LCSS Google Sheet and the InChILayersExplorer, an Excel spreadsheet that breaks an InChI into its layers. While the InChI OER is of value to a broader chemistry community, this paper seeks to reach out to chemical educators and provide them with an understanding of InChI and its role in the practice of science.

Keywords: IUPAC; International Chemical Identifier (InChI); Open Education Resource (OER); PubChem; Laboratory Chemical Safety Sheet (LCSS)

1 Introduction

Today's students will be entering a world and workplace where decision-making and the practice of science will become ever increasingly intertwined with big data, and yet this topic is noticeably missing in the undergraduate chemistry curriculum. In 2009 Microsoft Research published "The Fourth Paradigm: Data-Intensive Scientific

***Corresponding author: Robert E. Belford**, Department of Chemistry, University of Arkansas at Little Rock, Little Rock, AR 72204, USA, E-mail: rebelford@ualr.edu. https://orcid.org/0000-0002-4933-6379
Andrew P. Cornell, Department of Chemistry, University of Arkansas at Little Rock, Little Rock, AR 72204, USA, E-mail: apcornell@ualr.edu. https://orcid.org/0000-0003-1636-4784
Sunghwan Kim, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, E-mail: kimsungh@ncbi.nlm.nih.gov. https://orcid.org/0000-0001-9828-2074
Jordi Cuadros, Department of Quantitative Methods, IQS Universitat Ramon Llull, Barcelona, 08017, Spain, E-mail: jordi.cuadros@iqs.url.edu. https://orcid.org/0000-0001-6513-9140
Ehren C. Bucholtz, Department of Basic Sciences, University of Health Sciences and Pharmacy in St. Louis, St. Louis, MO 63110, USA, E-mail: ehren.bucholtz@uhsp.edu. https://orcid.org/0000-0002-3261-2947
Harry E. Pence, Department of Chemistry and Biochemistry, State University of New York at Oneonta, Oneonta, NY, 13820, USA, E-mail: Harry.Pence@oneonta.edu. https://orcid.org/0000-0002-0412-9018
Rudy Potenzone, InChI Trust, Cambridge CB2 1EZ, UK, E-mail: rudypoten@me.com. https://orcid.org/0000-0002-2051-9518

Discovery", an anthology of essays on data science in memory of the data scientist Jim Gray (Hey et al., 2009). The introduction "Jim Gray on eScience: A Transformed Scientific Method" was based on a 2007 talk he gave in Mountain View California, which defined four paradigms of science.

- 1st Paradigm: Empirical Science (thousands of years old)
- 2nd Paradigm: Theoretical Science (centuries old)
- 3rd Paradigm: Computational Science (decades old)
- 4th Paradigm: eScience/data exploration (years old)

Roughly 220 years earlier Lavoisier stated in his 1790 book, "Elements of Chemistry, in a New Systematic Order, Containing All the Modern Discoveries" (Lavoisier, 1790/2009), that:

> As ideas are preserved and communicated by means of words, it necessarily follows that we cannot improve the language of any science, without at the same time improving the science itself; neither can we, on the other hand, improve a science without improving the language or nomenclature which belongs to it.

At the turn of the second millennium, it was clear that a new form of chemical representation was evolving that aligned with the third and fourth paradigms of science and could be used by machines to both store data and perform complex computations. Digitized molecular representation essentially took three different forms: molecular graphs, line notation, and index numbers.

- A molecular graph was the most data-intensive format, which at a minimum contained two data tables, an atom table (identifying the atoms in a molecule) and a connection table (identifying the bonds).
- Line notations used rules to code the information of the molecular graph into a one-dimensional string of characters, with the Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988, 1990; Weininger et al., 1989) being the most common. Issues of canonicalization arose as different software would often generate different SMILES for the same molecules (caffeine was shown to have up to 4160 different SMILES strings (Sayle, 2014)), and there were even issues with canonical SMILES as the string depended on the canonicalization algorithm.
- The simplest way of representing a molecule on a computer was using an index number or alphanumeric string like a Chemical Abstract Services (CAS) registry number (https://www.cas.org/support/documentation/chemical-substances/faqs), which requires a lookup table and had no structural information coded into it.

Lavoisier's statement was prophetic and as the chemical sciences evolved beyond empirical and theoretical paradigms there was an interdependent and concurrent need for the chemical nomenclature to evolve. That is, the Fourth Paradigm Science could not effectively contribute to scientific discovery if communication across databases and software agents was impeded when different identifiers were used to describe the same chemical. The International Union of Pure and Applied Chemistry (IUPAC) is the international authority on chemical nomenclature and terminology whose recommendations are made public through the IUPAC Color Books (https://iupac.org/what-we-do/books/color-books/) and the journal Pure and Applied Chemistry (https://www.degruyter.com/journal/key/pac/html). IUPAC recognized the need to extend the realm of standardized nomenclature into computer representations and in March of 2000 Ted Becker, Alan McNaught, and Stephen Heller organized a meeting at the US National Academy of Sciences to explore this (Boucher et al., 2018). They subsequently teamed up with the US National Institute of Standards and Technologies (NIST), which was already working on the problem, and in January of 2001 the IUPAC-Chemical Identifier (IChI) project was initiated (https://iupac.org/project/2000-025-1-800/). In 2004, IChI was renamed to InChI (which stands for International Chemical Identifier), and in April 2005, InChI Version 1 was released. In 2009, the not-for-profit InChI Trust (https://www.inchi-trust.org/) was formed, which works with the IUPAC Division VIII InChI Subcommittee (https://iupac.org/body/802/) to develop, test, and advance the functionality of InChI in an effort to align with the needs of the evolving sciences.

Various aspects of InChI, including its specification, are well documented elsewhere (Heller et al., 2013, 2015). There were essentially two goals with the initial release of InChI. First, it was to enable communication across databases and software agents. The objective was not to replace the identifiers a database used, but to allow that identifier to be converted to an InChI, and thus if two databases used different identifiers for the same chemical, the data could be aligned as they both convert to the same InChI. The second is to come up with an open-source

standard canonical representation where only one InChI would be generated for a specific molecule, and that anyone could use. This is not to say that you cannot have a canonical SMILES, but they are only canonical if the same algorithm generated them. Thus, by having a unique canonical identifier that the different (often proprietary) identifiers of specific databases and software agents can interconvert between, the InChI enables data exploration in the chemical sciences. As Lavoisier so aptly stated above, *"neither can we, on the other hand, improve a science without improving the language or nomenclature which belongs to it"* and InChI can improve the practice of the chemical sciences in the age of big data.

The present article provides an overview of the InChI Open Education Resource (OER) (https://www.inchi-trust.org/oer/), which aims to assist students, educators, and researchers in learning about and benefiting from the use of InChI. While definitions of the term OER may vary depending on the contexts, it refers in this paper to teaching, training, learning, or research materials that are in the public domain or that have been released under an open license that permits their free use by any person. The OER includes a wide range of materials, such as journal articles, blog posts, presentation files, videos, spreadsheets, and computer scripts. Rising costs of textbooks, coupled with increased college tuition in the past decade, have created obstacles for students to access and succeed in higher education and the OER has emerged as a promising solution to this problem. Various forms of OERs exist. For example, there are open textbook sites like LibreTexts and OpenStax, which provide free online textbooks in a wide range of subjects for college and high school. Resources like the Massachusetts Institute of Technology (MIT) OpenCourseWare provide the public with lecture notes, assignments, exams, and other materials used in the courses offered by the school. Khan Academy is also a popular OER platform, which offers free online courses for a wide range of students from the pre-kindergarten through college level. Massive Open Online Courses (MOOC) are another important example of OERs.

