Towards Artificial Intelligence in Chemistry via the Rational Design of Chemical Compounds and Automated Discovery

Course outline

Lesson 1 (lecture)

Rational Design, Serendipity, and Virtual Screening. Course Overview.

The significance of Chemistry in modern economy. Heuristic approach, serendipity. The concept of accelerated serendipity proposed by David W. C. MacMillan. Disadvantages of serendipity-driven research.

What is the rational design in chemistry? The concept of virtual screening. High throughput experimentation. Robots in chemistry.

Artificial intelligence in Chemistry and rational design.

Applications of rational design in chemistry: notable examples. Linear free energy relationships. Medicinal chemistry. Materials science. Catalysis. Organic synthesis. Predicting reactions. Predicting reaction conditions.

Discussion: course structure.

References to additional resources and literature. International scientific journals publishing articles on the topics of rational design and automated chemical discovery (Nature & Science, ACS, RSC, Wiley, Elsevier, and others, e.g.). RSS feeds of ASAP articles.

Lesson 2 (lecture with interactive demonstrations)

Rational Design Tools 1: Python 3 and Some Common Libraries

Python 3. Some characteristics of the programming language. A few facts about the history of Python. Philosophy of the language, Python Enhancement Proposals. PEP 20, PEP 8.

Anaconda. History and installation. pip as an alternative key package manager.

Jupyter Notebook. A short introduction to the basic features.

NumPy and SciPy. Features. Installation in Anaconda.

Scikit-learn, Pandas, Matplotlib. Features, installation.

Lesson 3 (lecture with interactive demonstrations)

Rational Design Tools 2: Specialized Libraries, Cheminformatics and Quantum Chemistry Software

PyTorch. A brief description of features. Installation.

Visualizing molecules with free software. Avogadro and Jmol.

Specialized Python libraries for cheminformatics. RDKit, Mordred, OpenBabel.

Computational chemistry and quantum chemistry. Electronic orbitals. Hamiltonian. Further reading on quantum chemistry. Advantages and disadvantages of quantum chemical methods. Molecular dynamics and metadynamics.

Open/free online resources.

Lesson 4 (lecture with interactive demonstrations)

Rational Design Tools 3: Table Chemical Data, Molecular File Formats, and Chemical Databases

Data formats. Recording experimental and computational data. Tabular formats .xslx, CSV and TSV. JSON and XML.

Representing the structure of molecules. 2D: canonic SMILES and modifications of canonic SMILES, InChi. 3D: XYZ, MDL Molfile, PDB.

Visualizing molecular dynamics trajectories.

Chemical databases. Public chemical databases with computational and experimental data. SciFinder and Reaxys. Common DB problems. The FAIR guiding principles.

Lesson 5 (lecture with interactive demonstrations)

Illustrated Discussion of Some Machine Learning and Data Analysis Methods

Machine learning tasks: regression, classification, clustering.

Suggested books and courses on machine learning (ML).

ML project pipeline.

Underfitting and overfitting. Bias-variance trade-off.

Principal component analysis. Analyzing Pandas dataframes.

Data analysis and ML practice: demonstration with a Jupyter Notebook.

Linear regression. Algorithm of k-nearest neighbors and support vector machines. Random forest. CatBoost.

Some remarks on deep neural networks applications in chemistry.

Lesson 6 (lecture with interactive demonstrations)

Free Energy Relationships and Linear Regression

Thermochemical equations and thermodynamics. Thermodynamic potentials. Equations for thermodynamic functions.

Key reaction parameters. ΔG of chemical reactions. Free energy of activation.

Rate of reaction. Arrhenius equation. Eyring-Evans-Polanyi transition state theory. Free energy surface.

Equations of Hammet, Taft, and others. Substituent effect as a descriptor. Abraham's equation. Suggested reading: modern applications of the Abraham equation.

The works of Matthew Sigman et al. as an example of the modern application of (linear) free energy relationships in chemistry.

Implementing linear regression analysis to a relevant chemical problem with Jupyter Notebook.

Lesson 7 (lecture)

Chemical Structure and Molecular Descriptors

A classification of molecular descriptors by dimensionality. Descriptors as mathematical objects: scalars, vectors, matrices, tensors.

States of matter and the corresponding descriptors. Scale of chemical structural features.

Topological indices; substituent effect; quantum chemical; chemical hardness; ionization potential; electron affinity; Tolman's electronic and steric factors; STERIMOL; molecular electric dipole moment; quadrupole moment; connectivity matrix; distance matrix; Coulomb matrix.

What is a good descriptor? What is a good descriptor for atomistic ML?

Lesson 8 (workshop)

Representation of Chemical Structures With the Methods of Cheminformatics and Computational Chemistry

Creation of three-dimensional models of organic substances in the Avogadro editor. Automatic generation of chemical structures in 2D and 3D using PyBel.

Molecular geometry optimization with molecular mechanics and the GFN2-xTB method (black-box level only).

Practice with chemical structure files, XYZ, MOL/SDF, PDB formats, and their conversion using OpenBabel.

Lesson 9 (workshop)

Computation of Structural Descriptors and Properties of Molecules

Computation and analysis of molecular features (descriptors) using Pandas and RDKit. Morgan fingerprints. Tanimoto similarity. DScribe library.

Optional homework. Install and study the features of the smi2gcs library. Apply this library to compute molecular descriptors at home.

Lesson 10 (lecture)

Rational Design in Medicinal Chemistry

Classic computational approaches to drug design. QSAR/QSPR; history of the approach.

Target properties. Toxicity. Therapeutic effect. IC50 and affinity.

Biological activity *in vitro* and *in vivo*.

Mechanisms of biological activity. Lock-and-key model, non-covalent and covalent bonding. ADMET.

