GDAŃSK UNIVERSITY OF TECHNOLOGY

Subject card

Subject name and code	CHEMICAL INFORMATICS, PG_00070339		
Field of study	InfoBioChem		
Date of commencement of studies	February 2026	Academic year of realisation of subject	2025/2026
Education level	second-cycle studies	Subject group	Obligatory subject group in the field of study Subject group related to scientific research in the field of study
Mode of study	Full-time studies	Mode of delivery	blended-learning
Year of study	1	Language of instruction	Polish Polish language
Semester of study	1	ECTS credits	3.0
Learning profile	general academic profile	Assessment form	assessment
Conducting unit	Department of Physical Chemistry -> Faculty of Chemistry -> Faculties of Gdańsk University of Technology		
Name and surname of lecturer (lecturers)	Subject supervisor	dr hab. Agnieszka Gajewicz-Skrętna	
	Teachers	dr hab. Agnieszka Gajewicz-Skrętna dr inż. Miłosz Wieczór	

Lesson types

Lesson type	Lecture	Tutorial	Laboratory	Project	Seminar	SUM
Number of study hours	30.0	0.0	30.0	0.0	0.0	60

E-learning hours included: 12.0

eNauczanie source addresses:

Moodle ID: 4993 CHEMOINFORMATYKA
https://enauczanie.pg.edu.pl/2025/course/view.php?id=4993

Learning activity and number of study hours

Learning activity	Participation in didactic classes included in study plan	Participation in consultation hours	Self-study	SUM
Number of study hours	60	5.0	10.0	75

Subject objectives

This course covers fundamental chemometric concepts, including methods for analyzing experimental data in chemical, biological, and biotechnological research. Students will learn the principles of experimental design, as well as how to process, analyze, and interpret measurement data. The course also covers multivariate analysis, dimensionality reduction, and model validation, providing examples of chemometric applications in research and industry.

Learning outcomes

Course outcome	Subject outcome	Method of verification
[K7_W01] knows and understands the methods, techniques, and tools used to solve bioinformatics tasks, including molecular modeling.	The student demonstrates the mathematical knowledge necessary for properly describing datasets statistically. They also possess the physics knowledge required to analyze data related to technical and technological problems.	[SW1] Assessment of factual knowledge [SW2] Assessment of knowledge contained in presentation [SW3] Assessment of knowledge contained in written work and projects
[K7_W03] knows and understands methods for statistical analysis of biological phenomena, biotechnological or chemical processes, and specialized bioinformatics tools.	Upon completion of the course, students will understand the theoretical foundations and operating principles of the most important data mining techniques and machine learning methods. Students will also be able to identify and discuss applications of chemoinformatics methods in pharmacy, chemistry, and environmental protection.	[SW1] Assessment of factual knowledge [SW2] Assessment of knowledge contained in presentation [SW3] Assessment of knowledge contained in written work and projects
[K7_U05] is able to select computer modeling methods and apply them to solve problems related to the operation and regulation of complex systems.	Upon completing the course, each student will be able to independently formulate a research problem and select the appropriate data mining techniques and/or machine learning methods to address it. Students can perform data projection, visualization, similarity analysis, and clustering, as well as interpret the results. Students can also develop regression and classification models, select the proper model hyperparameters, and predict the response variable based on a set of explanatory variables. Additionally, students understand the responsibility associated with analytical decision-making, can work independently or as part of a team, and recognize the importance of meeting agreed-upon deadlines.	[SU4] Assessment of ability to use methods and tools [SU5] Assessment of ability to present the results of task [SU3] Assessment of ability to use knowledge gained from the subject

Subject contents

Course content – lecture

The lecture covers the following topics::

- Introduction to chemoinformatics: definitions, historical background, and importance in pharmacology, chemistry, and environmental protection.
- Encoding of chemical structure in chemoinformatics: SMILES notation, molecular fingerprints, and related representation methods.
- Data mining techniques: projection and visualization methods for multivariate data, similarity analysis, grouping methods, and clustering techniques.
- Machine learning methods for classification and regression problems, such as linear regression and its variants (ridge regression, Lasso, elastic net, principal component regression, and partial least squares regression); k-nearest neighbors; classification and regression trees; random forest; support vector machines; and neural networks.
- Contemporary challenges in chemoinformatics: quantity and reliability of input data, the problem of imbalanced classification datasets, lack of transparency of algorithms, etc.
- Examples of chemoinformatics applications in computer-aided drug design (CADD), chemistry, and environmental protection.

