A Spanish Chemoinformatics GitBook for Chemical Data retrieval and Analysis using Python Programming

Fernanda I. Saldivar-González*, Diana L. Prado-Romero, B. Raziel Cedillo-González, Ana L. Chávez-Hernández, Juan F. Avellaneda-Tamayo, Alejandro Gómez-García, Luis Juárez-Rivera, José L. Medina-Franco*

DIFACQUIM research group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico

*Corresponding authors: FI. S-G fer.saldivarg@gmail.com; JL. M-F medinajl@unam.mx

Abstract

Searching, retrieving, and analyzing chemical information is one of the main tasks faced by students and professionals in chemistry-related scientific disciplines. Currently, freely available modules developed in programming languages, such as Python, allow efficient data management and facilitate obtaining information and knowledge from the data. This manuscript describes an electronic handbook generated on the GitBook platform to introduce the Python programming language and the Analysis, computational representation, and visualization of chemical data. This manual explores the most common molecular representations of low molecular weight organic compounds and their applications in various contexts. It also illustrates the acquisition of chemical information from large public molecular databases such as ChEMBL and PubChem and the Analysis and visualization of chemical information using concepts such as chemical space. The GitBook is freely available and is expected to foster open science and facilitate learning for chemistry students at the undergraduate and graduate levels and professionals interested in chemical data analysis and visualization.

Keywords: Chemoinformatics; Scientific Education; Latin America; Python; Spanish-speaking community; Open science; Handbook

Abbreviations: ADME, absorption, distribution, metabolism, and excretion; ECFP, extended connectivity fingerprint; EDA, exploratory data analysis; MACCS, Molecular ACCes System; SAR, structure-activity relationships; SPR, structure-property relationships; QSAR, quantitative structure-activity relationship.

Introduction

Chemoinformatics is one of the independent disciplines that has become a pillar during the development and design of new drugs and, therefore, indispensable in pharmaceutical chemistry. This area of knowledge allows solving problems in the management and presentation of information in chemistry by integrating different computational techniques and methods.[1] Chemoinformatics merges chemistry and informatics to solve tasks in chemistry. Figure 1 illustrates schematically the broad applications of Chemoinformatics in drug discovery and many other chemistry areas.[2]

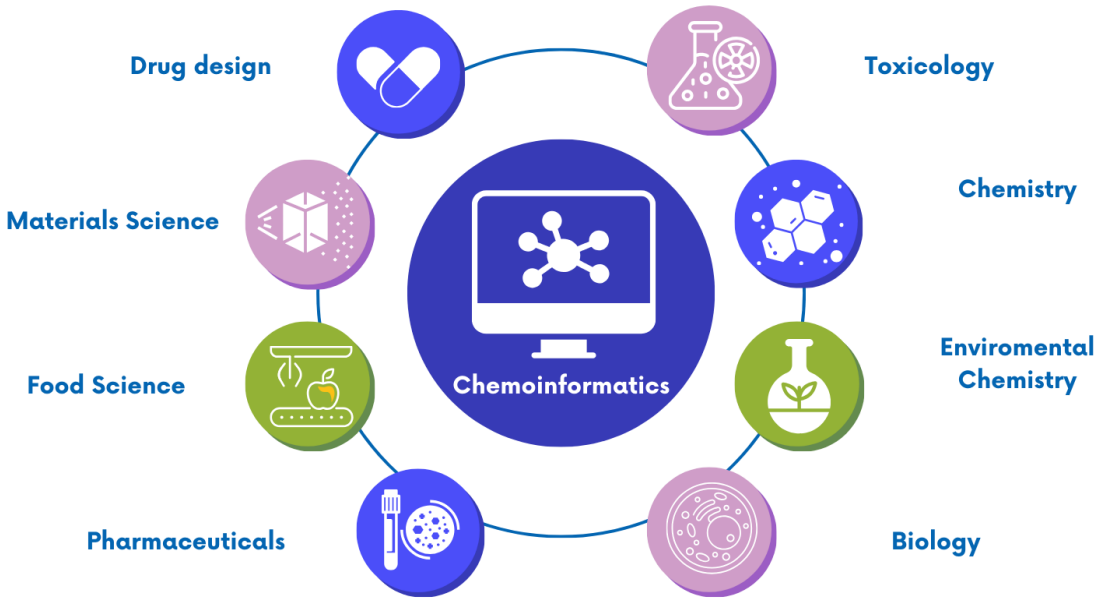


Figure 1. Overview of applications of Chemoinformatics.

