Machine Learning for Chemistry & Drug Design

IIIT, HYDERABAD

HUBDATA

CCNSB

U. Deva Priyakumar

@deva_priyakumar

Who can participate?

- Students, Researchers, Professionals with background in sciences who wish to understand methods of AI/ML in chemistry and biology.
- Students and professionals with background in computer science who wish to learn applications of AI/ML in natural sciences.

Prerequisites for tutorials

An exposure to any programming language (or a keen interest to learn programming skills) would be highly beneficial.

contact.ml4chem@ihub-data.iiit.ac.in

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Machine Learning for Chemistry & Drug Design

"Learning By Doing"

With AI/ML revolutionizing the field of Sciences, there is an increasing need for professionals having knowledge in both the domains. iHub-Data in collaboration with IIIT-H is offering a unique course on "Machine Learning for Chemistry", with an emphasis on drug discovery. The course features theoretical lectures by eminent faculty from the fields of Computer Science and Natural Sciences. It also includes a programming tutorial component to help develop practical skills. The course is ideally suited for students and researchers who may want to develop interdisciplinary skills in solving computationally intensive problems involving natural sciences.

HUBDATA

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

@IHUB_Data

Instructors

Dr. Maitreya Maiti

Dr. Girish Varma

Dr. Charu Sharma

Prof. C. V. Jawahar

Prof. S. Bapi Raju

Prof. U. Deva Priyakumar

People You Should Know :-)

Prof. C. K. Raju

Siddhartha Laghuvarapu

Sarvesh Mehta

Sriram, Bhuvanesh, Arihant,…

Course Format

- 1.5 hr. Lectures every Wednesday: 5:00 to 6:30 PM IST
  - Assumes no background in Statistics and ML
  - Will discuss ML methods for qualitative understanding
  - First step towards ML applications in molecular science
- 3 hr. Tutorials/Practicals every Friday: 5:00 to 8:00 PM IST
  - Focus on ‘Learning by Doing’
  - Understand the given codes -> practice and implement
  - Not meant for solving one’s specific research problem
- Evaluations: Quiz & Performance in tutorials/practicals
  - Cash prize for top performers; Research/Visiting Fellowships;…

Know Your Peers

Scientific Research

Typical Computations

Machine Learning

Rise of the Machines!

Schleder et al., *J. Phys.: Mater*. **2019**

J. Chem. Sci. **2022** - Feature article on ML4Chemistry

Short Quiz!

@deva_priyakumar

COMPUTERS AND BIOMEDICAL RESEARCH 6, 411–421 (1973)

Cybernetic Methods of Drug Design. I. Statement of the Problem—The Perceptron Approach

S. A. HILLER, V. E. GOLENDER, AND A. B. ROSENBLIT

FIG. 1. A three-layer perceptron.

TABLE 1

ANTICONVULSION ACTIVITY OF 1,3-DIOXANES

No.	R_1	R_2	R_3	Isomer	Activity (antagonism to corasol)
1	C_2H_5	H	iso-C_3H_7	trans	+I
2				cys	+I
3	C_3H_7	H	CH_3	trans	–I
4				cys	–I
5	C_3H_7	H	iso-C_3H_7	trans	+I
6				cys	–I
7	C_6H_{13}	H	CH_3	trans	+I

Learning set	Learning		Test	
	Reliability of recognition	Confidence interval	Reliability of recognition	Confidence interval
$n_1 = 22$	86	7	68	10
$n_2 = 24$	89	6	71	13
$n_3 = 26$	85	5	76	9

Computational Chemistry

Type II. Limited Understanding

Type I. Limited by Resources

Type III. Not Straightforward

Type IV. Novel Problem Statements

Type I. Limited Resources

Eg. Molecular Quantum Mechanics

Quantum Mechanics

Molecules = boxes filled with electrons

Electrons form standing waves in these boxes.

$$\psi\,(x, y, z)$$

The wavefunction contains ALL information about the molecule.

$$\mathcal{H}\,\psi \;=\; E\,\psi$$

Schödinger equation

Simplest Box

Particle in a
One-dimensional box

$$\hat{H}\psi = E\psi$$

$$\hat{H} = -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}$$

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = E\psi$$

$$\psi_n(x) = \left(\frac{2}{L}\right)^{1/2} \sin\left(\frac{n\pi x}{L}\right)$$

$$E_n = \frac{n^2h^2}{8mL^2} \qquad n = 1, 2, \ldots$$

Particle in a
One-dimensional box

$$\psi_n(x) = \left(\frac{2}{L}\right)^{1/2} \sin\left(\frac{n\pi x}{L}\right)$$

Boxes come in all kinds of shapes and sizes!

Particle in a Two-dimensional box

$$\psi_n(x) = \left(\frac{2}{L}\right)^{1/2} \sin\left(\frac{n\pi x}{L}\right)$$

$$E_n = \frac{n^2 h^2}{8mL^2} \qquad n = 1, 2, \ldots$$

$$\psi_{n_1,n_2}(x,y) = \frac{2}{(L_1 L_2)^{1/2}} \sin\frac{n_1 \pi x}{L_1} \sin\frac{n_2 \pi y}{L_2}$$

$$E_{n_1,n_2} = \left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2}\right)\frac{h^2}{8m}$$

Hydrogen Atom

Energy Levels

1D box

2D box

Hydrogen atom

Molecular Boxes

Schrödinger eq. -> solve - approximations -> wavefunction -> molecular properties

QM Methods

Hartree-Fock (HF)

Moller-Plesset Perturbation Method
(MP2, MP3, MP4,…)

Configuration Interaction
(CISD, FCI,…)

Coupled Cluster Method
(CCSD, CCSD(T),…)

Multiconfigurational SCF
(CASSCF, CASPT2)

Semi-empirical
methods

AM1
MNDO
PM3
.
.
.

