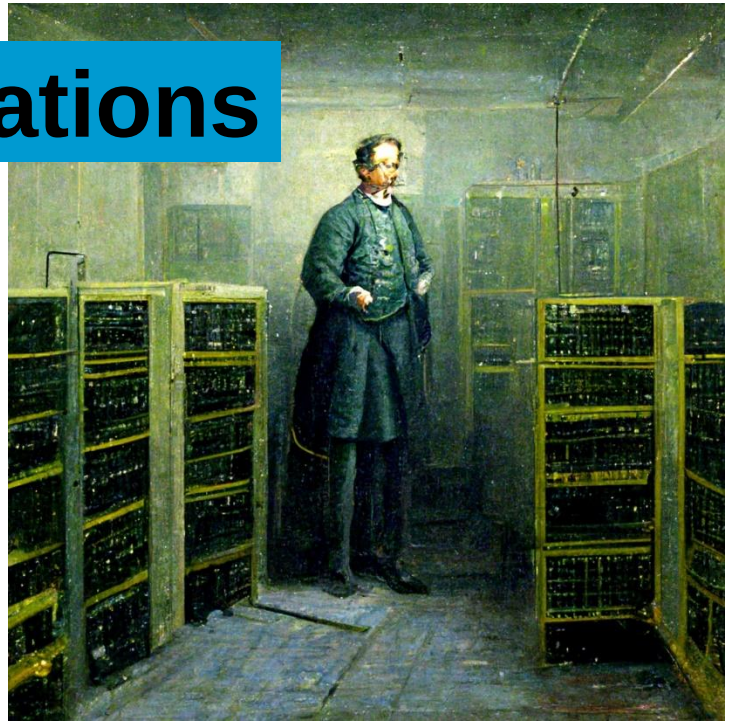
Universität Konstanz

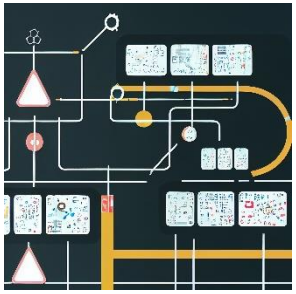
Machine Learning in Organic Chemistry Applications and Limitations



M. Sc. Lukas Holz

26.10.2022

What is Machine Learning and how does it work?



- **Traditional Programming:**
- Input: raw data + program developed by human
- Produces: Output data

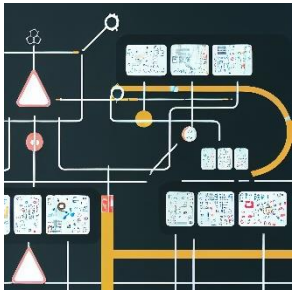
- **Machine learning:**
- Input: raw data + output data
- Produces: program (developed by computer, i.e. machine learned)

{Screenshot showing abstract of Samuel2000, highlighting the phrase „a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program. Contains hyperlink to respective publication}

Image on previous slide courtesy of the Midjourney AI (https://www.midjourney.com/home/) under creative commons license,
Using the prompt: /imagine Friedrich Wohler standing in a server room

Grimson, W. Eric L., “11. Introduction to Machine Learning”, *YouTube*, MIT OpenCourseWare, 19.05.2017, www.youtube.com/watch?v=h0e2HAPTGF4
Samuel, A. L., Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **2000,** *44* (1.2), 206-226.

What is Machine Learning and how does it work?



Typical examples of „learnable“ things

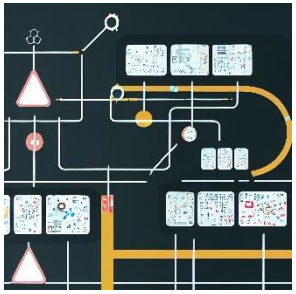
1) Given an image of a pet, ML may help to decide whether it‘s a cat or a dog (image recognition)
2) Build an algorithm that deciphers how handwritten digits look like (see MNIST data set):

{Image showing a sample of the MNIST dataset from Shah2020
Contains hyperlink to respective website}

3) Predict which shows or movies a user might want to watch based on their past viewing behaviour and personal info such as date of birth, region, etc. (e.g. Netflix)

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141–142.
Shah, N., https://medium.com/@niranjanshah474/building-neural-network-from-scratch-for-digit-recognizer-using-mnist-dataset-30397be28f5e, 23.09.2020

What is Machine Learning and how does it work?