One of Chemoinformatics's most widely acknowledged accomplishments is its contribution to providing access to chemical information within databases.[3] The vast volume of data related to chemical compounds, encompassing their physical, chemical, and biological properties, has promoted the creation of databases designed for the efficient storage and electronic dissemination of this information. To further enhance the utility of these databases, various computational methods have been devised. These methods facilitate more effective information retrieval by enabling comprehensive searches based on complete structures, substructures, and similarity. This approach streamlines data mining and enhances the overall efficiency of database searches.

In the same context, different types of molecular representations have enabled improved searches and expanded applications in various areas of chemistry. Examples include the development of new chemical compounds (*de novo* design),[4,5] properties prediction such as absorption, distribution, metabolism, and excretion (ADME),[6] structure-activity relationships (SAR), and structure-properties relationships (SPR),[7] and development of new chemical descriptors.[8] Notable examples of molecular descriptors commonly employed in drug discovery applications include structural fingerprints and molecular properties. Fingerprints encode a molecule's molecular fragments or functional groups, such as Molecular ACCes System (MACCS) Keys[9] and extended connectivity fingerprint (ECFP).[10] Typical molecular properties calculated are whole molecular properties of pharmaceutical interest that are part of empirical rules to assess drug-likeness.[11]

Chemical space analysis, molecular docking, and applying similarity concepts are prominent topics in computer-aided drug design (CADD). These topics are widely utilized in the pharmaceutical industry, universities, and research centers, demonstrating high-frequency drug discovery and development applications. The chemical space is the conceptual basis of Chemoinformatics.[12] Several definitions of chemical space are reviewed elsewhere.[13] Recently, the notion of the chemical multiverse has emerged, representing a collection of chemical spaces for a given set of compounds, each characterized by a distinct set of descriptors.[14] This concept finds applications in drug design, encompassing diversity analysis, SAR and SPR analysis, and the design of molecular libraries.[13]

Applications of the similarity concept, such as SAR/SPR analysis and activity landscapes, have been very useful in describing changes in biological activity associated with changes in chemical structures and the subsequent design of new compounds with improved activities.[15] Often, these differences can be confirmed with tools such as molecular docking, which take into account the mechanism of action at te structural level. However, their application is limited to the availability of information related to the therapeutic target.

In order to explore the topics mentioned above further, the following is a Spanish handbook that addresses concepts and tools of Chemoinformatics with applications in drug design. In recent years, there have been several contributions to teaching Chemoinformatics[16–18] and machine learning for chemists[19] in an organized and formal manner. However, it is still necessary to improve the teaching and

dissemination of the applications of Chemoinformatics in Latin America.[20] This is to promote international collaboration and technological competence by equipping professionals with advanced skills in handling computational tools for chemical data analysis. This not only benefits academic research but also the chemical and pharmaceutical industry.

In this context, the goal of developing an electronic handbook on Chemoinformatics in Spanish is to strengthen the understanding of its basic principles in chemistry students and professionals and to contribute to the ability of users to handle and interpret computational techniques associated with this scientific discipline in the context of bioactive compounds. This, in turn, will contribute to the formation of students and researchers who want to learn and benefit from the appropriate use and implementation of computational methods for their professional development.