While the aforementioned OERs cover a wide range of subjects from humanity to science, some OERs are discipline- or topic-specific. Many of them are community-based and take advantage of Web 2.0 Content Management Systems (CMS) to advance the teaching and learning of the specific area through the sharing of relevant instructional resources. Some examples include the Analytical Sciences Digital Library (ASDL) (Chrzastowski & Scheeline, 2006; Larive, 2009), the Interactive Online Network of Inorganic Chemists' Virtual Inorganic Pedagogical Electronic Resources (IONiC/VIPEr) (Jamieson et al., 2011), and Organic Education Resources (OrganicER) (Leontyev et al., 2020).

The InChI OER (https://www.inchi-trust.org/oer/) is similar to these topic-specific, community-based OERs. InChI OER contains open-access content useful for chemical education at the upper-undergraduate and graduate levels, although it also has materials targeting other audiences, including scientists interested in data-intensive chemical research. In an effort to develop a comprehensive resource, the decision was also made to extend the OER database to non-OER content. The non-OER content is not displayed by default and the user must click the non-OER button on the OER site to include it. This paper was originally presented at an online conference hosted by the Committee on Computers in Chemical Education (CCCE) of the American Chemical Society (ACS) Division of Chemical Education (CHED) (Belford et al., 2020). The present paper is modified from its original version to reflect discussions with chemical educators during the conference as well as technical updates since made.

2 What is InChI?

InChI is a structure-based textual chemical identifier (line notation) designed to be unique and to encode substance identity at different levels of granularity. Anyone who knows the formula of a chemical entity will be able to use the free software provided by the InChI Trust to obtain its InChI (see Figure 1). In addition, the InChI API library is also freely provided. There are also multiple free molecular editors like PubChem Sketcher (Ihlenfeldt et al., 2009) and MolView (Bergwerf, n.d.) that can generate InChI strings from chemical structures or names or *vice versa*, along with resolvers like OPSIN (Open Parser for Systematic IUPAC Nomenclature) (Lowe et al., 2011) and the NCI Chemical Identifier Resolver ("NCI/CADD Chemical Identifier Resolver," n.d.) Several of these can be found in the InChI OER by using the tags "Molecular Editor" or "Resolver".

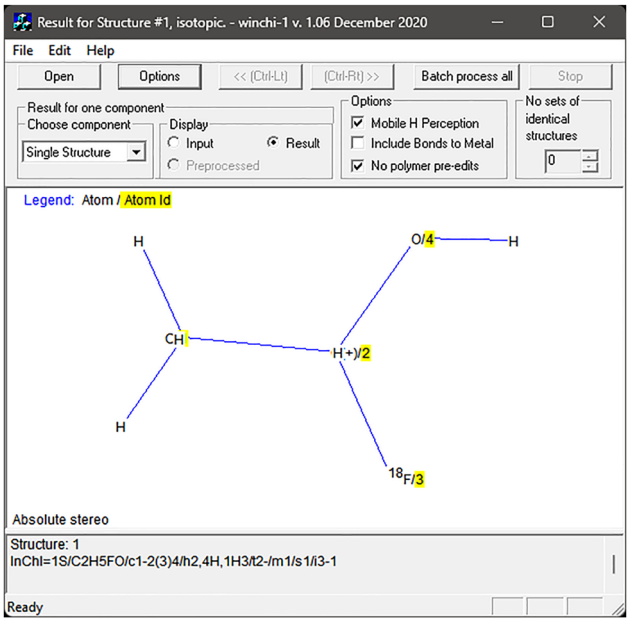


Figure 1: Windows version of the InChI software.

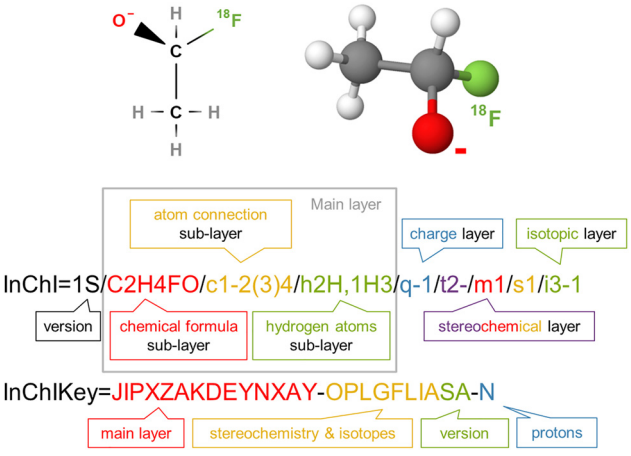


Figure 2: Layers for the standard InChI and InChIKey of (1*R*)-1-(^{18}F)fluoroethanolate. Note each layer or sublayer is separated by a forward slash.

The capability of the InChI to represent a chemical to different levels of details is based on its layered structure. The standard InChI has multiple layers (and several sublayers within the core layers). It starts with "InChI = 1S/", which indicates the version of the InChI being used, in this case, version 1. "S" stands for standard. After the prefix, the layers and sublayers are encoded, separated with forward slashes, "/". An example is shown in Figure 2. It corresponds to (1*R*)-1-(^{18}F)fluoroethanolate, the anionic form of an isotopically marked derivative of fluoroethanol.

The four layers in the standard InChI are a **main layer**, which includes the chemical formula, the connectivity, and the position of the hydrogens; a **charge layer**, which encodes changes in charge or protonation; a **stereochemical layer**, where double bond (Z/E) stereochemistry and tetrahedral stereochemistry are reflected; and an **isotopic layer**, to indicate isotopical changes. Additional layers, like a fixed-hydrogens layer, to locate mobile hydrogens if required, a reconnected layer, used when metals are present, or a polymer layer, devoted to encoding polymeric entities, are also considered in the current version of the InChI documentation (https://www.inchi-trust.org/download-latest-inchi-standard-software/).

Figures 3 and 4 show a set of InChI for related compounds to illustrate these layers and show how they help reflecting on chemical identity. Figure 3 includes neutral compounds which are different forms of 1-fluoroethanol.

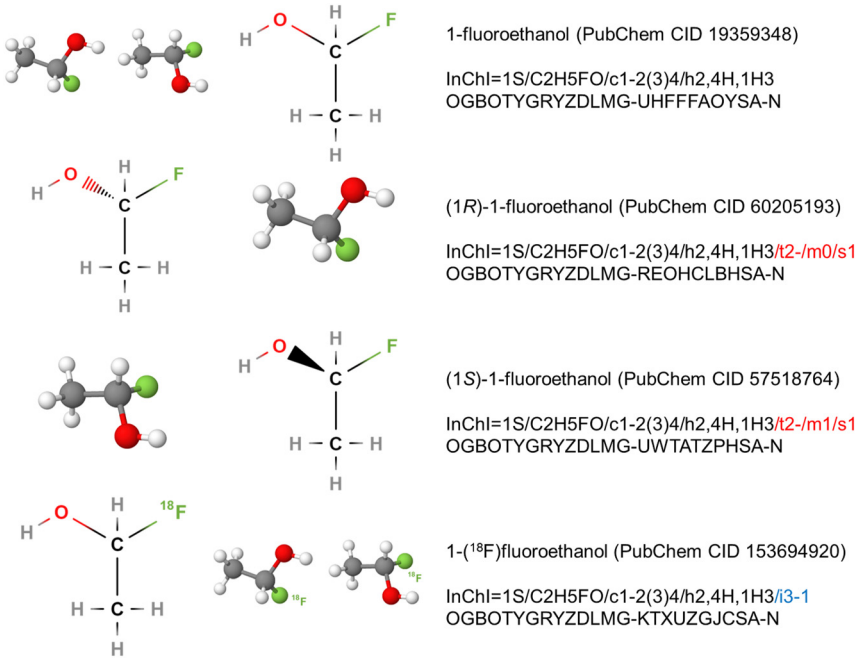


Figure 3: InChIs and InChIKeys of neutral compounds related to 1-fluoroethanol. To highlight the changes in the InChIs, the stereochemical layer is shown in red and the isotopic layer is in blue.