Solubility, logP. Drug-likeness.

2D and 3D QSAR. CoMFA. Virtual screening in drug design. The notion of applicability domain. Problems in QSAR studies.

Taste as a complex phenomenon.

Chemometrics. "Electronic tongue" and its application in medicinal chemistry.

QsarDB.

Lesson 11 (workshop)

Design of Substances With Target Properties in Medicinal Chemistry

Case studies from QsarDB.

Using Mordred for molecular descriptor calculations.

Case studies (using Jupyter Notebook)

Estimation of aqueous solubility of organic compounds by using the general solubility equation.

Modelling of antiproliferative activity measured in HeLa cervical cancer cells in a series of xanthene derivatives.

Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis.

Optional homework. Challenging problem for ML enthusiasts: reproduce as much as you can the following study. Site of Metabolism Prediction Based on ab initio Derived Atom Representations. Use the smi2gcs library for molecular descriptor calculations.

Lesson 12 (lecture)

Rational Design of Materials and Recognized Issues of Machine Learning in Chemistry

Descriptors of materials structure and target properties. Coulomb Matrix. Ewald Sum Matrix. Sine Matrix. Smooth Overlap of Atomic Positions (SOAP). DScribe library in use. Radial Distribution Functions (RDFs). Electronic band theory, band gap and d-band center. Defect-governed material properties.

Molecular mechanics methods in a nutshell. Atomistic machine learning. The concept of ML (NN) potentials.

Suggested reading: review articles. Recent advances and applications of machine learning in solid-state materials science [npj Comput. Mater. 2019, 5, 83]. The Role of Machine Learning in the Understanding and Design of Materials [J. Am. Chem. Soc. 2020, 142, 20273].

Using ML/DL to predict materials properties: case studies. Rational design of batteries [Nat. Energy 2019, 4, 383]. How to accurately predict materials properties using models trained on small datasets [npj Comput. Mater. 2018, 4, 25]? How being ignorant of crystal structure, we still can predict material properties [Nat. Commun. 2020, 11, 6280].

Materials properties predicted using ML.

Recognized issues of machine learning in chemistry and material science.

Lesson 13 (lecture)

Rational Design of Catalysts

What is a catalyst? What is catalytic activity? Metrics of activity. Catalyst activity vs. stability (to deactivation). Deactivation via catalyst poisoning.

Why multivariate optimization is important in catalysis.

Activity vs. stability of a catalyst.

ML in the rational design of catalysts: case studies. Feature importance for multiparametric optimization [Adv. Synth. Catal. 2004, 346, 1844]. Molecular volcano plots for homogeneous catalyst prediction [Chem. Sci., 2018, 9, 7069]. Brønsted−Evans−Polanyi Relation and ML for fast prediction of a heterogeneous catalytic mechanism [Nat. Commun. 2017, 8, 14621]. Coupling high-throughput experimentation with ML for catalyst prediction [Science 2018, 360, 186].

Lesson 14 (workshop)

Predicting Properties of Materials and Catalysts

Reproducing case studies from the literature.

Predicting properties of compounds with DScribe and kernel ridge regression.

Machine learning meets volcano plots: computational discovery of cross-coupling catalysts.

Lesson 15 (lecture with interactive demonstrations)

Predicting Organic Reactions with Machine Learning

Overwhelming diversity of organic chemistry. Objectives for ML in organic chemistry. Predicting products from reactants (reagents fixed and annotated); predicting reaction conditions; predicting products from reagents; predicting syntheses (retrosynthetic pathways).

Reasons for the overwhelming diversity of organic chemistry. Isomerism of functional groups. Isomerism of the carbon framework. Spatial isomerism. Conformational isomerism.

Classification(s) of organic reactions. Reaction efficiency metrics.

Using ML/DL to predict organic reactions: case studies. Using machine learning to predict suitable conditions for organic reactions [ACS Cent. Sci. 2019, 5, 1572]. Predicting the regioselectivity of electrophilic aromatic substitution reactions using machine learning [Digital Discovery, 2022, 1, 108]. Predicting outcomes of complex organic chemistry reactions [Chem. Sci., 2018, 9, 6091], [ACS Cent. Sci. 2019, 5, 1572].

Additional Recourses: open datasets of organic reactions.

Optional homework. Challenging problem for ML enthusiasts: reproduce as much as you can the following study. Predicting the regioselectivity of electrophilic aromatic substitution reactions using machine learning [Digital Discovery, 2022, 1, 108]. Use the smi2gcs library for molecular descriptor calculations.

Lesson 16 (lecture with interactive demonstrations)

Machine Learning for Organic Synthesis. Robots in Chemistry. Deep Learning in Drug Design.

Organic synthesis and its importance. Complexity of organic synthesis: from Aspirin to Taxol. The complexity of total synthesis. Retrosynthetic analysis. Synthetic strategies.

Case studies. The application of DL to synthesis planning [Chem. Sci., 2020,11, 3316].

Robots in chemistry. Rational design and innovation. Closed-loop approach and high-throughput experimentation.

The Chemputer: hardware and software, Chemical Description Language (XDL), Chemical Assembly Programming Language, analytical devices, reactivity analysis with an ML model. Discovering new reactions with the chemputer. Further reading:

automation in chemistry and the Chemputer.

Some remarks on deep learning in drug design. A deep learning approach to antibiotic discovery [Cell 2020, 180, 688]. Reinforced adversarial neural computer (RANC) for de novo molecular design [J. Chem. Inf. Model. 2018, 58, 1194].

Lesson 17 (presentations of course projects)

Lesson 18 (presentations of course projects)

Presentations of course projects conducted in small groups. You have to set up a machine learning model on your choice based on examples from the workshops and optional homeworks.