GitBook structure and content

The Chemoinformatics handbook is implemented within the GitBook platform, and it is organized into nine chapters that seek to promote the acquisition of basic concepts of Chemoinformatics and to develop competences for the search, acquisition, and analysis of chemical information employing programming and open access computational tools. The handbook covers a broad and diverse range of topics, including a general introduction to Python concepts and packages and basic commands on Linux. It also covers the search and analysis of chemical information using different databases and software for bibliometric visualization. It guides users on the applications of different molecular representations of low molecular weight organic compounds and further provides proficiency in utilizing databases pertinent to drug research through the use of Application Programming Interfaces (APIs), and instructs on building compound databases annotated with biological activity. Additionally, it introduces Exploratory Data Analysis (EDA) for examining physicochemical properties, exploring correlations within datasets, and visualizing the chemical space. The concept of a chemical multiverse is introduced and exemplified through a chemical art gallery.[21] Applications of the similarity concept to conduct structure-activity relationships, such as quantitative structure-activity relationships (QSAR) and activity landscape modeling, are also covered. The GitBook also exemplifies published examples of chemical library enumeration in more detail.[22,23] This can prove highly beneficial for investigating proposed methodologies

in chemical synthesis, facilitating the exploration of an affordable chemical space through the utilization of open-access Chemoinformatics tools. Finally, a general overview of molecular docking is provided, including steps to conduct protein-ligand molecular docking studies using open-access programs.