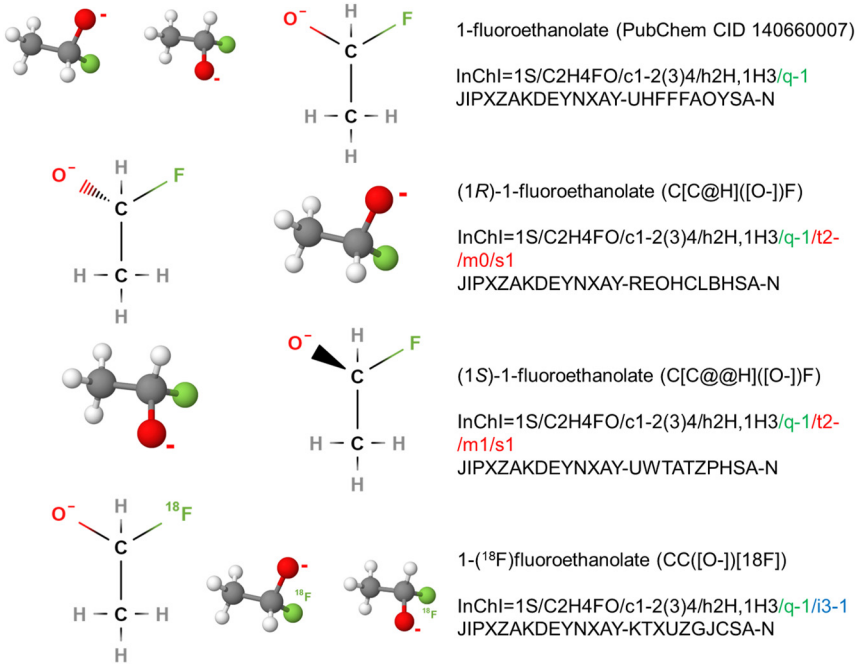


Figure 4: InChI and InChIKeys of charged species related to 1-fluoroethanol. To highlight the changes in the InChIs, the charge layer is in green, the stereochemical layer is shown in red, and the isotopic layer is blue.

Stereochemically unspecified compounds and their stereoisomers have the same main layers. The stereochemical layer allows one to distinguish between them. Isotopologues can be identified by the isotopic layer. Charged species may have different main layers than the neutral counterparts as shown in Figure 4.

The layers of the InChI notation can help discuss what identity means in chemical compounds and clarify what we refer to every time we use chemical identifiers.

The InChILayersExplorer (Figure 5) is an Excel spreadsheet that allows one to explore and exemplify the layers in the InChI. This tool is available at the InChI OER via the URL: https://www.inchi-trust.org/inchi-post/inchilayersexplorer-an-spreadsheet-to-tech-and-learn-the-structure-of-an-inchi/.

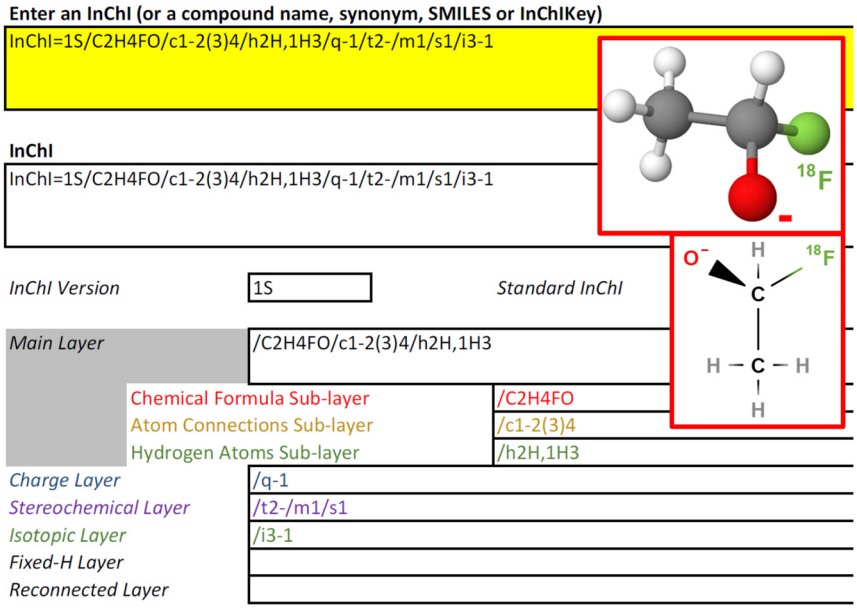


Figure 5: Screenshot of InChILayersExplorer that can be obtained in the InChI OER. Red inserts, which are not part of the spreadsheet application, show the 2D and 3D depictions of the chemical represented by the analyzed InChI, (1*R*)-1-(^{18}F)fluoroethanolate. The InChILayersExplorer can be accessed via the URL: https://www.inchi-trust.org/inchi-post/inchilayersexplorer-an-spreadsheet-to-tech-and-learn-the-structure-of-an-inchi/.

When the user provides an InChI string as an input, the InChILayerExplorer analyzes and splits it into individual layers, as shown in Figure 5 (with (1*R*)-1-(^{18}F)fluoroethanolate as an example). This tool also supports other types of chemical identifiers (like a chemical name, SMILES, or InChIKey), which are converted to the corresponding InChI string.

2.1 What is InChIKey?

One of the disadvantages of InChI is that it can get very long as the molecular size increases. For example, the InChI string of adrenocorticotropic hormone (also known as corticotropin; PubChem Compound ID 16132265) is 2073-characters long. It is not feasible to use such long InChI strings for casual chemical information searching (e.g., using popular search engines through a web browser). This concern led to the development of InChIKey, which is a hashed representation of the InChI (see Figure 2). Hashing is a mathematical process that converts data of arbitrary size into fixed-size values. The InChIKey is a 27-character-long string generated from an InChI string of variable length by using the SHA-256 cryptographic hash function of the SHA-2 family (U.S. National Institute of Standards and Technology, 2002). Because the InChIKey has a consistent length, regardless of the molecular size, and is almost always shorter than InChI except for very simple molecules, making it a more suitable identifier for searching the internet and indexing databases. For this reason, popular search engines including Google and Bing support web search by InChIKey.

