Comparative Analysis of Drug Compound Representation and Perceived Reactivity: Chemception Embedding versus SMILES Code

Cognitive Systems: Language, Learning, and Reasoning

Students:
Swapnil Jha
Milena Voskanyan

Name of the Module: Project Module 12 ECTS

Name des Moduls: PM 2 - Project in Machine Learning	Anzahl der Leistungspunkte (LP): 12
Modulart (Pflicht- oder Wahl-pflichtmodul):	Wahlpflichtmodul
Inhalte und Qualifikationsziele des Moduls:	*Qualifikationsziele* wie PM 1 *Inhalte* Die Studierenden erarbeiten sich zunächst ein spezialisiertes Gebiet der aktuellen Forschung im Bereich des maschinellen Lernens. Sie erschließen sich dazu die Literatur dieses Gebiets selbstständig und diskutieren Fragen im Seminar. Auf dieser Grundlage definieren Teams von Studierenden dann eigene, inhaltlich klar umgrenzte Forschungs-, Experimental- oder Entwicklungsprojekte. Sie bearbeiten diese Projekte und präsentieren abschließend ihre Ergebnisse. Bei der Auswahl der inhaltlichen Gebiete orientieren sich die Dozentinnen und Dozenten an Forschungsthemen der aktuellen Literatur.

Model Training Methodology

Generation of SMILES Codes

↓

Generation of image of chemical based on SMILES Code using the RDKit Library

↓

Training of Chemception Embedding based CNN Model with image data as input.

↓

Potential Training of further alternative network architectures.

↓

Comparative analysis of model based on SMILES Code, Chemception, and potential alternative architectures.

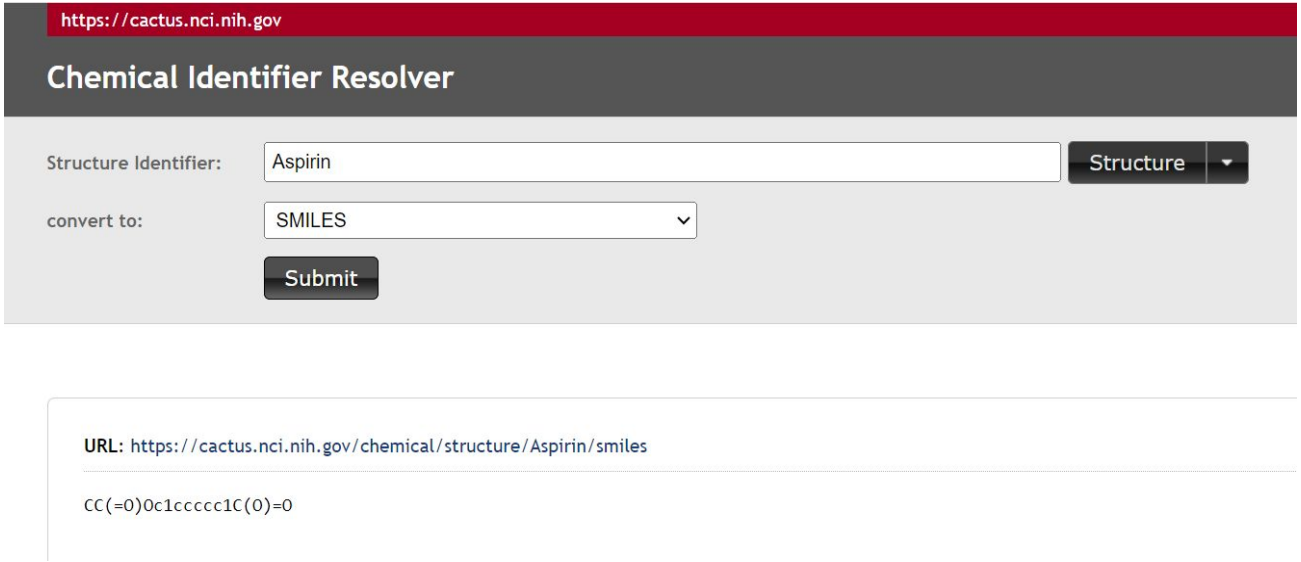
Task List

- Generation of SMILES code for sensitivity prediction.
  - Will take up to 5 part time working days as it will involve a lot of manual data annotation due to non-standardized identifiers used in input data.
  - Will be discussed in further detail in next slides.
- Exploration of Chemception encoding in CNNs.
- Evaluation of model performance using validation metrics such as MAE, R^2, and RMSE.
- Implementation and comparison of alternative network architectures.
- Assessment of different models and their data understanding.
- Compilation of a comprehensive report on model evaluation, optimization, data analysis, and future scope.

SMILES

1. **What is SMILES Code?**
   - SMILES (Simplified Molecular Input Line Entry System) is a notation system that represents a chemical structure in the form of a line of text.
2. **Purpose of SMILES Code**
   - To provide a standardized way to encode molecular information and facilitate the sharing and searching of chemical databases.
3. **Applications of SMILES Code**
   - Used widely in cheminformatics software for drug design and chemical analysis.
   - Essential for virtual screening and compound database management.
4. **Benefits of SMILES Code**
   - Simplifies the representation of complex chemical structures.
   - Enhances computational efficiency in chemical database searches.