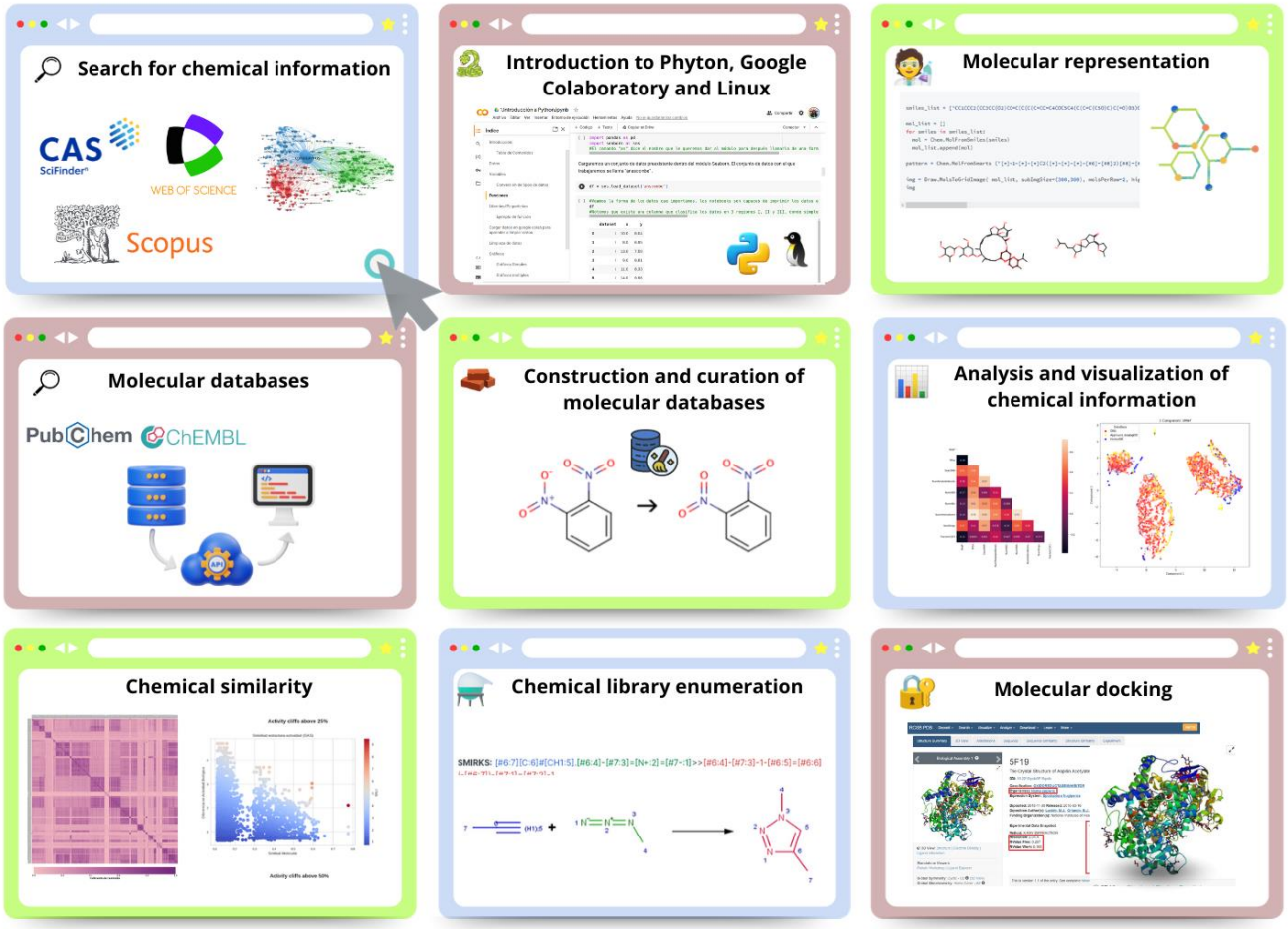


Figure 2. Snapshots of the nine main chapters of the GitBook.

Table 1 summarizes the contents and main objective of each handbook chapter. In each chapter, the general and specific objectives are mentioned, followed by an introduction of the most essential concepts of each topic and their relevance. Subsequently, procedures are developed with 'applicable' examples in research. The chapters end with exercises to reinforce the topics learned. The related Notebooks to explain each topic were developed on Google Colaboratory.

Table 1. Contents of the Chemoinformatics GitBook.

Chapter	Content	Objectives
1. Search for chemical information.	● Scifinder-n. ● Web of Science. ● Scopus. ● CAS Source Index (CASSI). ● Bibliometrics: Tools and Software.	● Enhance the skills of students and professionals in the chemical field to seek scientific information proficiently. ● Become familiar with diverse scientific information types and various tools or engines employed to search for chemical information. ● Evaluate and choose information based on distinct search criteria. ● Recognize the varied forms of scientific publications, understanding their structure and content. ● Acquaint oneself with tools and software employed in bibliometric Analysis and visualization.
2. Introduction to Phyton, Linux, and Google Collaboratory.	● Fundamentals of programming. ● Data cleaning. ● Installation of the environment in local and WSL-based. ● Basic commands on Linux.	● Introduce basic Python definitions and functions. ● Introduce the concept of packages. ● Learn how to import, manage, and clean existing datasets. ● Introduce basic commands in Bash programming language for Linux interaction from the terminal. ● Introduce Bash basic commands for managing, processing, and analyzing data from the Linux terminal.
3. Molecular representation.	● SMILES ● SMARTS ● InChi/Inchi keys	● Introduce the most common molecular representations of low molecular weight organic compounds and their applications in various contexts. ● To learn how to convert compounds between the different molecular representations as appropriate. ● Introduce the use of RDKit, py3Dmol, and smilesDrawer packages to manage chemical structures. ● Apply the knowledge learned to compare structures, filter databases, and visualize molecules with specific characteristics.
4. Molecular databases.	● PubChem ● ChEMBL ● DugBank ● ZINC ● ChemSpider	● Acquire proficiency in utilizing databases pertinent to drug research, including ChEMBL, PubChem, DrugBank, and ZINC. ● Identify the specific categories of information accessible within each resource, enabling streamlined and efficient information retrieval. ● Become familiar with using Application Programming Interfaces (APIs) to access information from public databases programmatically.
5. Construction and curation of molecular databases.	● Construction of compound databases.	● Build compound databases annotated with biological activity. ● Acquire knowledge about the molecular characteristics necessary for subsequent *in silico* studies. ● Identify and eliminate molecules that could potentially alter computational calculations. ● Curate compound databases using RDKit and Molvs modules.