2.2 InChI for chemical data sharing and reuse

As more scientific data generated from government-funded research became increasingly available to the public, scientific communities and funding agencies recognized a need for a digital data ecosystem that promotes the sharing and reuse of scholarly data to maximize research investments. This led to the development of a set of

guidelines for good data management and stewardship, called the FAIR data principles, where the term "FAIR" is the acronym for Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016). According to the FAIR principles, scientific data must be:

- **Findable**: both humans and computers should be able to find the data.
- **Accessible:** both data and the metadata should be readily accessible using standard methods.
- **Interoperable**: the data must be readily usable with other applications.
- **Reusable:** the data and metadata must be clearly defined so that they can be easily reused in other experiments.

Most methods of chemical nomenclature do not comply with these requirements as well as the InChI nomenclature does, because of its interoperability and machine readability. The InChI format and algorithm are non-proprietary and the InChI software is freely provided to everyone. In addition, because InChI makes it easier to programmatically link chemical structures in one online resource with the same structures in another, it has been adopted in many large chemical databases and computer programs that deal with millions of chemical structures.

2.3 Adoption of InChI in chemical education

There have been ongoing efforts to teach InChI in chemistry courses (Kadtsyna et al., 2022; Kim et al., 2021). These courses cover various use cases of InChI. For instance, students learn how to use chemical-structure drawing programs (e.g., ChemDraw, MarvinSketch) to draw chemical structures and convert them into their InChI strings, which, in turn, can be used as a query to search public online databases (e.g., PubChem and ChemSpider) for the desired molecular properties. Another example is the conversion between InChIs and other chemical identifiers (such as chemical names, CAS registry numbers, and SMILES strings) using free web-based tools, including the National Cancer Institute Chemical Identifier Resolver (https://cactus.nci.nih.gov/chemical/structure) (Peach & Nicklaus, 2018). Activities involving the Open Parser for Systematic IUPAC Nomenclature (OPSIN; https://opsin.ch.cam.ac.uk/) (Lowe et al., 2011) are also covered in introductory courses. Writing a simple python or R script that uses InChIs to programmatically get chemical information from public chemical databases is an important topic covered in cheminformatics courses for upper-undergraduate and graduate students.

In addition, InChI presents new opportunities for instructional technology in chemical education. Recently, a web-based classroom response system (WBCRS) was developed, which enables automated assessment of and feedback on chemical structures drawn by students as answers to organic chemistry questions (Otálvaro, 2022). Another example is a work by the Bucholtz group, in which open-source optical structure recognition software is used to convert hand-drawn Lewis dot structures of organic molecules into their InChI strings with an aim to develop an automated grading system for hand-drawn chemical structures (Belford & Gupta, 2019; Bucholtz & Stephens, 2018).

2.4 Prospect of InChI in chemical education

P. J. Karol stated in his paper published in the Journal of Chemical Education in 2018 that "InChI is used by professional chemical database providers, publishers, chemistry software vendors, librarians, patent attorneys, and information specialists." and that "it is not and should not be the subject of any general chemistry exam (Karol, 2018)." While Karol's comments are still deemed valid in 2023, there have been some notable changes. The increased availability of scientific data in the public domain has led to large-scale investments in data science infrastructure from both public and private sectors, creating a huge demand for a new workforce equipped with skill sets for data-intensive projects. In recognition of this demand, many colleges and universities introduced data science degree programs at both undergraduate and graduate levels, typically involving multiple academic units in various areas, such as computer science, mathematics, statistics, biology, medicine, pharmacy, finance, and business.

The chemical education community is also making various attempts to provide training opportunities in this area. An example is the Cheminformatics OLCC, which was mentioned in this paper. Importantly, in 2023, the American Chemical Society (ACS) recently approved its guidelines for accreditation of Bachelor's degree programs, which gives a strong emphasis in data and analysis skills (https://www.acs.org/education/policies/acs-approval-program/guidelines.html) (ACS, 2023). According to the new ACS guidelines, accredited schools should introduce students to modern laboratory-record-keeping tools including laboratory information management systems (LIMS) and electronic laboratory notebooks (ELNs). In addition, the guidelines also suggest three markers of excellence in the data and analysis skills area:

- understanding of data compliance and integrity issues within a regulatory context.
- appropriate documentation, data analysis, and data management skills.
- programming skills.

Considering all these recent changes, data science training and education opportunities for chemistry majors are expected to become more popular. Inevitably, students will more likely be introduced to InChI, because it is a molecular representation standard used by major information resources. Of course, because InChI is a very specialized topic, it will still not be appropriate to teach in lower undergraduate classes, as pointed out by Karol. Rather, most courses covering InChI will be elective or special topics courses for upper-undergraduate and graduate students.

3 What is the InChI OER and why is it valuable?

There are several advantages to using OERs to add a discussion of InChIs to a traditional course. If the course textbook does not include a discussion of InChIs, the instructor can include as much information as he or she wishes by adding (and editing as needed) one or more OERs. If the topic is covered in the textbook, the OERs can allow the instructor to include alternative approaches to the topic. This does not increase the cost to the students, and students can access this information even after they have completed the course. Unlike a hardbound textbook, OERs are easy to expand, correct, or modify.

As mentioned in the Introduction, the InChI OER (https://www.inchi-trust.org/oer/) is a community-based OER, and is built on the WordPress Web 2.0 CMS. It is integrated into the InChI-Trust website (https://www.inchi-trust.org) and can be accessed through the "RESOURCES" tab on the Trust's website (Figure 6). The InChI OER is being developed by an IUPAC Task Group (https://iupac.org/project/2018-012-3-024/) involving support through the InChI Trust, the IUPAC Committee on Publications and Cheminformatics Data Standards (https://iupac.org/body/024/), the IUPAC Committee on Chemistry Education (https://iupac.org/body/050/) and IUPAC Division VIII, Chemical Nomenclature and Structure Representation (https://iupac.org/body/800/). The InChI OER task group consists predominantly of educators who are working with the Division VIII InChI Subcommittee and the InChI Trust in an effort to bring about a greater awareness of InChI and its impact on the practice of science.

Like other Web 2.0 education instructional resource sites, the InChI OER allows people to share and find resources. When content is uploaded to the OER it is tagged, and then those tags are exposed to the public as a filter, which can then be used to find material within the site. In the InChI OER, there are two types of tag filters (see Figure 7 and https://youtu.be/d5qsfUDnyZE). The first filters between OER and non-OER content, as there are

Figure 6: The "resources" tab on the header of the InChI trust webpage has a link to the InChI OER.

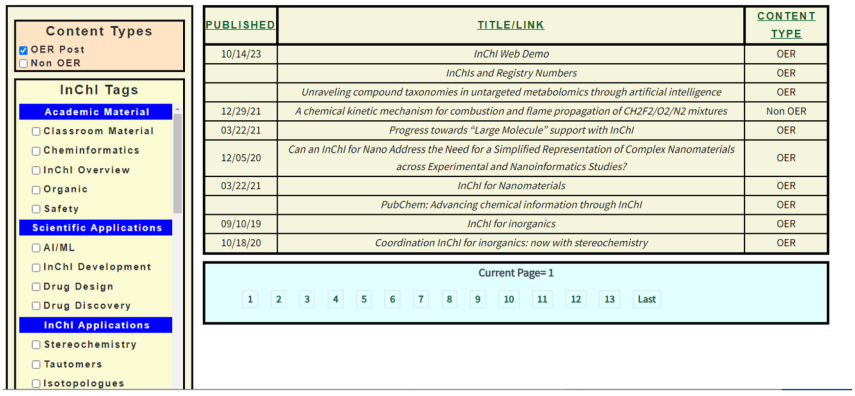


Figure 7: The InChI OER tag filter when no tags are selected.

valuable non-OER publications related to InChI, which can be found by simply clicking the non-OER radio button (clicking both gives you access to all content). The second filters content through the hierarchical InChI OER Taxonomy.