Basic process

1) Acquire and import raw data (typically in .csv format) into your programming environment

2) Clean data, i.e. remove duplicates; if data are text-based, convert them to numerical values; otherwise machine will learn bad patterns

3) Split whole data set into training and testing sets: e.g. a set of 1000 data entries gets split into 700-800 training and 200-300 testing sets; this is needed to ascertain the accuracy of your model

4) Create a model: select the type of machine learning algorithm to learn the data (different types see next slides)

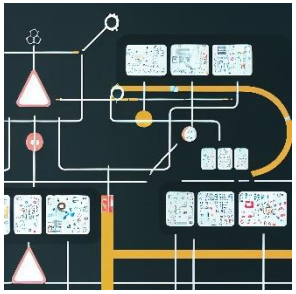
5) Train the model: feed the training data to your model

6) Predictions: ask model to analyze testing set

7) Evaluate and improve your machine learning model

Hamedani M., “Python Tutorial - Python Full Course for Beginners”, *YouTube*, Programming with Mosh, 18.02.2019, https://www.youtube.com/watch?v=_uQrJ0TkZlc&t=15183s

Which types of machine learning are there?



1) Supervised Learning

Useful for labelled training data such as: set of images of cats, dogs and birds
→ raw data = pixels of images of varying RGB values
→ discrete labels = [‚cat‘, ‚dog‘, ‚bird‘]
The human ‚expert‘ tells the algorithm which outputs to expect from the raw data

2) Unsupervised Learning

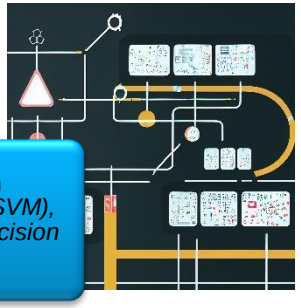
Can be eymployed when processing unlabelled training data; helpful in finding hidden patterns in given dataset, like:
→ customer segmentation (clustering by age, nationality, gender, …)
→ pre-processing of images, removing visual noise to increase picture quality (see Dimensionality Reduction)
No human intervention in learning process, no labelling, the algorithm is not told what to expect from the data

