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

Course outline

Class 1 (lecture)

The Notion of Rational Design of Chemical Compounds. Course Overview.

The classical approach to the development of new compounds (new materials, catalysts, biologically active compounds) and chemical reactions. Heuristic approach, serendipity. The concept of accelerated serendipity proposed by David W. C. MacMillan. Disadvantages of serendipity-driven research.

What is rational design in chemistry? Directed discovery or prediction based on the identification of statistical patterns in available data. Causality. Artificial intelligence and rational design.

What are the tools of rational design? Databases. Statistical models and machine learning. Objective: prediction of new compounds with predefined properties. The meaning of prediction in chemistry.

Applications of rational design in chemistry: notable examples.

Medicinal chemistry. The "funnel diagram": from tens of thousands of candidate molecules to a single drug, years of development, and billions of dollars/euros. QSAR/QSPR.

Linear free energy relationships are a classical field in chemistry that experiences a renaissance. Predicting the properties of solvents. Prediction of reactivity.

Catalysis: heterogeneous catalysis, homogeneous catalysis, electrocatalysis.

Materials science.

Organic synthesis. Predicting new reactions. Prediction of synthesis pathways.

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.).

Class 2 (lecture with interactive demonstrations in Jupyter Notebook)

Rational Design Tools 1: Python 3 and Popular Specialized Libraries

Python 3. 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. Possibilities and features. Installation in Anaconda.

Scikit-learn, Pandas, Matplotlib. Possibilities, features, installation.

Class 3 (lecture with interactive demonstrations in Jupyter Notebook)

Rational Design Tools 2: Specialized Python Libraries and Other Software

PyTorch/TensorFlow-Keras. A brief description of features and comparison. Installation.

Visualizing molecules with free software. Avogadro and Jmol.

Specialized Python libraries for chemoinformatics. RDKit, Mordred, OpenBabel, and others.

Computational chemistry methods: links to reviews and textbooks. The role of quantum chemical methods in rational design. Advantages and disadvantages of quantum chemical methods in application to rational design.

Open/free online resources.

Class 4 (lecture with interactive demonstration)

Rational Design Tools 3: Data Types, Their Representation, and Chemical Databases

Structure of molecules and materials: structural formulae, molecules in Cartesian coordinates, crystal structure.

Data recording 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.

Chemical databases. SciFinder and Reaxys. ACS Solvent Selection Tool. The Materials Project.

Class 5 (lecture with interactive demonstration in Jupyter Notebook)

Brief Discussion of Machine Learning and Data Analysis Algorithms With Examples

Machine learning tasks: regression, classification, clustering. Linear and logistic regression. Principal component method in data analysis (principal component analysis). Algorithm of k-nearest neighbors and support vector machines. Decision tree and random forest. Artificial neural networks.

Class 6 (lecture with interactive demonstrations in Jupyter Notebook)

Linear Regression Method and Free Energy Scaling Relationships

Equations of Hammet, Taft, and others. The substituent effect as a descriptor. Abraham's equation. The works of Matthew Sigman et al. as an example of the modern application of (linear) free energy scaling relationships in chemistry. Implementing linear regression with Jupyter Notebook.

Class 7 (lecture)

Chemical Structure and Structural Descriptors

Recap: structure of molecules and materials. Structural formulae, molecules in Cartesian coordinates, crystals.

Descriptors as mathematical objects: scalars, vectors, matrices, tensors.

Descriptors for organic and inorganic substances.

Descriptors of structure and properties of chemical systems.

Class 8 (workshop)

Representation of Chemical Structures With the Methods of Chemoinformatics

Creation of three-dimensional models of organic substances in the Avogadro editor: simple hydrocarbons, amino acids and sugars, caffeine, and ascorbic acid.

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.

Class 9 (workshop)

Computation of Structural Descriptors and Properties of Molecules

Computation and analysis of molecular features (descriptors) using Pandas, OpenBabel, RDKit, Mordred, and other libraries.

Class 10 (lecture)

Rational Design in Medicinal Chemistry

QSAR/QSPR concept, history of the approach.

Target properties. Biological activity *in vitro* and *in vivo*. Mechanisms of biological activity. Key-lock model, non-covalent and covalent bonding. Chemical activity. Solubility, logP. Lipinski's rule of five. Compound taste prediction.

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

Class 11 (workshop)

Design of Substances With Target Properties in Medicinal Chemistry

Case studies from QsarDB.

Solubility estimation with linear regression using Jupyter Notebook, Numpy, Scikit-learn, Pandas, and Matplotlib.

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.

Session 12 (lecture)

Design of Compounds with Target Properties in Materials Science and Fundamental Problems of Machine Learning

Descriptors of materials structure and target properties.

Literature review of the use of machine learning in the design of new materials (review of reviews at the factual level and the principle of modeling) [Ref1, Ref2].

The Materials Project, an open data set for creation of proof-of-concept models.

Challenges of machine learning. Underfitting and overfitting. "Garbage-in-garbage-out." A discussion of a relevant review.

The problem of interpretability of machine learning and neural network models.

Class 13 (lecture)

Rational Design of Catalysts

What is a catalyst? What is catalytic activity? Examples of catalysts: homogeneous and heterogeneous catalysis.

Descriptors of catalyst structure and target properties of catalysts. Activity vs. stability of a catalyst.

A review of the literature examples of the use of machine learning in the design of high-performance catalysts: facts and illustrations [Ref1, Ref2].

Application of Artificial Neural Networks for Catalysis: A Review.

Class 14 (workshop)

Rational Design of Catalysts

Regression task with classical machine learning and neural network methods to correlate activity of homogeneous catalysts with their structure.

Class 15 (lecture)

Organic Synthesis and Rational Design. Robots in Chemistry

Representation of organic chemical reactions in the specialized literature. Geometric progression and multistep organic synthesis: yield of the target substance. Multistage syntheses of pharmaceuticals.

Artificial intelligence tasks in organic synthesis: prediction of new chemical reactions, prediction of synthetic pathways, and prediction of optimal reaction conditions.

Data quality problems in organic synthesis.

Prediction of optimal chemical reaction conditions: a case study.

The concept of chemputer by Leroy Cronin.

Class 16 (lecture with interactive demonstration)

Organic Synthesis and Rational Design. Neural Networks in the Design of Antibiotics and Anticancer Drugs

Prediction of chemical reactions using artificial intelligence. Development of the IBM RXN for Chemistry system [Ref1, Ref2].

Demonstration of IBM RXN: reaction and synthesis prediction.

A Deep Learning Approach to Antibiotic Discovery.

Illustrating the concept of adversarial training: Reinforced Adversarial Neural Computer for de Novo Molecular Design.

Class 17 (lecture).

The problem of protein folding and neural networks

Proteins, their role in the body, and their structure. The dream of de novo protein design. How small changes in structure (amino acid sequence) lead to significant changes in properties: insulin analogs.

AlphaFold and AlphaFold 2. General remarks [Ref1, Ref2]. An analysis of the architecture.

Class 18 (exam, 2 academic hours; topic: Machine Learning)

A student must prepare a brief presentation of two to three slides (title slide not included) on a randomly selected topic from the list (see supplementary materials).

Class 19 (exam, 2 academic hours; topic: Chemoinformatics)

Two randomly selected questions from the list of test questions (see supplementary materials) must be answered. The student is expected to answer briefly (7-8 minutes), without preparation.