SYLLABUS
"COMPUTATIONAL MOLECULAR DESIGN WITH MACHINE LEARNING"

National Academy of Sciences of Ukraine

Yukhnovskii Institute for Condensed Matter Physics

Specialty: 104 Physics and astronomy (E5 Physics and astronomy)

Course lecturer:
Dr. Maksym Druchok
maksym@icmp.lviv.ua
ICMP, Svientsitskii St.1, Lviv, Ukraine

Course description

This course offers a focused introduction to modern machine learning techniques used in computational molecular design. It emphasizes methods for predicting the properties of small organic molecules and generating novel molecular structures with specific desired characteristics. PhD students will learn how to represent molecules using formats like SMILES, compute chemical descriptors, and apply ML for property prediction. The course also explores generative models for autonomous molecule generation and optimization. Each topic is reinforced through hands-on coding exercises in Python and real-world datasets. No prior experience in machine learning or chemistry is required, though basic familiarity with programming will be beneficial.

Course goals

The general goal for students is to learn about models and methods currently employed in computational studies of small organic molecules, with a focus on predicting their properties and generating new molecules using machine learning. This will be achieved through a combination of classroom lectures and practical sessions. To reinforce the acquired skills and concepts, theoretical ideas will be introduced alongside practical applications using real molecular datasets. By the end of this course, students are expected to learn how to:

- represent and manipulate small organic molecules computationally using cheminformatics tools,
- apply classical and neural network-based machine learning models to predict molecular properties,
- use generative models to sample chemically valid molecules,
- implement generative strategies for property-driven molecule generation.

Course breakdown structure

Components of the course	Total hours
Number of credits/hours	3/90
All classes, including:	48
● lectures, hrs	16
● practical sessions, hrs	16
● seminars, hrs	16
Self-work, including:	42
● individual scientific ● study, hrs	18
● preparation to lectures, practical sessions, seminars, and exam, hrs	24
Exam	1

Percentage of class work – 53.3%

Lectures

	Topic	Hours
1	**Introduction and python for molecular informatics.** Overview of the course: goals and applications in molecular discovery. Python essentials: data types, control flow, functions, libraries. Introduction to cheminformatics. Molecular structure representations: SMILES, InChI, and molecular graphs. Canonical SMILES, handling duplicates. Installing and using RDKit. Computing descriptors and fingerprints with RDKit.	4
2	**Machine learning for molecular property prediction.** Overview of supervised learning: regression and classification. Train/test splits, cross-validation, metrics. Feature engineering with molecular descriptors. Hyperparameter tuning. Classical ML. Neural networks: architecture, activation functions, loss functions. Molecules as graphs.	4
3	**Generative modeling overview and SMILES-based generation.** What is generative modeling? Applications in drug discovery. Introduction to RNNs/LSTMs for SMILES generation. Overview of model types: VAEs, GANs, etc. SMILES generation as a language modeling problem. Sampling techniques	4

	and decoding issues. Evaluation: validity, uniqueness, novelty of generated molecules.	
4	**Reinforcement Learning and Monte Carlo Tree Search for molecular design.** Motivation: optimizing molecules for desired properties. Basics of reinforcement learning: states, actions, rewards. Overview of Monte Carlo Tree Search (MCTS). Applying MCTS to SMILES or graph expansion. Multi-objective optimization. Case study: property-optimized molecule generation.	4
	Total hours	16

In-class practical sessions

	Topic	Hours
1	**Preparing data and extracting molecular descriptors with RDKit.** Acquire training datasets. Cleaning the data. Compute molecular descriptors and fingerprints using RDKit.	4
2	**Property prediction with Scikit-learn.** Build a data feeding pipeline for the training. Train and evaluate classical ML models. Tune hyperparameters and monitor training/validation loss. Interpret model performance metrics.	4
3	**Building a neural network for molecular property prediction.** Construct a feedforward neural network using PyTorch. Use molecular descriptors or fingerprints as input.	4
4	**Training a neural network for molecular property prediction.** Tune hyperparameters and monitor training/validation loss. Run predictions. Visualize the inference results.	4
	Total hours	16

Seminars

	Topic	Hours
1	Data quality and bias in molecular machine learning.	4
2	Molecular representations beyond SMILES.	4

3	State-of-the-art studies in prediction of molecular properties.	4
4	Generative models for molecules: A survey of approaches.	4
	Total hours	16

Grading system explained

Max grades in points					
Semester			**Exam**		**Total**
Lecture engagement	**In-class practical sessions**	**Seminars**	**Written component**	**Oral component**	
5	35	20	0	40	100

Grade cut-offs

88% A
80% B
70% C

Recommended reading materials for self-study

1. Ramsundar, B., Eastman, P., et al. **Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More.** O'Reilly Media, Inc. (2019)
2. Goh G.B. et al. **Deep Learning in Computational Chemistry.** J. Comput. Chem. 38 1291-1307 (2017)
3. Gomez-Bombarelly R. et al. **Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules.** ACS Cent. Sci. 4, 268-276 (2018)
4. Hagg A., Kirschner K.N. **Open-Source Machine Learning in Computational Chemistry.** J. Chem. Inf. Model. 63, 4505 (2023)
5. Tang X. et al **A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation.** Brief. Inform. 25(4), bbae338 (2024)
6. Crucitti D. et al. **De novo drug design through artificial intelligence: an introduction.** Front. Hematol. 3. 1305741 (2024)