CHEM 4580/5580 Molecular Informatics

Instructor	Scott Reed
Office	Science Building 3071E
Phone	303-315-7634
E-mail	scott.reed@ucdenver.edu
Office Hours	Th 12:00 – 1:00 PM or by appointment by Zoom (https://ucdenver.zoom.us/j/3152346473)

Course Dates/Times	Fr 9:00AM - 10:40AM
Course Location	NC 1314
Course Format:	Lecture (with live coding. Laptop required)
Semester Hours:	3

Need to Know:
You should be comfortable with the first 3 sections of this tutorial:
https://education.molssi.org/python_scripting_cms/
If you are not, discuss a plan to work through these modules during the first week of the semester to stay on track. The first two weeks we will cover sections 4-6. Extra office hours will be provided for this on request.

Before the First Class:
This class will involve extensive in-class use of python. You will need to have a laptop in class that has python installed using Anaconda. Details on how to do this can be found here:
https://education.molssi.org/python_scripting_cms/setup.html
Please contact Prof. Reed before the first class if you need to borrow a laptop or need help installing python following the instructions at that link.
If you are ever unable to access a laptop with Anaconda on it, you can as a backup do the same work on a web browser at https://colab.research.google.com/ or https://chemcompute.org
Other programs can be used to access Jupyter notebooks such as Pycharm (academic version free) or Visual Studio Code (free). Both of these offer advantages but have higher learning curves. If you plan to continue coding after this class it is worth the time investment to learn these programs. Both can handle the Jupyter notebook format are suitable for class.

Background:

We are at a tipping point in the history of Chemistry. Prior to 1828, molecules exclusively came into existence by natural processes driven by chance, such as evolution. Since 1828 when Friedrich Wöhler synthesized urea, a new method of molecular creation became available. In this new method, a molecule is conceived of inside a human brain first and then it is prepared in the lab. That era has been dominant for a long time, but its dominance is coming to an end. In the next era of Chemistry, molecules will come into existence first as a computational representation. Consider that only approximately 10^8 out of the estimated 10^{63} possible drug-like molecules have ever been prepared. Consider also that the number of ways to combine amino acids into small proteins exceeds the number of atoms in the universe. These vast areas of

potential new molecules are most often created computationally before they are created experimentally. Databases of billions of molecules that have never been synthesized but are synthesizable are already available. Therefore, most molecules synthesized in the future will exist first as a computational representation. In this new era, molecules first exist in databases that already contain information on the molecule's properties and the best method for their preparation. All of this can occur before a scientist enters the lab. To study molecules without taking advantage of these resources is a dated and arguable ineffective way to study molecular science.

Course Description:

This course resides at the intersection between Chemistry, Biochemistry, and Data Science. The course covers fundamental concepts of Chemical and Biochemical Informatics and provides students with hands on experience in using computational tools to manipulate chemical and biochemical data. Students will learn fundamentals of data science, database management, data structure, data representation, data visualization, and data analysis as applied to Chemistry and Biochemistry.

The course does not require extensive programming experience. Examples explored in class and in homework will be built using Python code within Jupyter Notebooks or Google Colab notebooks such that students can explore new topics while remaining focused on the underlying molecular concepts and computer methods which allow them to manage large amounts of molecular information and to find relationships between the structure and properties of molecules. Data mining approaches will be explored as will classification algorithms and chemical similarity analysis methods. Students will learn about the applications of cheminformatics in drug discovery, such as compound selection, virtual library generation, virtual high throughput screening which can check for potential molecules that have the potential to be developed into drugs. Throughout this course we will use artificial intelligence and large language models to speed our access to and understanding of these topics.

While this course is not a pre-requisite for 4510 Computational Chemistry, CHEM 4640 AI in Chemistry and Biochemistry, or CHEM 4845 Molecular Modeling and Drug Design, the skills developed in this course will work synergistically with those courses and will allow you to get more from your experiences in those courses or from your experience in a research lab.

Computational tools are increasingly necessary to the study of Chemistry and Biochemistry. Students in this course will learn methods to represent chemical and biochemical structures that are machine readable and searchable and learn how to retrieve chemical information from application programming interfaces (APIs) for databases of chemical structures, properties, and experimental data. Students will learn how to perform calculations of molecular properties.

Topics covered include:

Representing Small Molecules on Computers
Database Resources in Cheminformatics

Searching Databases for Chemical Information
Quantitative Structure Property Relationships
Molecular Similarity
Computer-Aided Drug Discovery and Design
Machine-learning Basics
Python programming
Generating 2D and 3D structures
Chemical similarity analysis
The concept of pharmacophores
Virtual screening
Computer aided drug design

Chemical and Biochemical databases we will likely explore include:
Protein databases such as: PDB, Gnomad, uniport, and Alphafold.
Chemical databases such as: Pubchem, Cambridge Structural Database, and ZINC.

After building a basic set of skills, students will select a topic to explore independently in greater depth and to write a report and present to the class based upon their exploration. Students completing this course will have the tools they need to design independent projects in data analysis, data storage, and web based or native application design.

Course Objectives:
At the completion of the course, students will be better able to generate and manipulate molecular information. While this course will not focus on web application design, student will learn the basics that can form the foundation of a database driven web application.
At the completion of this course, students will be equipped to:

- Understand how to retrieve data from application programming interfaces
- Understand database schema, data structure, and data management.
- Understand how to interconvert between representations of molecules.
- Understand various chemical, biochemical, and genomic data types and how they are represented digitally.
- Understand methods to represent, visualize, and analyze chemical and biochemical data.
- Understand how to use relational database management system (DBMS) with SQL to access and store molecular data.

Text:
No textbook needs to be purchased for this course. We will use an on-line textbook prepared by the ACS Division of Chemical Education's Committee on Computers in Chemical Education. This Cheminformatics Online Chemistry Course (OLCC) is available through Canvas.
Additional References and reading material will be provided for some topics.
Students must bring a laptop to class each week or apply for a no-cost loan of a laptop for the semester from the department.

Letter grades are determined based on the numerical score. The standard correlation is as follows: 90.00 - 100 = A, 89.00 - 89.99 = A-, 87.00 - 88.99 = B+, 80.00 - 86.99 = B, 79.00 -79.99 = B-, 77.00 - 78.99 = C+, 70.00 - 76.99 = C, 69.00 -69.99 = C-

Assignments:

Homework	30%
Code Review	10%
Midterm Exam	20%
Project	20%
Final Presentation	20%

Homework assignments will involve weekly coding assignments embedded in Jupyter notebooks. Assignments can be submitted as ipynb files thorugh email to scott.reed@ucdenver.edu. An option to submit through Github will also be provided. No late assignments accepted.

Each student will complete an independent project. Project plan presentations in the middle of the semester will describe an informatics resource and a plan for using that resource.

Students will be required to include in their final project an operational Jupyter Notebook and will be required to provide feedback to other students through the final project design process. Code Review will also involve providing regular feedback to classmates.