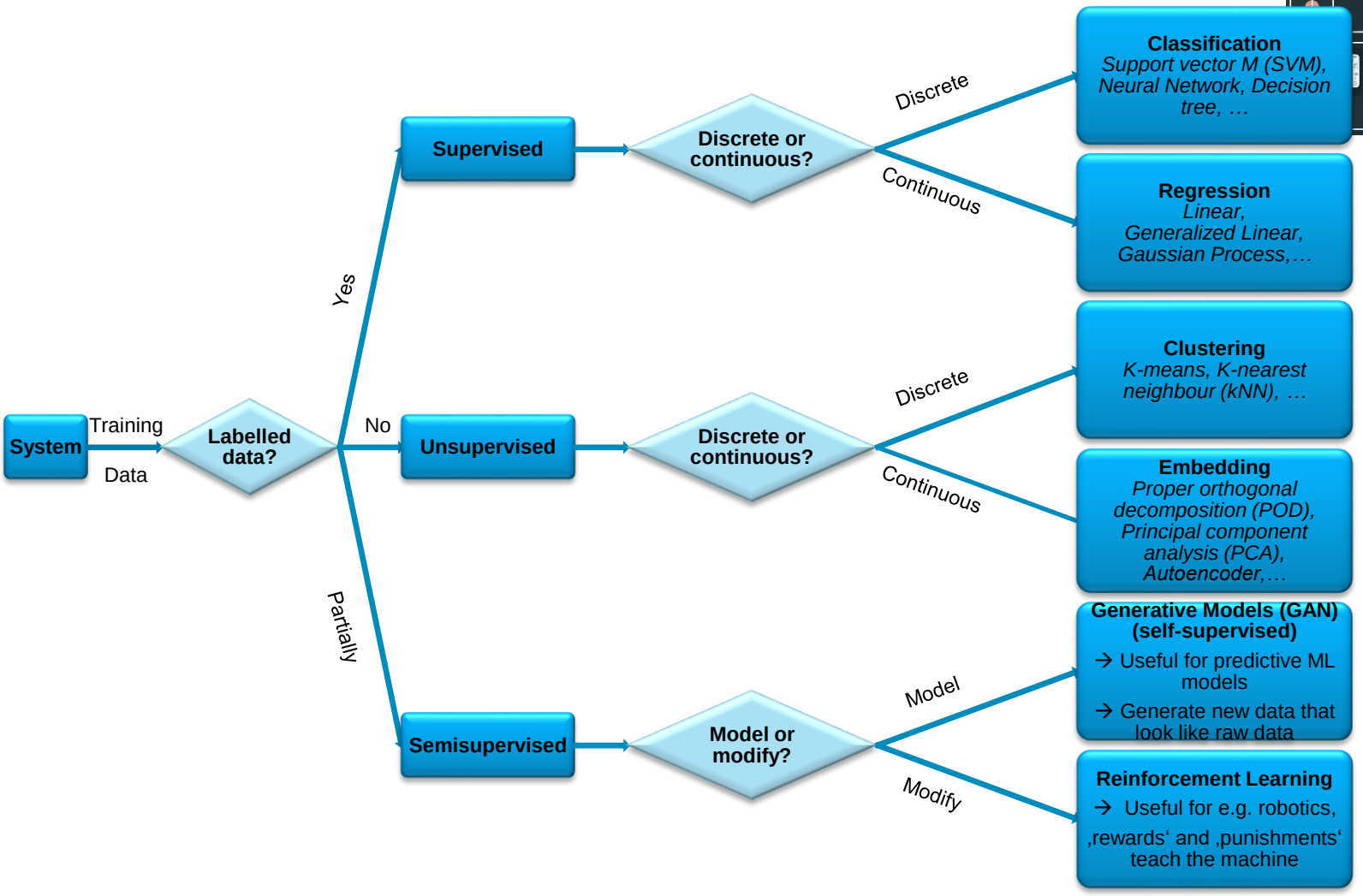
3) Semisupervised Learning (compromise between supervised and unsupervised)

Take this approach when analyzing partially labelled training data, e.g. given a dataset of a million images wherein only a few thousand images are labelled

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
Keen M., “Supervised vs. Unsupervised Learning“, *YouTube*, IBM Technology, 27.07.2022, https://www.youtube.com/watch?v=W01tIRP_Rqs

Which approach should I take for my data set?

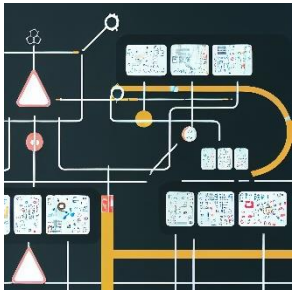




Brunton S., "Types of Machine Learning 2", *YouTube*, Steve Brunton, 06.06.2019, https://www.youtube.com/watch?v=0_lKUPYEYyY

"A.I. learns to walk", *YouTube*, Code Bullet, 20.04.2019, https://www.youtube.com/watch?v=K-wIZuAA3EY

What are applications & limitations to Organic Chemistry?