3.1 InChI OER taxonomy

The InChI OER Taxonomy has 5 broad categories of tags and the web interface allows you to perform Boolean "AND" searches of the content uploaded to the OER. Table 1 defines the current tags and a brief description of their purpose, and as this is a dynamic web site we expect for the taxonomy and content to evolve over time.

When you click an item in the taxonomy filter, you get a page with information on that item. It includes a direct link to the original source if the content is non-OER or published in an open-access journal. Otherwise, the content is directly uploaded to the InChI Trust website and available to the public through the InChI OER (all content directly uploaded to the site is OER).

When the InChI OER loads, you see a tag filter on the left and a display of hits on the right (Figure 7), and you need to click the non-OER if you want to see all content cataloged by the site. By simply clicking the checkboxes of the web interface you can perform a Boolean "AND" filter of multiple tags.

Figure 7 shows the tag filter before any tag has been chosen. Once you choose a tag, the filter adjusts to only show papers with that tag, and the tags that are on those papers. Figure 8 shows there are currently nine items that have the tag "Classroom Material", and only papers that are tagged classroom material show in the filter.

When you further refine the filter by combining the tag "Classroom Material" with the tag "Organic" the filter is reduced to three items that have been uploaded to the OER (Figure 9).

Choosing the content from Figure 8 titled "InChILayersExplorer", the OER will bring up a page dedicated to that specific resource (the spreadsheet in Figure 5). Depending on the resource, different items may be shown on

Table 1: InChI OER taxonomy used to categorize InChI OER items.

Category	Description
Academic material	Information of relevance to different disciplines. The first tag is classroom material, which when combined with a discipline can assist in finding material related to the discipline
Scientific applications	Part of the goal of InChI is to advance digital nomenclature and thus support the advancement of scientific discovery. Here we are identifying resources related to fourth paradigm science.
InChI applications	Many of these are actual topics of the IUPAC InChI subcommittee working groups and represent ways InChI is advancing.
Content type	This allows you to filter by content type, like spreadsheet or video.
Miscellaneous	Useful tags that do not really fit the above categories.

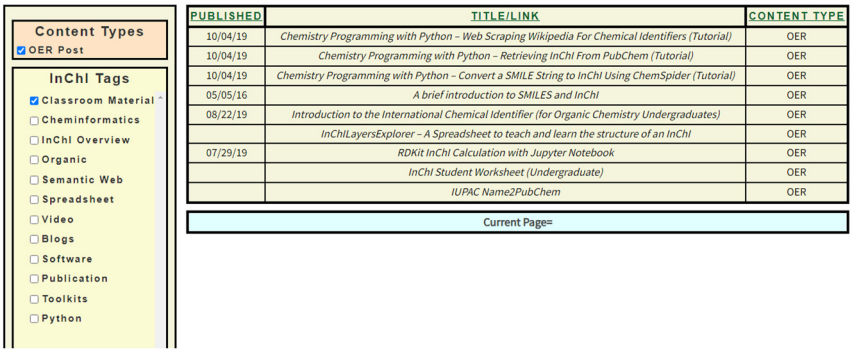


Figure 8: Tag taxonomy filtered to the tag “Classroom Material”.

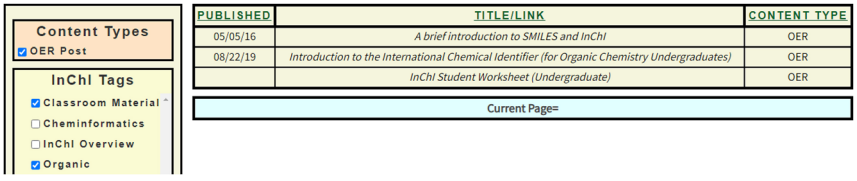


Figure 9: Tag taxonomy filtered to tags “Classroom Material” and “Organic”.

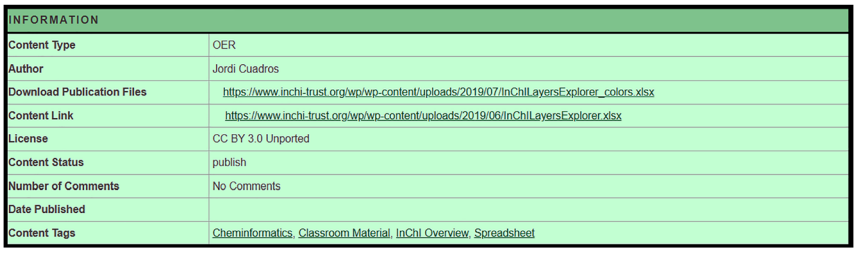


Figure 10: The information box associated with the InChILayersExplorer. Note you can download the spreadsheet in Figure 5 from the third field in the box.

the page, but all will have an information box (e.g., Figure 10 for the InChILayersExplorer). Some will contain links to download files, while others may link to the original host that published the content. The last piece of information, located at the bottom, includes the tags with which that material has been tagged. Clicking a tag will pull up all other resources tagged with the same term.

For the long-term sustainability of the InChI OER project, the Zenodo InChI OER community page was created to share important materials for the development and maintenance of InChI OER (https://zenodo.org/communities/inchi-oer/). These materials include the source code used to create the InChI OER website (Cornell, 2023a), the InChI OER manuals for developers, content contributors, and users (Cornell, 2023b), and the InChI OER taxonomy structure diagram (Cornell, 2023c).

3.2 Exemplar of the InChI OER: application to safety and the stockroom

In this section, we will look at how one can go to the InChI OER and learn how to create a Google Sheet that will take your chemical stockroom inventory and connect it to PubChem Laboratory Chemical Safety Summaries (LCSS) (https://pubchem.ncbi.nlm.nih.gov/lcss) (Kim et al., 2015). PubChem (https://pubchem.ncbi.nlm.nih.gov) (Kim et al., 2023) is a public compound database and one of the largest “big data” repositories in the chemical sciences, which has information on 115 million chemical compounds from 941 data sources as of October 2023 (https://pubchem.ncbi.nlm.nih.gov/docs/statistics). PubChem aggregates data from multiple sources in a chemical compound’s summary page and maintains the provenance of that data by linking to the original source. The

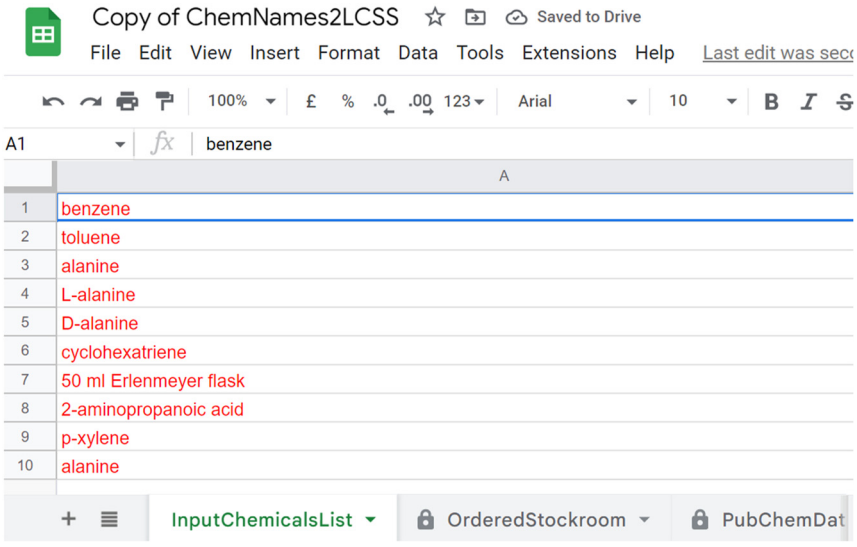


Figure 11: Screen capture of the Google Sheet "ChemNames2LCSS" showing three tabs.