Density Functional Theory: 100s of Methods

Limitations

- Large computational time for high levels of theory even for molecules with few atoms
- Cannot handle large molecules unless severe approximations are made (eg. semi-empirical QM methods)
- Hence limited to small molecules
- Dynamics of molecules (eg. solvated drug-like molecules)
- Proteins/nucleic acids out of scope for the most part

Morawietz & Artrith, JAMD, 2021

BAND-NN

Predicting accurate DFT energies of small organic molecules

J. Comp. Chem. **2020**, *41*, 790

J. Comp. Chem. **2022**, *xxxx*

Type I Problems are Abundant in Chemistry

- Energies
- Other molecular properties
- Protein-Ligand binding free energies
- Solvation free energies
- ADME properties of drug-like molecules
- Protein-protein interactions
- Protein stability due to mutations, different environments…
- Spectroscopic properties
- …

MoleculeNet

Quantum Mechanics

- QM7/QM7b: Electronic properties(atomization energy, HOMO/LUMO, etc.) determined using *ab-initio* density functional theory(DFT).
- QM8: Electronic spectra and excited state energy of small molecules calculated by multiple quantum mechanic methods.
- QM9: Geometric, energetic, electronic and thermodynamic properties of DFT-modeled small molecules.

Physical Chemistry

- ESOL: Water solubility data(log solubility in mols per litre) for common organic small molecules.
- FreeSolv: Experimental and calculated hydration free energy of small molecules in water.
- Lipophilicity: Experimental results of octanol/water distribution coefficient(logD at pH 7.4).

Biophysics

- PCBA: Selected from PubChem BioAssay, consisting of measured biological activities of small molecules generated by high-throughput screening.
- MUV: Subset of PubChem BioAssay by applying a refined nearest neighbor analysis, designed for validation of virtual screening techniques.
- HIV: Experimentally measured abilities to inhibit HIV replication.
- PDBbind: Binding affinities for bio-molecular complexes, both structures of proteins and ligands are provided.
- BACE: Quantitative (IC50) and qualitative (binary label) binding results for a set of inhibitors of human β-secretase 1(BACE-1).

Physiology

- BBBP: Binary labels of blood-brain barrier penetration(permeability).
- Tox21: Qualitative toxicity measurements on 12 biological targets, including nuclear receptors and stress response pathways.
- ToxCast: Toxicology data for a large library of compounds based on *in vitro* high-throughput screening, including experiments on over 600 tasks.
- SIDER: Database of marketed drugs and adverse drug reactions (ADR), grouped into 27 system organ classes.
- ClinTox: Qualitative data of drugs approved by the FDA and those that have failed clinical trials for toxicity reasons.

Type II. Limited Understanding

Eg. Proteins: Sequence to Structure

Protein Folding Problem
Alphafold

Protein Sequence
SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC
Neural Network
Databases
Distance Predictions
Angle Predictions
Score
Gradient Descent
Structure
deepmind.com

Won CASP competition - predicted 25/43

deepmind.com

Performance of Alphafold & Alphafold2

deepmind.com

Other Type II Problems?

Type III. Not Straightforward Problems
Type IV. Novel/New Problem Statements

Type IV

Type III

Identify Molecules with Desired Properties

Best Synthetic Route?

AI Enabled Robotic Synthesis

Automated Characterization

Type III

Inverse Problems

Molecular Generation

- Generate molecules de novo (SMILES)
- Target Properties:
  - Log P value in the range from 0 to 5 (drug like)
  - Easily synthesizable (empirical synthetic accessibility score)

Ertl et al. *J. Cheminf*. **2009**, 8

Molecular Generation (reinforcement learning)

NO ENTRY
ENTER
NO ENTRY
d
lo
Loss
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
6
8
Predicted logP values

Easy
Hard
Optimized
Unbiased
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
Predicted SA Scores

Organic Retrosynthesis

Retrosynthesis

Target
(1)
(2)
(3)
(4)
(5)
(6)

Other Type III & IV Problems?

Scientific Research

What to expect from the course?

- What is Machine Learning (supervised methods, training practices, metrics)
- ML for Chemistry & drug design: Problems & existing solutions
- Molecular Featurization
- Deep Learning
  - NN, back propagation, loss functions, hyper parameters
  - CNN, RNN, LSTM, GNN
  - RL, Autoencoders
  - Case studies on DL for Chem (prediction, classification, generation,…)

1.5 hr Lecture every week

3 hr Tutorials/Practicals

Summary

- Introduction to the course
- Where does ML methods fit in the “Experiment - Theory - Computation” paradigm of Scientific Research?
    - ML useful not only for solving problems for which solutions exist (Eg. predicting molecular properties)
    - but also problems that are typically not addressed using conventional methods
        - Exploration of chemical space - eg. molecular generators
        - Predicting multi-step retro synthetic pathways
        - Predicting structure from spectra
    - Expectations from the course

Questions?

Added after the lecture

Further Reading

- Artificial intelligence: machine learning for chemical sciences: https://link.springer.com/article/10.1007/s12039-021-01995-2

- Best practices in machine learning for chemistry: https://www.nature.com/articles/s41557-021-00716-z

If you are not able to access to any of these, please reach out to one of the GAs

I am currently enrolled in the following program:

Mentimeter

110

Which state in India are you from currently? If you are from outside India, please type your country name.

Mentimeter

114

When (do you think) was the first paper on ML applied to a chemistry/drug design task published?

Mentimeter

114

If this was 50 years ago, Why is it only in the last 5 years there is a surge in the activity in this area?

Mentimeter

106