{Title of Grzybowski2019 in React.Chem.Eng.
Contains hyperlink to respective publication}

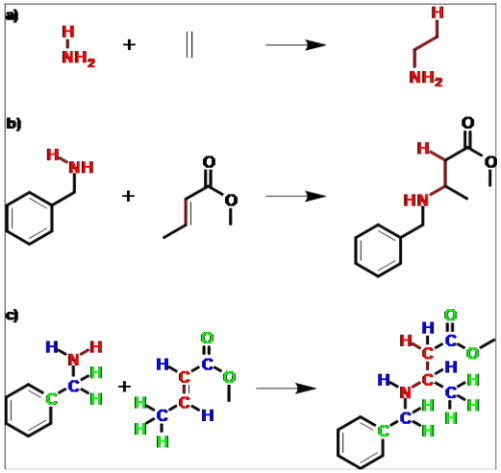
Motivation: provide a framework for translating synthetic-organic knowledge into reaction rules understandable by a machine

a) Defining reaction core and its environment

reaction rules must account for:
- all-important flanking groups
- atoms taking place in reaction (red)
- neighbouring atoms of radius 1 (directly connected to partaking atoms, blue)
- neighbouring atoms of radius 2 (atoms connected to atoms of radius 1, green), and so forth

b) Limitations of automatic rule extraction

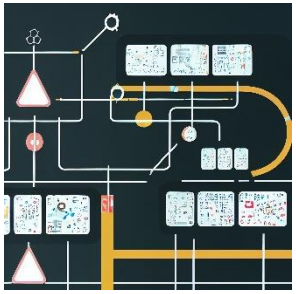
- popular reaction types (e.g. Wittig olefinations, Suzuki couplings) have many (> 10^5) precedents for reaction rules
- specialized reactions (e.g. stereospecific C-H insertion of carbenes yielding tertiary alcohols) being uncommon (~20 examples in Reaxys) deems automatic extraction of meaningful statistics and reaction rules impossible
- multistep cascade reactions and mechanisms are hard to account for in ML model (only 30 examples for anionic [4+2]-cycloaddition)



{cropped screenshot of Figure 2.c) from Grzybowski2019}

Grzybowski, B. A. *et al*., "The logic of translating chemical knowledge into machine-processable forms: a modern playground for physical-organic chemistry", *React. Chem. Eng.*, **2019**, 4, 1506-1521

What are the applications & limitations to Organic Chemistry?



c) Mechanism-based rule coding

- uses SMARTS (SMILES arbitrary target specification) notation for reaction rules
- e.g. Diastereoselective addition of vinylmagnesium bromides to Michael acceptors, catalyzed by Cu(I)
- example „[CH2,O:7]“: either methylene or oxygen are allowed at position 7 (‚A‘ in the ring) as cyclic enones and unsat. lactones both are valid substrates (also shown in decision tree)

{screenshot of Figure 11. from Grzybowski2019, Highlighting the SMARTS code for including either Methylene or oxygen into ring}

d) The importance of structural context

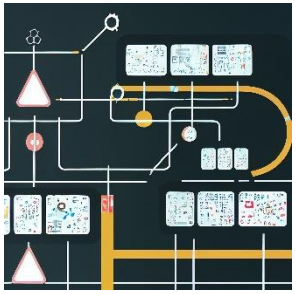
- reaction rule must account for prohibited motifs, e.g. intermediates that violate Bredt’s rule and other highly strained structures EXCEPT that highly strained structure is aim of the retrosyn analysis

e) Accounting for non-local electronic effects

usage of Hammett’s constants and proton affinities as basis for acumulative electronic and mesomeric effects of substituents on (het-)aromatic systems

Grzybowski, B. A. *et al*., “The logic of translating chemical knowledge into machine-processable forms: a modern playground for physical-organic chemistry”, *React. Chem. Eng.*, **2019**, 4, 1506-1521
Daylight Chemical Information Systems, https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, visited 24.10.2022

What are the applications & limitations to Organic Chemistry?



