**Data Chemistry OLCC Proposed Course Content**

**Unit 1: Introduction to Data Chemistry and Course Overview**

*Objectives:*

* Review the history of data science.
* Review applications of data science in chemistry.
* Review current challenges and unsolved issues in data chemistry.
* Provide an overview of the course.

*Activities:*

* Complete the interactive PubChem training course developed by the National Library of Medicine (<https://www.nlm.nih.gov/oet/ed/pubchem/tutorial/index.html>).
* Complete Jupyter notebooks that cover basic Python programming topics (data types, variables, conditions, loops, etc.).

*Materials:*

* An introductory material will be written based on the previous OLCC materials available at LibreTexts:
  + <https://chem.libretexts.org/Courses/University_of_Arkansas_Little_Rock/ChemInformatics_(2017)%3A_Chem_4399_5399/1.1%3A_Introduction_to_Cheminformatics>
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/01%3A_Introduction/1.01%3A_Introduction>
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/01%3A_Introduction/1.02%3A_Brief_History_of_Cheminformatics>
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/01%3A_Introduction/1.03%3A_Introduction_to_Data_and_Databases>

**Unit 2: Understanding Public Chemical Databases**

*Objectives:*

* Review private and public chemistry databases developed for various purposes (e.g., Chemical Abstract Service & SciFinder, PubChem, PDB, DrugBank, and many others.)
* Review the roles of public databases in “open” science and big data era.
* Describe data organization in PubChem.
* Review information available in PubChem.

*Activities:*

* Complete the PubChem tutorial published in Current Protocols ((<https://doi.org/10.1002/cpz1.217>). The protocols in this tutorial will provide an overview of search and retrieval using PubChem, which will be covered in Units 3 and 4.
* Complete Jupyter notebooks that cover data visualization using the PubChem Element data.
* Learn technical aspects of PubChem.
  + Data organization in PubChem: What are PubChem substances and compounds?
  + PubChem Data Source page.
  + “Versioning” in PubChem data.
  + Live and non-live records in PubChem: PubChem records can be live, dead, and resurrected.
  + Legacy records in PubChem

*Materials:*

* The existing materials will be used with some updated content.
  + <https://chem.libretexts.org/Courses/University_of_Arkansas_Little_Rock/ChemInformatics_(2017)%3A_Chem_4399_5399/4%3A_Understanding_Public_Chemical_Databases>
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/03%3A_Database_Resources_in_Cheminformatics>

**Unit 3: Chemical Representations**

*Objectives:*

* Review various chemical names used for representing small molecules, including common names, systematics names, brand names, and international non-proprietary names (INN).
* Learn how to represent small molecules using line notations such as SMILES, SMARTS, InChI, and InChIKeys.
* Understand the limitations of commonly used chemical representations.
* Review issues with interchangeability between chemical representations and its impact on data integration between different sources.

*Activities*

* Search PubChem (and a few other public databases) interactively using various types of queries including chemical names and chemical line notations (including SMILES, SMARTS, InChI, InChIKey).
* No programming topics for this week. Students who do not have strong programming skills will use this week as a makeup time.

*Materials:*

* The existing materials will be used after minimal changes.
  + <https://chem.libretexts.org/Courses/University_of_Arkansas_Little_Rock/ChemInformatics_(2017)%3A_Chem_4399_5399/2.1%3A_Chemical_Representations_on_Computer_Part_I>
  + <https://chem.libretexts.org/Courses/University_of_Arkansas_Little_Rock/ChemInformatics_(2017)%3A_Chem_4399_5399/2.2%3A_Chemical_Representations_on_Computer%3A_Part_II>
  + <https://chem.libretexts.org/Courses/University_of_Arkansas_Little_Rock/ChemInformatics_(2017)%3A_Chem_4399_5399/2.3%3A_Chemical_Representations_on_Computer%3A_Part_III>
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/02%3A_Representing_Small_Molecules_on_Computers>

**Unit 4: Searching chemical databases**

*Objectives:*

* Understand diverse contexts of chemical “identity” (different tautomeric states, stereochemistry, salt forms, etc.).
* Understand the concept of 2-D and 3-D molecular similarity.
* Understand the concept of substructures and superstructures.

*Activities:*

* Perform various types of chemical structure searches including identity, 2-D/3-D similarity, superstructure/substructure searchers.
* Learn how to refine previous search results based on molecular properties (such as molecular weight, XLogP, heavy atom counts).
* Learn how to perform a complex queries by combining multiple search results using Boolean operators (AND, OR, and NOT)
* Retrieve chemicals that belong to a given class using the PubChem Classification Browser.
* Generate an image of the structure of a given chemical using the PubChem imaging service.
* Review how to find chemical information for a given gene, protein, pathway, cell line, or taxon.