SMILES


https://cactus.nci.nih.gov
Chemical Identifier Resolver
Structure Identifier:
Aspirin
Structure
convert to:
SMILES
Submit
URL: https://cactus.nci.nih.gov/chemical/structure/Aspirin/smiles
CC(=O)Oc1ccccc1C(O)=O

https://cactus.nci.nih.gov

Chemical Identifier Resolver

Structure Identifier: ABT737 Structure

convert to: SMILES

Submit

A lot of Identifiers have informal or test names, as a result manual annotation is required, in this instance from https://www.biomol.com/de/produkte/chemikalien/biochemikalien/abt-737-cay11501-1

Similarly, a lot of such entries will need to be manually searched and annotated.

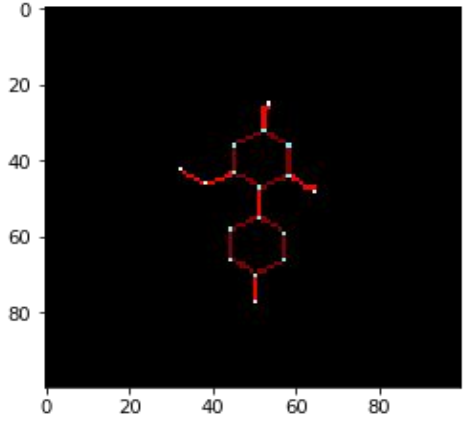
URL:

Sorry, your structure identifier could not be resolved (the request returned a HTML 404 status message)

Chemception

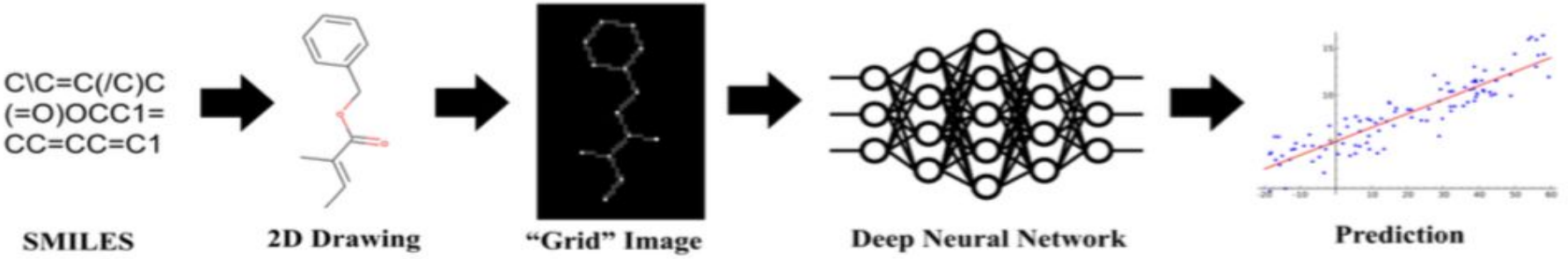
- What is Chemception?
- Purpose of Chemception
- Applications of Chemception
- Benefits of using Chemception

What is Chemception



- A deep learning model adapted from the Inception architecture used in image recognition.
- Designed to analyze and interpret chemical structures through convolutional neural networks.

Purpose of Chemception



- To predict chemical properties, activities, and interactions by processing 2D images of molecular structures.

- Facilitates faster and more accurate chemical analysis compared to traditional methods.

Applications of Chemception

- **Drug discovery**: Predicting the effectiveness and safety of new drugs.
- **Environmental science**: Assessing the toxicity and environmental impact of chemicals.
- **Material science**: Developing new materials with desired properties.

Benefits of Chemception

- Increases the speed and reduces the cost of chemical analysis.
- Enhances the accuracy of predictions, leveraging vast amounts of data.
- Provides a non-invasive method for chemical property prediction.

Validation Metrics

- **MAE** indicating Mean Absolute Error.
- **R^2** quantifies the proportion of variability observed in the dependent variable that can be attributed to the combined influence of the independent variables
- **RMSE** indicates how well a functional curve is fitted to the available data, or how much, on average, a forecast deviates from the (historical) data/actual observations.

References

- https://www.cancerrxgene.org/
- https://cactus.nci.nih.gov/chemical/structure
- https://www.investopedia.com/terms/r/r-squared.asp
- https://h2o.ai/wiki/auc-roc/#:~:text=AUC%2DROC%20is%20a%20performance,between%20positive%20and%20negative%20classes.
- https://de.statista.com/statistik/lexikon/definition/303/root_mean_square_error/
- https://www.cheminformania.com/wp-content/uploads/2017/11/Chemception-Demo_std_11_2.png
- https://depth-first.com/images/posts/20190204/chemception.png
- https://arxiv.org/abs/1509.09292
- https://arxiv.org/abs/1706.06689
- https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.9b00520
- https://github.com/prassepaul/mlmed_transfer_learning
- https://github.com/prassepaul/mlmed_ranking
- https://www.uni-potsdam.de/fileadmin/projects/ambek/Amtliche_Bekanntmachungen/2014/ambek-2014-05-200-216.pdf