Machine Learning in Drug Discovery: Applications and Techniques

MAGDALENA WIERCIOCH

Jagiellonian University

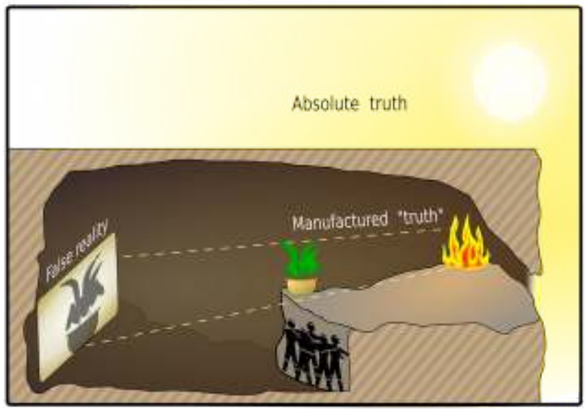
May 5th, 2022

Artificial Intelligence in Research and Applications Seminar

Outline

- Research interests
- Data representation
- Drug discovery
- Methods

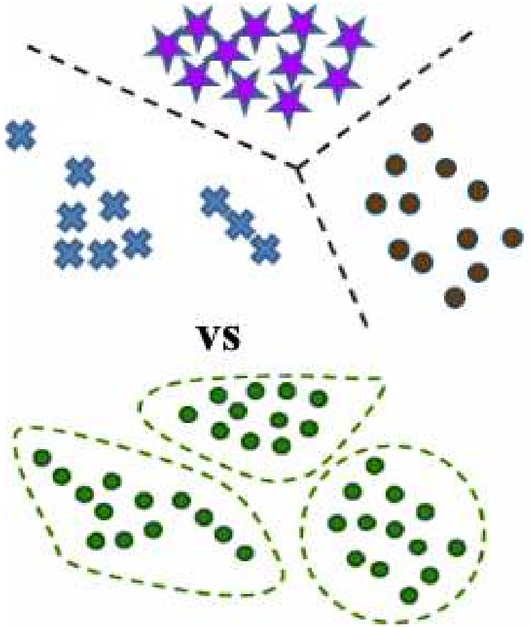
Research interests



source: Wikimedia Commons (Therea Knotts)