*Materials:*

* The existing materials will be used after minimal changes.
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/04%3A_Searching_Databases_for_Chemical_Information>

**Unit 5: Exploratory Data Analysis of PubChem BioAssay Data**

*Objectives:*

* Explain various types of quantitative measures of bioactivity of chemicals (e.g., IC50, EC50, Ki & Kd).
* Explain the heterogeneity of bioactivity data in PubChem.
* Explain the stages of biological assay experiments.

*Activities:*

* Learn how to bioactivity data for small molecules tested in assays archived in PubChem.
* Write a python script to compute descriptive statistics (e.g., mean, median, mode, standard deviation, minimum, maximum, etc.) for the bioactivity data downloaded from PubChem.

*Materials:*

* New materials will be created from scratch.

**Unit 6: Building an Interactive Dashboard using Tableau**

*Objectives:*

* Build an interactive dashboard that visualizes bioactivity data from PubChem.

*Activities:*

* Use Tableau (<https://www.tableau.com/>) to create an interactive dashboard that visualizes bioactivity data for a given target.
* Share the created dashboard with others through the Tableau Public website (<https://public.tableau.com/app/discover/>).

*Materials:*

* New materials will be created from scratch.

**Unit 7: Quantitative Structure-Activity/Property Analysis (QSAR/QSPR)**

*Objectives:*

* Explain what QSAR and QPAR are.

*Activities:*

* Write a python script that builds a QSAR model from public data.

*Materials:*

* The existing materials available on Libertexts will be used:
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/05%3A_5._Quantitative_Structure_Property_Relationships>

**Unit 8: Molecular Similarity**

*Objectives:*

* Explain what molecular fingerprints are.
* Explain the pros and cons of the structural keys and hashed fingerprints.
* Explain the similarity metrics used to quantify structural similarity between molecules.

*Activities:*

* Write a python script that tests molecular fingerprints and similarity coefficients.
* Perform 2-D and 3-D similarity searches to find structurally similar molecules for a given query (revisiting the topic covered in Unit 4).

*Materials:*

* The existing materials available on Libertexts will be used:
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/06%3A_Molecular_Similarity>

**Unit 9: Computer-Aided Drug Discovery and Design**

*Objectives:*

* Describe the overall process of drug development.
* Explain the role of virtual screening in drug discovery and development.
* Describe the differences between structure-based and ligand-based drug discovery approaches.

*Activities:*

* Perform a ligand-based virtual screening against the PubChem Compound database.
* Perform a molecular docking experiment.

*Materials:*

* The existing materials available on Libertexts will be used:
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/07%3A__Computer-Aided_Drug_Discovery_and_Design>

**Unit 10: SQL with python**

*Objectives:*

* Describe basic terminologies related to databases.
* Understand the basic syntax of the Structured Query Language (SQL)

*Activities:*

* Build a chemical database using SQLite (<https://www.sqlite.org/>).
* Access the database using SQL queries.
* Write a python script that access the database using SQL queries.

*Materials:*

* New materials will be created from scratch

**Unit 11: Linked Open Data and Knowledge Graph**

*Objectives:*

* Explains how knowledge can be encoded in a machine-readable format using the resource description framework (RDF).
* Understand the basic syntax of the SPARQL Protocol and RDF Query Language.

*Activities:*

* Access PubChemRDF data using the PubChemRDF REST-ful interface (https://pubchem.ncbi.nlm.nih.gov/docs/rdf#section=4-RESTful-Interface).
* Access a pre-built virtuoso database containing PubChemRDF data using SPARQL queries.

*Materials:*

* New materials will be created.

**Unit 12: Artificial Intelligence 1: Unsupervised Machine Learning**

*Objectives:*

* Explain what artificial intelligence, machine learning, and deep learning are:
* Describe the differences between unsupervised and supervised machine learning.
* Explain what chemical structure clustering is.

*Activities:*

* Perform a chemical structure clustering of a set of molecules using KNIME as well as a python script.

*Materials:*

* An extended version of the existing material will be created:
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/08%3A_Machine-learning_Basics/8.01%3A_Machine_Learning_Basics>

**Unit 13: Artificial Intelligence 2: Supervised Machine Learning**

*Objectives:*

* Explain what supervised learning is.
* Explain commonly used supervised learning algorithms.
* Explain the limitations of supervised learning approach.

*Activities:*

* Build a bioactivity prediction model from PubChem’s bioactivity data, using KNIME as well as a python script.

*Materials:*

* An extended version of the existing material will be created:
  + <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/08%3A_Machine-learning_Basics/8.01%3A_Machine_Learning_Basics>

**Unit 14: Artificial Intelligence 3: Generative AI and Large Language Models**

*Objectives:*

* Explain what generative AI and large language models (LLMs) are.
* Explain what prompt engineering is.
* Explain the limitations of large language models.
* Review the applications of generative AID and LLMs in chemistry.

*Activities:*

* To be determined, based on the availability of computational resources suitable for LLMs.
* This is the last unit of the course, likely before the final exam. A brief review of all topics covered during the course is highly recommended.