6. Analysis and visualization of chemical information.	● Calculation and Analysis of molecular descriptors. ● Visualization of chemical space.	● Introduce Exploratory Data Analysis (EDA) for chemical data. ● Employ visual methodologies to examine physicochemical properties crucial to pharmaceutical applications, along with descriptors linked to molecular complexity. ● Explore potential correlations among various variables within the dataset. ● Utilize chemical space visualization methods to generate comprehensive profiles of chemical databases. ● Introduce the concept of a chemical multiverse and showcase it through a Chemical Art gallery.
7. Chemical similarity.	● Molecular representation (fingerprints). ● Similarity functions. ● QSAR. ● Activity landscapes.	● Introduce the concept of chemical similarity and its applications in drug design. ● Acquire a fundamental understanding of the critical components used to assess the similarity between chemical compounds. ● Study structure-activity relationships through QSAR and activity landscape modeling.
8. Chemical library enumeration.	● Using known synthesis reaction schemes and available reagents. ● Using transformation rules retrieved from the literature.	● Illustrate examples of virtual chemical library enumeration. ● Gain proficiency in employing SMARTS and SMIRKS for encoding chemical reactions and transformations.
9. Molecular docking.	● Ledock. ● AutoDockVina.	● Give a general, non-exhaustive overview of what a molecular docking study is. ● Explain the steps for a protein-ligand molecular docking study with two open-access programs.

Implementation and discussion

A free online workshop was recently conducted on the UAMedia platform (https://www.uamediadigital.com/cursos-online) to use and disseminate the GitBook among chemistry students and professionals. The workshop lasted 20 hours and covered the content summarized in Table 1, excluding molecular docking, which was subsequently included based on attendee requests. The participants came from 11 Mexican academic institutions and 14 universities in Central and South America. Of 67 attendees who answered the survey, most of them were academics (35.8%) and graduate students (31.3%). In Figure 3, the demographic distribution of attendees is presented, along with some responses to questions related to GitBook implementation. Notably, for most participants, while possessing fundamental knowledge of Chemoinformatics, the workshop, and the handbook were their first approaches to programming with Python.

An inquiry was made regarding the viability of autonomously implementing the material presented in GitBook without instructor guidance. Additionally, the feasibility of utilizing GitBook as a teaching resource was explored for respondents who identified themselves as teachers. The responses indicated a positive outlook for both queries. It is important to mention that since the GitBook content is at an introductory level, it can be incorporated into any undergraduate or graduate chemistry program. While most examples during the course focused on drug design, its application extends beyond and can be effectively utilized in diverse fields such as materials science or food chemistry, among others.

As for the perceived difficulty in understanding the concepts and related codes in the GitBook (Figure 4), this increases as practical applications such as chemical space visualization, similarity applications in a drug design context, and chemical library enumeration are implemented.