{Title of Newhouse2021 in ChemRXiv
Contains Hyperlink to respective publication}

Motivation: design a computer-aided strategy to optimize key steps in total synthesis

Key step to be optimized: radical 6-*endo*-trig cyclization

{cropped screenshot of Figure 1.A from Newhouse2021,
Showing the retrosynthetic analysis of clovan-2,9-dione
*From intermediate **8**}*

Start:

- plotting reported yields vs. Computed free energies (DFT) of 125 literature-known 6-*endo*-trig cyclizations (figure **B**)

→ no correlation, hence outcome determined by many more factors

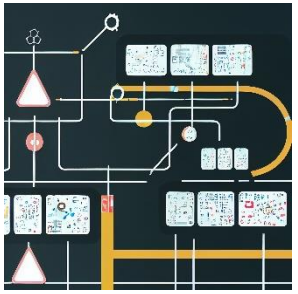
{screenshot of Figure 2.B from Newhouse2021}

Next steps:

- restrict 6-*endo*-trig data set to sp^3-centered radicals undergoing cyclization onto pendant olefin (→ 99 rxns)
- calculate physical descriptors (molecular, atomic, steric, …) using unrestricted DFT calculations (→ 340 descriptors for each reaction)
- 340 descriptors for 99 rxns may cause ‚overfitting' → use correlation filtering and PCA dimension reduction to transform 340 descriptors into 20 orthogonal parameters

Newhouse T *et al*., "A Neural Network Model Informs Total Synthesis of Clovane Sesquiterpenoids", *ChemRxiv*. Cambridge: Cambridge Open Engage; 2021; This content is a preprint and has not been peer-reviewed.

What are the applications & limitations to Organic Chemistry?



- Benchmark several supervised ML models against each other (figure **C**)

{screenshot of Figure 2.C from Newhouse2021}

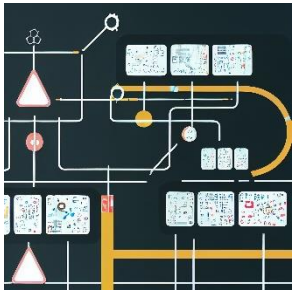
- → Neural network model (NNET) shows highest goodness-of-fit ($R^2 = 0.82$)

- control experiments to prove soundness of NNET model:
  replace chemically meaningful descriptors with random values (figure **D**)
- This yields low correlation between predicted and reported yields ($R^2 = 0.02$)
- → predictions made my NNET model rely on chemically meaningful descriptors/properties rather than random chance

{screenshot of Figure 2.D from Newhouse2021}

Newhouse T *et al*., “A Neural Network Model Informs Total Synthesis of Clovane Sesquiterpenoids”, *ChemRxiv*. Cambridge: Cambridge Open Engage; 2021; This content is a preprint and has not been peer-reviewed.

What are the applications & limitations to Organic Chemistry?



- With trained NNET model, three intermediates (**7-9**) were tested:

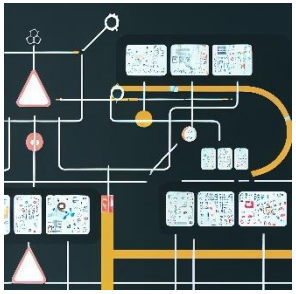
{cropped screenshot of Figure 3.A from Newhouse2021}

- Synthetically useful predicted yield of intermediate **7** expected with polarized alkene
- Intermediate **8** was chosen as synthetic target as the corresponding disconnection seemed innovative
- Synthetic application of „learned radical reactivity“ reaches clovan-2,9-dione in 5 steps and excellent agreement of predicted and experimental yield for 6-*endo*-trig cyclization :

{screenshot of Figure 3.B from Newhouse2021}

Newhouse T *et al*., “A Neural Network Model Informs Total Synthesis of Clovane Sesquiterpenoids”, *ChemRxiv*. Cambridge: Cambridge Open Engage; 2021; This content is a preprint and has not been peer-reviewed.

What are the applications & limitations to Organic Chemistry?