PubChem LCSS is a subset of the information contained in a chemical's compound summary page focused on hazard and safety information, as exemplified by the LCSS for Benzene (PubChem Compound ID 241):

https://pubchem.ncbi.nlm.nih.gov/compound/241#datasheet=LCSS

The PubChem LCSS is modeled after the format described in *Prudent Practices in the Laboratory: Handling and Management of Chemical Hazards* by the National Academy of Science's National Research Council (NRC, 2011).

Most chemists query the web for information through the search interface of their web browsers, but you can also navigate the web through calls in a spreadsheet to an Application Program Interface (API). If one goes to the InChI OER and selects the tags "Spreadsheet" and "Safety", they filter the content for Google sheets related to safety, where they can find and make a copy of "ChemNames2LCSS" sheet. This spreadsheet allows users to paste a list of up to 1000 chemicals (think stockroom inventory) and instantly generate links to the LCSS of the chemicals.

Initially, when you load a list of words they are in red, which means there is no LCSS connected to that word (Figure 11).

As the Google Sheet processes the list, some of the words turn black, meaning that the word is a unique chemical and there is now an LCSS connected to that chemical (Figure 12). If the word becomes shaded it means it is redundant to a name already linked and if it stays red, it is not identified as a chemical.

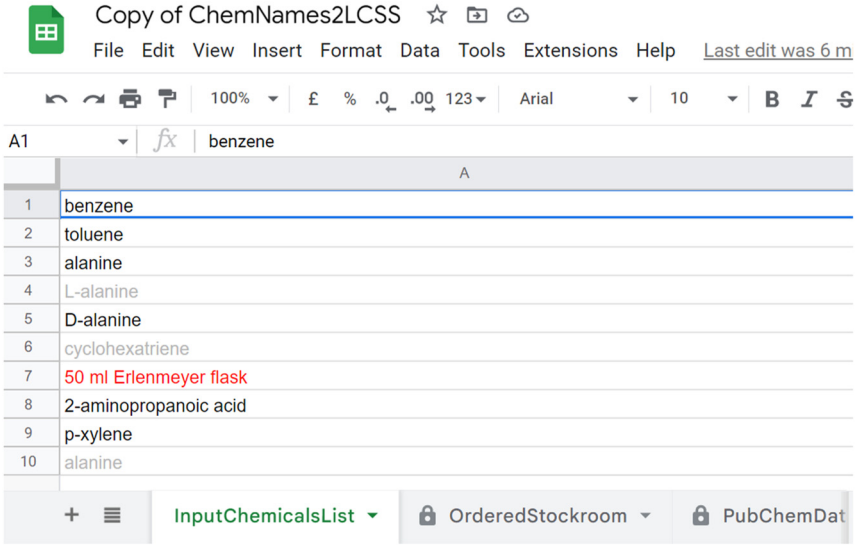
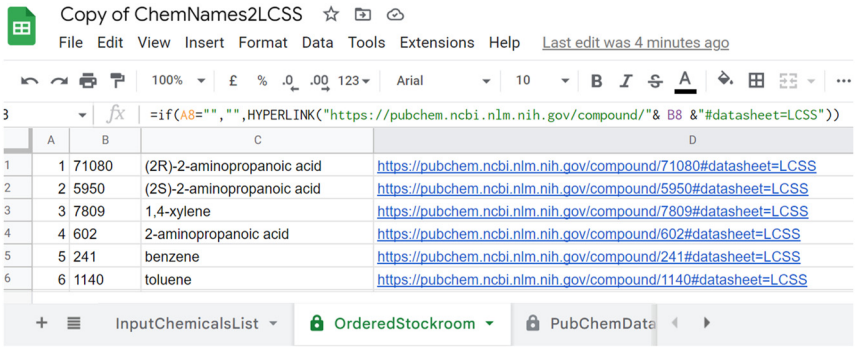
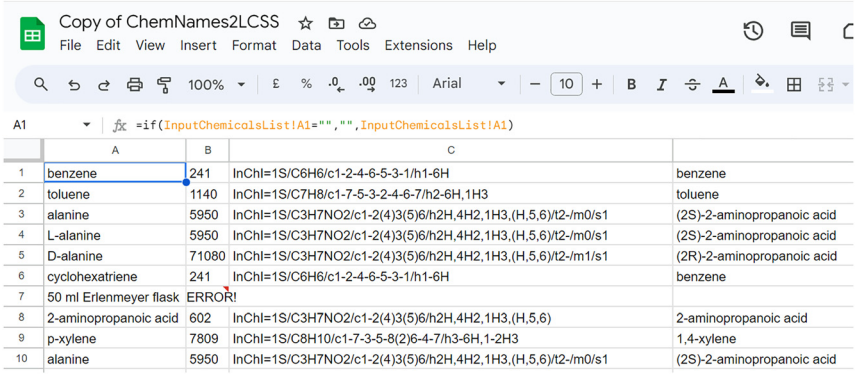


Figure 12: The InputChemicalsList after processing. Of the 10 entries, only 6 are unique chemicals.



	A	B	C	D
1	1	71080	(2R)-2-aminopropanoic acid	https://pubchem.ncbi.nlm.nih.gov/compound/71080#datasheet=LCSS
2	2	5950	(2S)-2-aminopropanoic acid	https://pubchem.ncbi.nlm.nih.gov/compound/5950#datasheet=LCSS
3	3	7809	1,4-xylene	https://pubchem.ncbi.nlm.nih.gov/compound/7809#datasheet=LCSS
4	4	602	2-aminopropanoic acid	https://pubchem.ncbi.nlm.nih.gov/compound/602#datasheet=LCSS
5	5	241	benzene	https://pubchem.ncbi.nlm.nih.gov/compound/241#datasheet=LCSS
6	6	1140	toluene	https://pubchem.ncbi.nlm.nih.gov/compound/1140#datasheet=LCSS

Figure 13: "Ordered Stockroom" tab provides an alphabetized list of the identified 6 unique chemicals with links to LCSS.



	A	B	C	
1	benzene	241	InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H	benzene
2	toluene	1140	InChI=1S/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3	toluene
3	alanine	5950	InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m0/s1	(2S)-2-aminopropanoic acid
4	L-alanine	5950	InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m0/s1	(2S)-2-aminopropanoic acid
5	D-alanine	71080	InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m1/s1	(2R)-2-aminopropanoic acid
6	cyclohexatriene	241	InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H	benzene
7	50 ml Erlenmeyer flask	ERROR!		
8	2-aminopropanoic acid	602	InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)	2-aminopropanoic acid
9	p-xylene	7809	InChI=1S/C8H10/c1-7-3-5-8(2)6-4-7/h3-6H,1-2H3	1,4-xylene
10	alanine	5950	InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m0/s1	(2S)-2-aminopropanoic acid

Figure 14: The PubChemDataCollection tab was used to process the data. Note that there are many more columns to this tab.