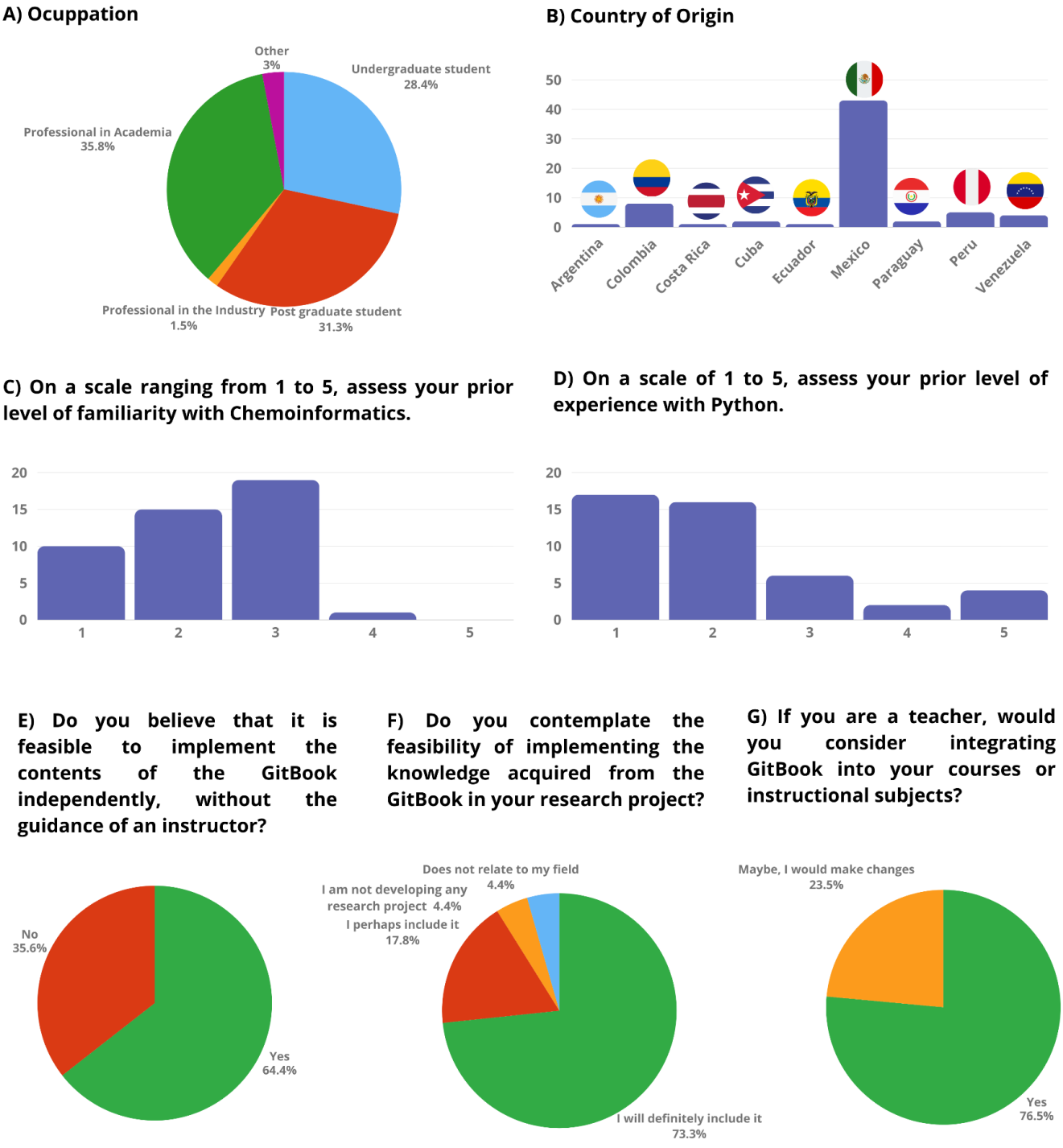
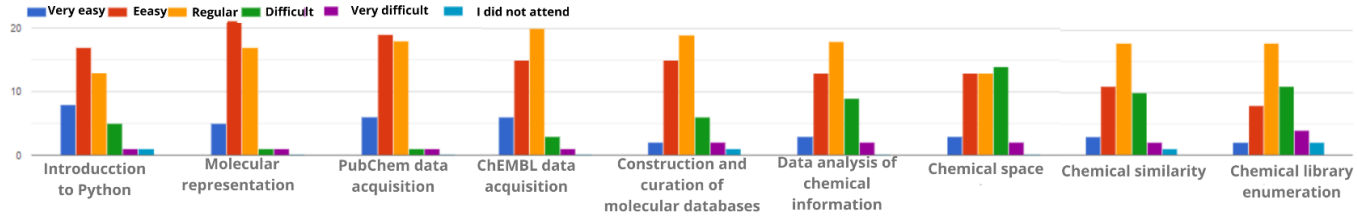


Figure 3. Information about attendees at a Chemoinformatics Workshop and answers to questions about implementing the Chemoinformatics GitBook. **A)** Pie chart indicating the occupation of attendees; **B)** Demographic distribution of attendees. **C)** Distribution of previous knowledge concerning Chemoinformatics, and **D)** Python language**. E), F),** and **G)** Pie charts indicating answers to questions related to implementing GitBook.

E) On a scale ranging from very easy to very difficult, assess the level of complexity in comprehending the concepts associated with each topic.



F) On a scale from very easy to very difficult, evaluate the level of complexity associated with comprehending the code related to each topic.