{Title of Novitskiy2022 in J. Org. Chem.
Contains Hyperlink to respective publication}

Motivation: provide a hybrid ML/DFT approach for simulating NMR spectra of (large) natural products (NPs)

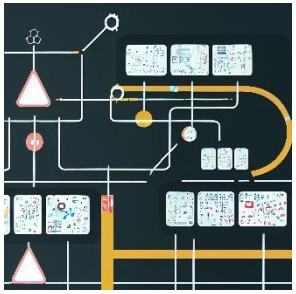
- DFT may predict NMR of NPs with high accuracy
- DFT for large NPs computationally expensive and time-consuming
- Computer-Assisted Structure Elucidators (CASE) generate and assess large numbers of structural candidates quickly but with low accuracy
- → Machine Learning-augmented Density functional theory as ‚Goldilocks zone' of acceptable accuracy and time for calculation

{screenshot of Figure 1 from Novitskiy2022}

Basic process:
- geometry optimization with ‚cheap' functional and basis set
- Nuclear magnetic shielding DFT calc. → crude chemical shifts / correct by ML-derived corrections yields augmented chemical shift predictions
- Fermi contact DFT calc. → natural bond orbital (NBO) corrections / correct by ML-derived corrections yields augmented coupling constants

Novitskiy, I. M., Kutateladze, A. G., "DU8ML: Machine Learning-Augmented Density Functional Theory Nuclear Magnetic Resonance Computations for High-Throughput In Silico Solution Structure Validation and Revision of Complex Alkaloids", *J. Org. Chem.*, **2022**, 87, 4818-4828

What are the applications & limitations to Organic Chemistry?



- Using DU8ML, authors could predict structures of nearly 170 reported alkaloids‘ NMR Spectra with short calculation times (under 150 min per structure)

- 35 misassigned structures where identified and revisions (matching the experimental NMR data) were proposed
  e.g. cyctetryptomycin A revised structure

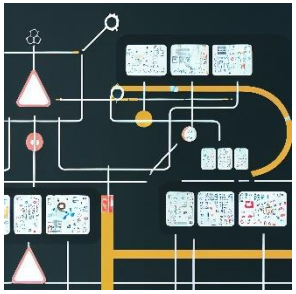
{screenshot of Figure 2 from Novitskiy2022}

{screenshot of Figure 12 from Novitskiy2022}

- (c)rmsd (*solvent-corrected* root-mean-square-deviation) value as NMR-analogue for goodness-of-fit

Novitskiy, I. M., Kutateladze, A. G., “DU8ML: Machine Learning-Augmented Density Functional Theory Nuclear Magnetic Resonance Computations for High-Throughput In Silico Solution Structure Validation and Revision of Complex Alkaloids”, *J. Org. Chem.*, **2022**, 87, 4818-4828

What are the major Limitations to Machine Learning?

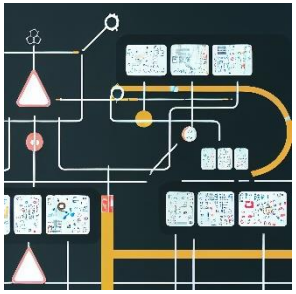


1) **Ethical limitations***
   Self-driving cars; who‘s to blame in case of accident?
2) **Deterministic problems**
   ML for weather forecast → ML can‘t understand underlying physics of weather system, might predict nonsensical weather conditions (e.g. negative rainfall)
3) **Lack of (quality) data**
   - creating a model from low quality data, or biased data creates a low quality or biased model
   - overfitting the model, i.e. training the model to also account for noise in sampling data → yields poor results when testing with new data
4) **Lack of interpretability**
   ML algorithm may become a black box that ‚performs‘ well while also being too complex for human understanding → typical problem in Deep Learning
5) **Lack of reproducibility**
   Small changes in data or differences in software environment between ML-algorithm designer and user may lead to failure of the ML model.
   → Can be attenuated by building reproducibility into machine learning; proper documentation of the ML design process

*lacks applicability in Organic Chemistry

Postindustria, Inc., https://postindustria.com/what-are-the-major-limitations-of-machine-learning-algorithms/, 18.03.2022
EliteDataScience, https://elitedatascience.com/overfitting-in-machine-learning, 06.07.2022
Interpretable AI, LLC, https://www.interpretable.ai/interpretability/what/, visited: 23.10.2022
DecisivEdge, https://www.decisivedge.com/blog/the-importance-of-reproducibility-in-machine-learning-applications/, visited: 23.10.2022

Conclusions & wrap-up



Machine learning can be a useful tool for

- Modelling systems with many variables
- Creating retrosynthetic analyses of organic compounds
- Help with structure elucidation of complex natural products

However, …

- Another skill with two separate, steep learning curves (learning Python or MatLab code & meaningful, reliable ML workflow)
- May have severe drawbacks and pitfalls depending on the problem at hand

Universität Konstanz

Thanks for your attention!
Questions or comments?