Course content – laboratory

The laboratory classes cover the following topics:

- Overview of major open-source chemoinformatics tools, e.g., *ChemmineR; RDKit; KNIME; Enalos+ KNIME nodes; OEChem; CDK; Open Babel*, etc.
- Practical introduction to data projection and visualization techniques, similarity analysis, and grouping/ clustering methods, including correlation matrices, heatmaps, k-means clustering, partitioning around medoids (PAM), hierarchical clustering, and principal component analysis (PCA).
- Practical introduction to regression and classification methods, including linear regression and its variants (ridge regression, LASSO, elastic net, principal component regression, and partial least squares regression), k-nearest neighbors, classification and regression trees, random forest, support vector machines, and related approaches.
- Final project: development and presentation of the results of a chemoinformatics analysis involving the use of selected data mining techniques and machine learning methods to address a defined research problem.

Prerequisites and co-requisites

<table>
<tr><td>Assessment methods and criteria</td><td colspan="2"><table>
<tr><th>Subject passing criteria</th><th>Passing threshold</th><th>Percentage of the final grade</th></tr>
<tr><td>Lecture: written exam consisting of multiple-choice test questions and open-ended questions</td><td>50.0%</td><td>50.0%</td></tr>
<tr><td>Laboratory classes: project-based assignment</td><td>50.0%</td><td>50.0%</td></tr>
</table></td></tr>
<tr><td rowspan="3">Recommended reading</td><td>Basic literature</td><td>• Bąk A, Polański J. Podstawy chemoinformatyki leków. Wydanie drugie rozszerzone. 2018, Wydawnictwo Uniwersytetu Śląskiego, ISBN:9788380128965.
• Morzy T, Eksploracja danych. Metody i algorytmy. 2013, Wydawnictwo Naukowe PWN. ISBN:9788301171759.
• Szeliga M. Data Science i uczenie maszynowe. 2017, Wydawnictwo Naukowe PWN. ISBN:9788301192327.</td></tr>
<tr><td>Supplementary literature</td><td>• Engel T, Gasteiger J. Chemoinformatics: Basic Concepts and Methods (1st Edition). 2018, Wiley-VCH. ISBN:9783527693788.
• Engel T, Gasteiger J. Applied Chemoinformatics: Achievements and Future Opportunities. 2018, Wiley-VCH. ISBN:9783527806546.</td></tr>
<tr><td>eResources addresses</td><td></td></tr>
<tr><td>Example issues/ example questions/ tasks being completed</td><td colspan="2">Example theoretical topics:
• Methods for dimensionality reduction in data analysis.
• What does model hyperparameter optimization mean?
• Differences between supervised and unsupervised learning.
• Differences between regression and classification
• Properties of principal components
• List and discuss the assumptions behind three selected techniques for handling class imbalance in classification.
• What is ADMET analysis?
• Applications of data mining methods in drug design
• What is virtual screening?
• Explain the FAIR data principles in chemoinformatics
Example computational topics:
• Classify the given chemical compounds using a decision tree model and the k-nearest neighbors method
• Compare the performance of the two aforementioned classification methods
• Draw a dendrogram and interpret it using similarity analysis
• Calculate the cumulative percentage of variance explained by the first three principal components
• Encode a chemical molecule in SMILES format
• Based on the descriptive statistics provided for regression models that predict drug absorption rates, rank the models from best to worst performance.</td></tr>
<tr><td>Practical activites within the subject</td><td colspan="2">Not applicable</td></tr>
</table>

Document generated electronically. Does not require a seal or signature.