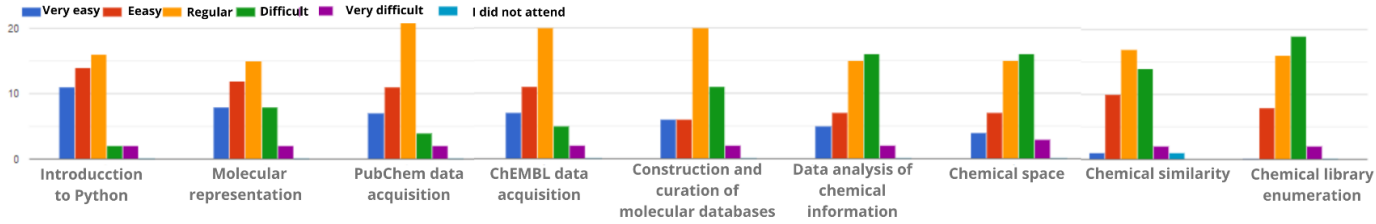


Figure 4. Perceived difficulty in understanding the concepts and related codes for each topic.

Conclusions

Chemoinformatics is a scientific discipline that has emerged in response to the need to manage, classify and interpret chemical information efficiently. One of the main applications of Chemoinformatics has been in drug development since it has facilitated the integration of chemical and biological data to generate information and, ultimately, helpful knowledge (for instance, predictive models of biological activity). The herein Chemoinformatics handbook, constructed within the GitBook platform and organized into nine chapters, is a valuable educational resource for Spanish-speaking students and professionals entering the Chemoinformatics field, showcasing real drug design applications. The fundamental concepts and applications can be adapted and extended to various chemical-related disciplines. Evaluation within a group predominantly comprised of teachers and graduate students indicated a positive perception of the handbook's utility for implementation in undergraduate and graduate chemistry programs. The emphasis on teaching Chemoinformatics through open-access tools and the fact that the manual is published in Spanish aligns with the broader goal of democratizing science and cultivating interest among students and professionals in the chemical field. This approach facilitates learning in programming, machine learning, and computational techniques and contributes to optimizing drug development costs through informed prioritization and rationalization of various drug development processes.

Acknowledgments

We thank Norberto Sanchez-Cruz for his advice on the use of AutoDock Vina. The technical assistance and enthusiastic participation of Jesús Armando Rufino Valencia and Maria de la Luz Seseña Alcalde (School of Chemistry, UNAM) is greatly acknowledged. FISG, DLPR, RCL, ALCH, JFAT, and AGG thank the Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCyT), Mexico, for the Postgraduate grants 848061, 888207, 1099206, 847870, 1270553, and 912137 respectively.

References

(1) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer, 2007.

(2) López-López, E.; Bajorath, J.; Medina-Franco, J. L. Informatics for Chemistry, Biology, and Biomedical Sciences. *J. Chem. Inf. Model.* **2021**, *61* (1), 26–35. https://doi.org/10.1021/acs.jcim.0c01301.

(3) Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21* (2), 151. https://doi.org/10.3390/molecules21020151.

(4) Mouchlis, V. D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A. G.; Aidinis, V.; Lynch, I.; Greco, D.; Melagraki, G. Advances in de Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci.* **2021**, *22* (4). https://doi.org/10.3390/ijms22041676.

(5) Chávez-Hernández, A. L.; López-López, E.; Medina-Franco, J. L. Yin-Yang in Drug Discovery: Rethinking de Novo Design and Development of Predictive Models. *Front. Drug Discov.* **2023**, *3*. https://doi.org/10.3389/fddsv.2023.1222655.

(6) Daina, A.; Michielin, O.; Zoete, V. SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Sci. Rep.* **2017**, *7*, 42717. https://doi.org/10.1038/srep42717.

(7) Ragno, R.; Esposito, V.; Di Mario, M.; Masiello, S.; Viscovo, M.; Cramer, R. D. Teaching and Learning Computational Drug Design: Student Investigations of 3D Quantitative Structure–Activity Relationships through Web Applications. *J. Chem. Educ.* **2020**, *97* (7), 1922–1930. https://doi.org/10.1021/acs.jchemed.0c00117.

(8) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of

Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57* (8), 1757–1772. https://doi.org/10.1021/acs.jcim.6b00601.