Image courtesy of the Midjourney AI (https://www.midjourney.com/home/), under creative commons license,
Using the prompt: /imagine Justus Liebig sitting at his desk and coding in Python

Getting started with Python and Machine Learning & Further Reading

- Hamedani, M., „Python Tutorial - Python Full Course for Beginners", *YouTube*, Programming with Mosh (www.youtube.com/watch?v=_uQrJ0TkZlc)

- Playlist: "MIT 6.0002 Introduction to Computational Thinking and Data Science, Fall 2016", MIT OpenCourse Ware, *YouTube*, (https://www.youtube.com/playlist?list=PLUl4u3cNGP619EG1wp0kT-7rDE_Az5TNd)

- Playlist: „Neural networks", 3Blue1Brown, *YouTube* (https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi)

- Brownlee, J. „How to choose a feature selection method for machine learning", https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

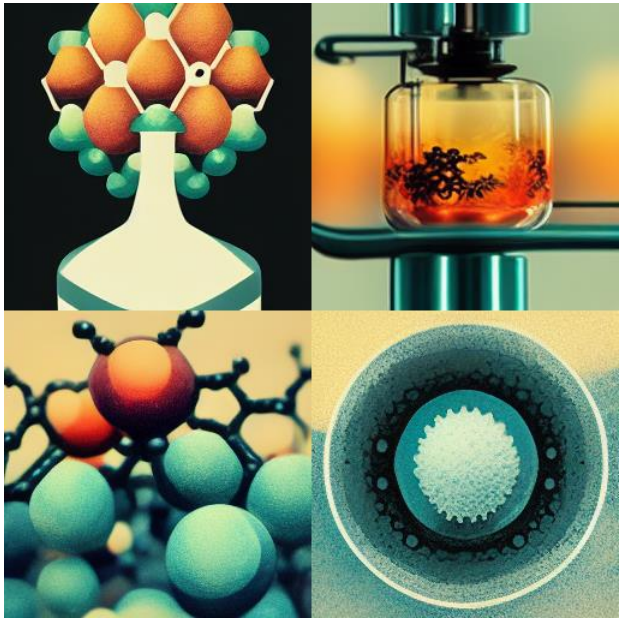
- Kim, D. E., „Total Synthesis of Paspaline A and Emindole PB Enabled by Computational Augmentation of a Transform-Guided Retrosynthetic Strategy", *J. Am. Chem. Soc.*, 2019, 141 (4), 1479-1483

- Badowski, T. *et al.*, „Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning", *Angew. Chem. Int. Ed.*, 2020, 59, 725 –730

- Dral, P. O., „Quantum Chemistry in the age of Machine Learning", *J. Phys. Chem. Lett.*, 2020, 11, 2336-2347

Image courtesy of the Midjourney AI (https://www.midjourney.com/home/), under creative commons license, Using the prompt: /imagine Democritus of Abdera staring into a server room

Some unexpected and questionable results of an image fabricating AI



Midjourney AI prompt:
/imagine organic chemistry but its machine learned

Midjourney AI prompt:
/imagine these are the missing links between machine learning and organic chemistry

All images created with the Midjourney AI under creative commons license

Some unexpected and questionable results of an image fabricating AI

Midjourney AI prompt:
/imagine Organic chemistry natural products synthesis machine learning

Midjourney AI prompt:
/imagine lsd trip bad drug trip machine learning limitations organic chemistry

All images created with the Midjourney AI under creative commons license