If you click the "OrderedStockroom" tab you get a unique list of chemicals in alphabetical order with links to the LCSS (Figure 13).

The question becomes, what was the role of the InChI in this endeavor? One of the first steps in processing data is to clean it up, and for example, a stockroom inventory will have multiple listings of the same chemical, along with content that may not be a chemical. If you look at the PubChemDataCollection tab (Figure 14), you see a list of InChIs (Column C), which would identify each unique chemical. The script used the InChI to identify the unique chemicals, which were then sorted in alphabetical order and listed in the middle tab (Figure 13) with a link to the PubChem LCSS. Note how Row 7 (Erlenmeyer flask) produced an error and could not be processed, and so this item stayed red (Figure 12).

The use of InChI thus allowed us to identify redundant entries and synonyms, but as we will see in the next section on the future of InChI, additional layers like the description of mixtures (MInChI) are also of critical importance, as oftentimes the phase and concentration of a chemical influence its properties and safety concerns. That is, the role of digital representation of chemicals is an ongoing endeavor of IUPAC and the InChI Trust, and the layered nature of InChI allows continual development of its functions and features.

4 Future of InChI and the InChI OER

The extensible nature of the layered InChI notation allows the development of new features and the IUPAC InChI subcommittee in collaboration with the InChI Trust has multiple working groups devoted to developing features that can enable InChI to better serve the needs of science in the digital age (Goodman et al., 2022). An overview of the working groups can be found on the InChI Trust website (https://www.inchi-trust.org/inchi-working-groups/). If you are interested in contributing to any working group you can contact the chair of that working group, which is posted on the InChI Trust website.

The following lists some of the working groups that would be of interest to educators and some of the resources available within the OER.

- ***Extended Stereochemistry*** (https://www.inchi-trust.org/extended-stereochemistry/): This group is seeking to improve stereochemical handling. While InChI can describe stereochemistry well, continual improvement is being made in various areas such as atropisomers and polyhedral stereochemistry for organometallics and coordination compounds. Other representations such as Haworth and Fischer present challenges as InChI are generated from molfiles.
- ***Extended Tautomers*** (https://www.inchi-trust.org/extended-tautomers/): InChI has a set of rules to predict and resolve tautomers. This group has completed further development on rules to identify prototropic, valence, and ring-chain tautomers (Dhaked et al., 2020) and is discussing how to incorporate these into version 2.
- ***Isotopologues*** (https://www.inchi-trust.org/isotopologues/): The focus of this group is to enhance the isotopic layer to handle specific isotopologues, isotopomers, and partial isotopomers.
- ***Large Molecules*** (https://www.inchi-trust.org/large-molecules/): This group is searching for an approach to allow for biologics using monomer-based, pseudo-atoms for non-standard InChI. By adding monomer support, biopolymer InChI can be reduced in complexity and length, making it possible to handle even larger biological molecules. This would result in increases in speed for computation of large molecule InChI.
- ***Markush Structures*** (https://www.inchi-trust.org/inchi-markush/): This project addresses the description of variable structures by providing pseudo elements to mark a substitution point on a molecule corresponding to traditional Markush structures. VInChI is a variable InChI where multiple InChI are compacted into a single string describing how other molecules can be generated.
- ***Mixtures*** (https://www.inchi-trust.org/mixtures/): The Mixtures InChI (MInChI) project (Clark et al., 2019) defines a method to describe the chemical composition of a given mixed substance from an unambiguous machine-readable linear notation that can be resolved into a list of unique components.
- ***Nanomaterials*** (https://www.inchi-trust.org/nanomaterials/): The NInChI project (Lynch et al., 2020) aims to encode the composition, size, shape, and surface chemistry of nanomaterials/nanoforms. Since nanomaterials are particulates, these challenge traditional nomenclature limiting exchange of ideas between scientists in academia, industry, and regulatory agencies.
- ***Organometallics*** (https://www.inchi-trust.org/organometallics/): This group is investigating how disconnection and connection code affects organometallic compounds (Clark, 2020).
- ***Reactions*** (https://www.inchi-trust.org/reactions/): The Reaction InChI (RInChI) (Grethe et al., 2018) allows for a concise description of chemical reactions and facilitates the manipulation and analysis of reaction data.
- ***InChI QR Codes*** (https://www.inchi-trust.org/inchi-qr-codes/): This group is developing a QR code (2D bar code) version of the InChI, called QRInChI, to provide a quick way to access chemical information (Frey et al., 2022).

The InChI OER is maintained by the InChI Education working group and our hope is that it will grow as time evolves and more resources relating InChI to education are generated. Lavoisier's correlation of the interdependence of the growth of science with its nomenclature is as true today as it was in 1790, and our goal with the InChI OER is to help connect this new digital nomenclature to practicing scientists by making an open resource where practitioners can find and share resources. Unlike most OERs devoted to providing curriculum instructional material on well established topics, the InChI OER is trying to provide a platform for this new evolving nomenclature as it is being developed. If you wish to contribute content to the InChI OER, or participate in the InChI Education working group, you should contact Professor Belford (rebelford@ualr.edu).

5 Conclusions

InChI is a chemical identifier that encodes chemical structure information into a hierarchically layered line notation. Because InChI is open, non-proprietary, and freely available to everyone, it has been adopted in many chemical information resources. In addition, because of its machine readability and interoperability, InChI is now

commonly used to link identical chemical structures between different information resources that contain millions of chemical structures.

The InChI OER aims to provide the public with InChI-related training and education materials. Currently, it contains over 120 items, collected from various resources, and the majority of them are freely available to the public. The InChI OER provides an interface that uses a tag taxonomy to filter content and allow users to find desired materials, as described in this paper. In addition, we demonstrated the use of InChI in a spreadsheet program to automatically retrieve chemical safety information from a public database.

Acknowledgments: We would like to thank the following for their help and support in our efforts: Steve Heller, Alan McNaught, Richard Kidd, Jonathan Goodman, Leah McEwen, Ralph Stuart, Nathan Brown, Tina Qin, Vincent Scalfani, Martin Walker, and Steven Wathen.
Research ethics: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: This work was supported by IUPAC (project number: 2018-012-3-024) and an InChI Trust Fellowship. The work of SK was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.
Data availability: Not applicable.

References

ACS Guidelines for Bachelor's Degree Programs. (2023). Retrieved November 2, 2023, from American Chemical Society website: https://www.acs.org/education/policies/acs-approval-program/guidelines.html

Belford, R., Cuadros, J., Cornell, A., Gupta, T., & Bucholtz, E. (2020). InChI OER. Committee on Computers in Chemical Education (CCCE) Newsletter. https://confchem.ccce.divched.org/2020CCCENLP1

Belford, R. E., & Gupta, T. (2019). Introduction: Technology integration in chemistry education and research. In *ACS symposium series: Vol. 1318. Technology integration in chemistry education and research (TICER)* (Vol. *1318*, pp. 1–20). American Chemical Society.