(9) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(10) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(11) A Decade of Drug-Likeness. *Nat. Rev. Drug Discov.* **2007**, *6* (11), 853–853. https://doi.org/10.1038/nrd2460.

(12) Varnek, A.; Baskin, I. I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol. Inform.* **2011**, *30* (1), 20–32. https://doi.org/10.1002/minf.201000100.

(13) Saldívar-González, F. I.; Medina-Franco, J. L. Approaches for Enhancing the Analysis of Chemical Space for Drug Discovery. *Expert Opin. Drug Discov.* **2022**, *17* (7), 789–798. https://doi.org/10.1080/17460441.2022.2084608.

(14) Medina-Franco, J. L.; Chávez-Hernández, A. L.; López-López, E.; Saldívar-González, F. I. Chemical Multiverse: An Expanded View of Chemical Space. *Mol. Inform.* **2022**, *41* (11), e2200116. https://doi.org/10.1002/minf.202200116.

(15) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure–activity Landscapes. *Drug Discov. Today* **2009**, *14* (13), 698–705. https://doi.org/10.1016/j.drudis.2009.04.003.

(16) 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. Teaching Cheminformatics through a Collaborative Intercollegiate Online Chemistry Course (OLCC). *J. Chem. Educ.* **2021**, *98* (2), 416–425. https://doi.org/10.1021/acs.jchemed.0c01035.

(17) Sydow, D.; Rodríguez-Guerra, J.; Kimber, T. B.; Schaller, D.; Taylor, C. J.; Chen, Y.; Leja, M.; Misra, S.; Wichmann, M.; Ariamajd, A.; Volkamer, A. TeachOpenCADD 2022: Open Source and FAIR Python Pipelines to Assist in Structural Bioinformatics and Cheminformatics Research. *Nucleic Acids Res.* **2022**, *50* (W1), W753–W760. https://doi.org/10.1093/nar/gkac267.

(18) Walters, P. *Practical Cheminformatics*. https://practicalcheminformatics.blogspot.com/ (accessed 2024-01-10).

(19) Lafuente, D.; Cohen, B.; Fiorini, G.; García, A. A.; Bringas, M.; Morzan, E.; Onna, D. A Gentle Introduction to Machine Learning for Chemists: An Undergraduate Workshop Using Python Notebooks for Visualization, Data Processing, Analysis, and Modeling. *J. Chem. Educ.* **2021**, *98* (9), 2892–2898. https://doi.org/10.1021/acs.jchemed.1c00142.

(20) Miranda-Salas, J.; Peña-Varas, C.; Valenzuela Martínez, I.; Olmedo, D. A.; Zamora, W. J.; Chávez-Fumagalli, M. A.; Azevedo, D. Q.; Castilho, R. O.; Maltarollo, V. G.; Ramírez, D.; Medina-Franco, J. L. Trends and Challenges in Chemoinformatics Research in Latin America. *Artificial Intelligence in the Life Sciences* **2023**, *3*, 100077. https://doi.org/10.1016/j.ailsci.2023.100077.

(21) Gaytán-Hernández, D.; Chávez-Hernández, A. L.; López-López, E.; Miranda-Salas, J.; Saldívar-González, F. I.; Medina-Franco, J. L. Art Driven by Visual Representations of Chemical Space. *J. Cheminform.* **2023**, *15* (1), 100. https://doi.org/10.1186/s13321-023-00770-4.

(22) Saldívar-González, F. I.; Huerta-García, C. S.; Medina-Franco, J. L. Chemoinformatics-Based Enumeration of Chemical Libraries: A Tutorial. *J. Cheminform.* **2020**, *12* (1), 64. https://doi.org/10.1186/s13321-020-00466-z.

(23) Saldívar-González, F. I.; Navarrete-Vázquez, G.; Medina-Franco, J. L. Design of a Multi-Target Focused Library for Antidiabetic Targets Using a Comprehensive Set of Chemical Transformation Rules. *Front. Pharmacol.* **2023**, *14*, 1276444. https://doi.org/10.3389/fphar.2023.1276444.