Syllabus of BIOINF 595 (2025 Winter, Bioinformatics Program)

Course Name:

Machine Learning for Drug Discovery

Course Description:

The role of machine learning in molecular modeling and drug discovery is rapidly growing. This course introduces fundamental concepts and methods of chemoinformatics, structural biology, and machine learning. Topics covered include molecular representations, physics based protein modeling, ligand docking, chemical conformation generation, virtual screening, bioactivity data modeling, chemical data curation, drug discovery pipeline, machine learning bioactivity prediction, deep-learning architectures for molecular systems. We will discuss recent breakthroughs in applying scientific foundation models and generalizability of predictions for drug discovery. Emphasis of the classes is in understanding fundamental concepts in bioinformatics and practical application of data science tools and methods to problems in medicinal chemistry and their own research. By the end of the course students will be able to use ML models to tackle challenges in molecular modeling and drug discovery, including defining objectives as quantifiable tasks, curating chemical, structural, and bioactivity data, and appling state of the art machine learning methods.

Instructor

Matthew O'Meara, Email: maom@umich.edu
Terra Sztain-Pedone, Email: tsztain@umich.edu

Schedule and location:

Tuesday and Friday 11:30a - 1:00p
Room MedSci II 2813

Cross-listing

This course will be cross-listed with MEDCHEM

Textbook:

No textbook is required for this course. All teaching materials will be posted on the course website.

Prerequisites:

Familiarity with linear algebra and basic programming (BIOINF-575 or equivalent), or approval by instructor

Presentation, homework, & grades:

There will be homework assignments, including code writing and literature reading and presentation. Final grade consists of class participation (10%), homework/labs (30%), paper presentation (20%), and a final project. Key assignments will include reading and analyzing published papers and data sets, curating bioactivity data sets, and training structure and activity prediction models.

Paper Presentation Scoring:
Criteria each is scored 1-5 and the score is summed to 20%

- Clear statement of the problem
- Clear summary of the results and methods
- Clear description of the conclusions/limitations of the work
- Answering questions about the work

Final Project Scoring:
Criteria each is scored 1-5 and the score is summed to 40%

- Clear statement of the problem
- Methods detailed enough to be reproduced without any additional information
- Clear description of the results/conclusions/limitations of the work
- Informative figures conveying results

Format

- 3 pages not including references
- 11 pt font Arial (10 pt for figure legends)
- Sections: Abstract, Intro, Methods, Results, Conclusion, References

Schedule:

Day Number	Topic	By	Details	Format
Fri - 1/10/25	Course Intro	O'Meara + Sztain	Drug discovery pipeline	Lecture
Tue - 1/14/25	CADD intro	Sztain	Introduction to CADD Tasks	Lecture
Fri - 1/17/25	CADD intro Lab	Sztain	ChEMBL, ML intro	Lab
Tue - 1/21/25	Protein structure /	Sztain	Protein folding, MD	Lecture

	dynamics			
Fri - 1/24/25	MD tutorial	Sztain	openMM, MDtraj	Lab
Tue - 1/28/25	ML forcefields	Sztain	Learning QM, MM, CG forcefields	Lecture
Fri - 1/31/25	Bottom-up coarse graining tutorial	Sztain	CGSchnet, force matching	Lab
Tue - 2/4/25	Data science intro	O'Meara	Tidy data / grammar of graphics	Lecture
Fri - 2/7/25	Molecular structure data	O'Meara	Structure determination and databases	Lab
Tue - 2/11/25	Machine Learning	O'Meara	Supervised learning	Lecture
Fri - 2/14/25	Ligand-based virtual screening	O'Meara	Molecular fingerprints / chemical spaces	Lab
Tue - 2/18/25	Structure-based virtual screening	O'Meara	Physics based pose prediction / docking	Lecture
Fri - 2/21/25	Rosetta GALigandDock tutorial	O'Meara	Pose prediction vs. virtual screening	Lab
Tue - 2/25/25	Bioactivity data	O'Meara	Biophysics of bioactivity and assay interference	Lecture
Fri - 2/28/25	Bioactivity prediction lab	O'Meara	Apply ML method for bioactivity task	Lab
Tue - 3/4/25	break			
Fri - 3/7/25	break			
Tue - 3/11/25	Unsupervised ML	O'Meara	Clustering, embeddings, and explainable models	Lecture
Fri - 3/14/25	Representation learning tutorial	O'Meara	Dataset comparison and exploratory analysis of VS Hits	Lab
Tue - 3/18/25	Chemical synthesis data	O'Meara	Chemical properties and synthesis	Lecture
Fri - 3/21/25	Generative AI for Ligands	O'Meara	AI Ligand generation	Lab
Tue - 3/25/25	Protein language models	O'Meara	Sequence space and conditional sequence generation	Lecture

Fri - 3/28/25	Generative AI for Proteins	O'Meara	AI Protein structure prediction and design	Lab
Tue - 4/1/25	ML + MD enhanced sampling	Sztain	ML features, progress coordinates	Lecture
Fri - 4/4/25	Rave tutorial	Sztain	ML progress coordinates enhanced sampling	Lab
Tue - 4/8/25	Free energy methods	Sztain	Estimation of free-energy from MD	Lecture
Fri - 4/11/25	Free energy methods	Sztain	Estimation of free-energy from MD	Lab
Tue - 4/15/25	Pathway methods	Sztain	Weighted ensemble	Lecture
Fri - 4/18/25	Guest lecture			Lecture
Tue - 4/22/25	Guest lecture			Lecture

Topics covered:

Introduction to Computer-aided Drug Design

1. Drug discovery pipeline through the lens of computational modeling

Drug Discovery Data Modalities

1. Tidy data
    a. Manipulating data frames
    b. Grammar of graphics
    c. Organizing computational analyses
2. Chemical data
    a. Biophysics
        i. Protonation, stereochemistry, salt forms, aromaticity, purity
        ii. Experimental characterization
            1. Crystallography
            2. NMR
            3. MassSpectrometry
    b. File formats
        i. Small molecule formats: SMILES, SMARTS, InChi, mol2, sdf
    c. Machine Learning representations
        i. Fingerprints
        ii. Graphs
        iii. 3D coordinates
    d. Data sources
        i. Synthesis enumeration
        ii. Commercially available compounds, make-on-demand chemistry

iii. Databases: PubChem, ChEMBL, SureChEMBL

3. Protein structure data
    a. Biophysics
        i. Central dogma (DNA, RNA, Protein, post-translational processing)
        ii. Metastable states, kinetics, intrinsically disordered
        iii. Experimental characterization
            1. High-resolution
                a. X-ray crystallography
                b. CryoEM
                c. NMR
            2. Low-resolution
                a. CD
                b. ITC
                c. FRET
    b. File formats
        i. PDB, mmCIF, MD formats
    c. Machine Learning representations
        i. 3D coordinates
        ii. Internal degrees of freedom (Backbone, sidechain)
        iii. Frames
        iv. van der Mer
        v. Contact maps
        vi. Multiple sequence alignments
4. Bioactivity data
    a. Biophysics of bioactivity
        i. Dose-response models
        ii. Binding vs. functional outcomes
        iii. Coupled functional assays
        iv. Growth assays
        v. Kinetic models
        vi. Pooled assays
    b. Confounding effects
        i. In vitro vs in cell
        ii. Toxicity
        iii. Polypharmacology
        iv. Measured and unmeasured experimental measurement bias
    c. Experimental screening
        i. High-throughput Screening
        ii. DNA-encoded libraries
    d. Bioactivity databases
        i. ChEMBL, PubChem
        ii. IC50/EC50 vs Ki

iii. Uncertainty quantification
iv. ML Taskification
    1. Papers with Code
    2. HuggingFace
    3. Therapeutic Data Commons

Machine Learning

1. Dimensionality reduction
    a. Linear embedding
        i. Matrices and geometric rank
        ii. PCA
    b. Non-linear embedding
        i. t-SNE, UMAP,
        ii. Global-local trade-off
        iii. Quantifying distortion
2. Gradient boosted decision trees
    a. Quantifying prediction quality
    b. Generalization and cross-validation
    c. Hyper-parameter optimization and autoML
3. Deep learning principles
    a. Backpropagation
    b. Stochastic gradient descent
        i. Gradient estimators
        ii. Stochastic networks
    c. Training dynamics
        i. double descent
        ii. memorization vs generalization
        iii. scaling laws
    d. AI software stack
        i. GPU/TPUs
        ii. Low-level matrix math
        iii. Pytorch
        iv. Code, data, and model management
        v. High-performance computing
    e. Deep learning architectures
        i. Graph convolution neural networks
        ii. Transformer
        iii. Normalizing flows
            1. Denoising/diffusion
            2. Flow matching
        iv. Equivariance and symmetry
            1. SE(3)
            2. Permutation

Physics based molecular modeling

1. Degrees of freedom
    a. Sampling vs. scoring
    b. protein, ligand DOFs
2. Molecular mechanics force fields
    a. Energy components
        i. Torsional potentials
        ii. Van der Waals
        iii. Electrostatics
        iv. Desolvation
        v. Hydrogen bonding
    b. Case study Forcefields
        i. Rosetta
        ii. Amber
        iii. UCSF Dock
3. Sampling
    a. Molecular dynamics
    b. Monte-carlo markov chain
    c. Gradient descent
    d. Genetic algorithm
4. Ligand docking
    a. Pose prediction task
    b. Challenges
        i. Site flexibility
        ii. Explicit solvation
        iii. Co-factors
        iv. Protonation states
5. Virtual screening
    a. Drug discovery pipeline
        i. Screening
        ii. Lead optimization
        iii. Pre-clinical animal studies
        iv. Clinical trials phases
    b. Make-on-demand chemical libraries
    c. Docking model optimization
        i. retrospective decoy-discrimination
        ii. physics based hyperparameter optimization
    d. High-throughput virtual screening
        i. Unbiased comprehensive screening
        ii. Iterative screening
    e. Hit-picking

        i. Automated filtering
        ii. Human filtering
    f. Testing predictions
        i. Sourcing compounds and cost
        ii. Biochemical testing
6. Molecular dynamics
    a. Representation
        i. chemical degrees of freedom
        ii. water box
        iii. force-fields
    b. Simulation
        i. Initialization and equilibration
        ii. Thermostats
    c. Analysis
        i. state-clustering
        ii. reaction coordinates
        iii. distance vs. fluctuation analysis
    d. Enhanced sampling
        i. weighted ensemble simulation (WESPA)
        ii. parallel tempering
    e. Free energy estimation
        i. Free-energy perturbation

Machine learning for drug discovery

1. Compound based property prediction
    a. Discrete vs. continuous outcomes
    b. Out of distribution prediction / generalization
2. Conditional compound generation
    a. Quantification of quality
        i. Synthetic accessibility
        ii. Diversity
        iii. Task-specific scores
    b. Sourcing
        i. Analog-by-catalog
        ii. Retrosynthesis
3. Pose prediction
    a. Generation vs. scoring
    b. Multi-modality representations
4. Dynamical ensemble generation
    a. State space models
    b. Boltzmann generators