Bergwerf, H. (n.d.). MolView. Retrieved November 2, 2023, from MolView website: https://molview.org/

Boucher, R., Heller, S., Kidd, R., & Pletnev, I. (2018). What on Earth is InChI? Retrieved November 26, 2020, from IUPAC 100 Stories website: https://iupac.org/100/stories/what-on-earth-is-inchi/

Bucholtz, E. C., & Stephens, C. (2018). *Optical Structure recognition (OSR) as a means to grade student work in organic chemistry*. In *Presented at the 25th biennial conference on chemical education, South Bend, IN*, South Bend, IN.

Chrzastowski, T. E., & Scheeline, A. (2006). ASDL: The analytical sciences digital library taking the next steps. *Science & Technology Libraries*, *26*(3–4), 79–94.

Clark, A. M. (2020). Coordination InChI for inorganics: Now with stereochemistry. Retrieved January 19, 2023, from https://cheminf20.org/2020/10/18/coordination-inchi-for-inorganics-now-with-stereochemistry/

Clark, A. M., McEwen, L. R., Gedeck, P., & Bunin, B. A. (2019). Capturing mixture composition: An open machine-readable format for representing mixed substances. *Journal of Cheminformatics*, *11*(1), 33.

Cornell, A. (2023a). Boots7458/InChI-OER. *Zenodo*. https://doi.org/10.5281/zenodo.7679132

Cornell, A. (2023b). InChI OER manual. *Zenodo*. https://doi.org/10.5281/zenodo.7679158

Cornell, A. (2023c). *InChI OER taxonomy structure*. *Zenodo*. https://doi.org/10.5281/zenodo.7675975

Dhaked, D. K., Ihlenfeldt, W.-D., Patel, H., Delannée, V., & Nicklaus, M. C. (2020). Toward a comprehensive treatment of tautomerism in chemoinformatics including in InChI V2. *Journal of Chemical Information and Modeling*, *60*(3), 1253–1275.

Frey, J. G., Hartshorn, R. M., & McEwen, L. R. (2022). Specification of International Chemical Identifier (InChI) QR codes for linking labels on containers of chemical samples to digital resources (IUPAC Recommendations 2021). *Pure and Applied Chemistry*, *94*(10), 1195–1206.

Goodman, J., Blanke, G., Potenzone, R., & Heller, S. (2022). InChI open meeting. *Chemistry International*, *44*(3), 42–44.

Grethe, G., Blanke, G., Kraut, H., & Goodman, J. M. (2018). International chemical identifier for reactions (RInChI). *Journal of Cheminformatics*, *10*(1), 22.

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics*, *7*, 23.

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., & Pletnev, I. (2013). InChI – the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, *5*, 7.

Hey, T., Tansley, S., Tolle, K., & Gray, J. (2009). The fourth paradigm: Data-intensive scientific discovery. *Microsoft Research*. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

Ihlenfeldt, W. D., Bolton, E. E., & Bryant, S. H. (2009). The PubChem chemical structure sketcher. *Journal of Cheminformatics*, *1*(1), 20.

Jamieson, E. R., Eppley, H. J., Geselbracht, M. J., Johnson, A. R., Reisner, B. A., Smith, S. R., Watson, L.A., & Williams, B.S. (2011). Inorganic chemistry and IONiC: An online community bringing cutting-edge research into the classroom. *Inorganic Chemistry*, *50*(13), 5849–5854.

Kadtsyna, A. S., Chubarov, A. S., & Kadtsyn, E. D. (2022). Basic cheminformatics course for first-year chemistry students. *Journal of Chemical Education*, *99*(8), 2932–2942.

Karol, P. J. (2018). The InChI code. *Journal of Chemical Education*, *95*(6), 911–912.

Kim, S., Bucholtz, E. C., Briney, K., Cornell, A. P., Cuadros, J., Fulfer, K. D., Gupta, T., Hepler-Smith, E., Johnston, D. H., Lang, A. S. I. D., Larsen, D., Li, Y., McEwen, L. R., Morsch, L. A., Muzyka, J. L., & Belford, R. E. (2021). Teaching cheminformatics through a collaborative intercollegiate online chemistry course (OLCC). *Journal of Chemical Education*, *98*(2), 416–425.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolten, E. E. (2023). PubChem 2023 update. *Nucleic Acids Research*, *51*(D1), D1373–D1380.

Kim, S., McEwen, L., Stuart, R., Thiessen, P., Gindulyte, A., Zhang, J., Bolton, E. E., & Bryant, S. H. (2015). PubChem laboratory chemical safety summary. *ConfChem 2015 Fall Newsletter*. https://confchem.ccce.divched.org/2015FallCCCENLP3.

Larive, C. K. (2009). The analytical sciences digital library (ASDL). *Analytical and Bioanalytical Chemistry*, *395*(8), 2425.

Lavoisier, A. (2009). Elements of chemistry, in a new systematic order, containing all the modern discoveries. Project Gutenberg. https://www.gutenberg.org/files/30775/30775-h/30775-h.htm

Leontyev, A., Houseknecht, J. B., Maloney, V., Muzyka, J. L., Rossi, R., Welder, C. O., & Winfield, L. (2020). OrganicERs: Building a community of practice for organic chemistry instructors through workshops and web-based resources. *Journal of Chemical Education*, *97*(1), 106–111.

Lowe, D. M., Corbett, P. T., Murray-Rust, P., & Glen, R. C. (2011). Chemical name to structure: OPSIN, an open source solution. *Journal of Chemical Information and Modeling*, *51*(3), 739–753.

Lynch, I., Afantitis, A., Exner, T., Himly, M., Lobaskin, V., Doganis, P., Maier, D., Sanabria, N., Papadiamantis, A. G., Rybinska-Fryca, A., Gromelski, M., Puzyn, T., Willinghagen, E., Johnston, B. D., Gulumian, M., Matzke, M., Etxabe, A. D., Bossa, N., Serra, A., … Melagraki, G. (2020). Can an InChI for nano address the need for a simplified representation of complex nanomaterials across experimental and nanoinformatics studies? *Nanomaterials*, *10*(12), 2493.

National Research Council. (2011). *Prudent Practices in the Laboratory: Handling and Management of Chemical Hazards*. The National Academies Press. Updated Version.

NCI/CADD Chemical Identifier Resolver. (n.d.). Retrieved November 2, 2023, from Chemical Identifier Resolver website: https://cactus.nci.nih.gov/chemical/structure

Otálvaro, F. (2022). Merging drawing-based questions with automatic assessment in organic chemistry using smartphones. *Journal of Chemical Education*, *99*(8), 3044–3048.

Peach, M. L., & Nicklaus, M. C. (2018). Chemoinformatics at the CADD group of the national cancer Institute. In *Applied Chemoinformatics* (pp. 385–393). Weinheim, Germany: Wiley-VCH Verlag GmbH & Co.

Sayle, R. A. (2014). How do I write thee? Let me count the ways. Retrieved November 26, 2020, from https://nextmovesoftware.com/blog/2014/07/15/how-do-i-write-thee-let-me-count-the-ways/

Weininger, D. (1988). SMILES, a chemical language and information-system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, *28*(1), 31–36.

Weininger, D. (1990). SMILES. 3. Depict—graphical depiction of chemical structures. *Journal of Chemical Information and Computer Sciences*, *30*(3), 237–243.

Weininger, D., Weininger, A., & Weininger, J. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, *29*(2), 97–101.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018.