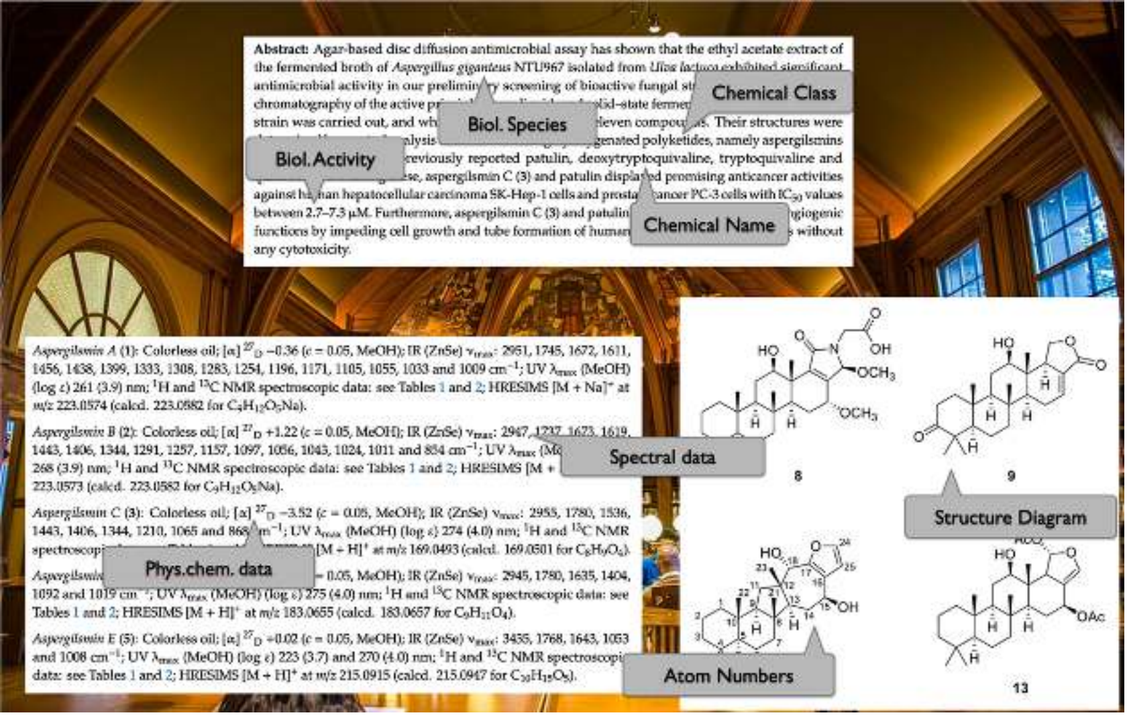
representation learning

hypothesis: the success of machine learning algorithms depends on data representation



supervised learning, unsupervised learning

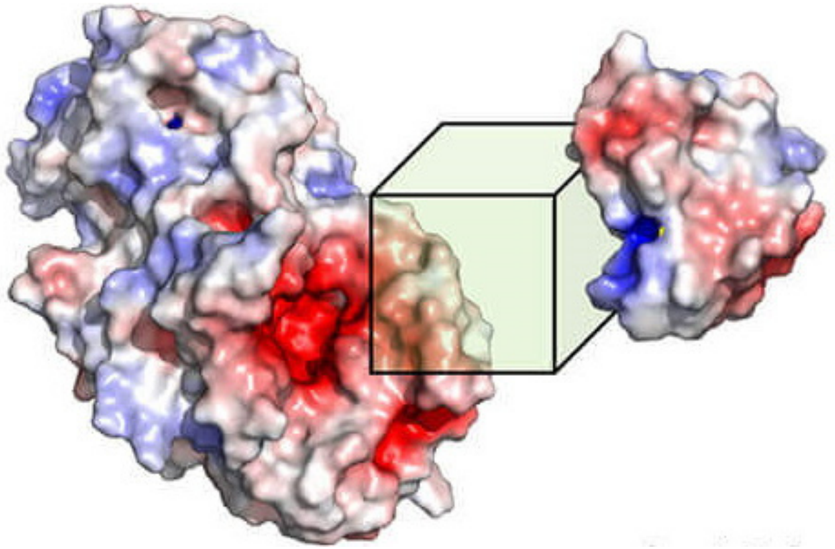
XAI: Explainable Artificial Intelligence



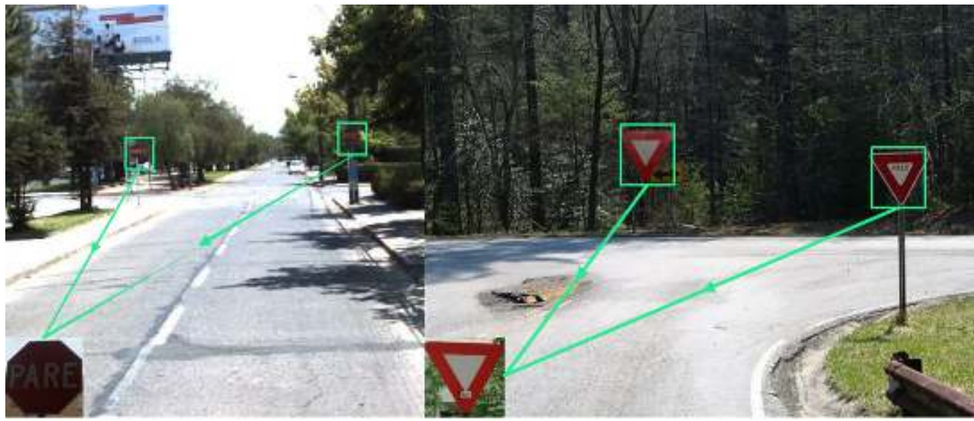
Rajan et al., 2020

Liu et al., 2019

source: Sansan



source: Wang et al., 2019



Villalón-Sepúlveda et al., 2017

Representations

source: Wikipedia

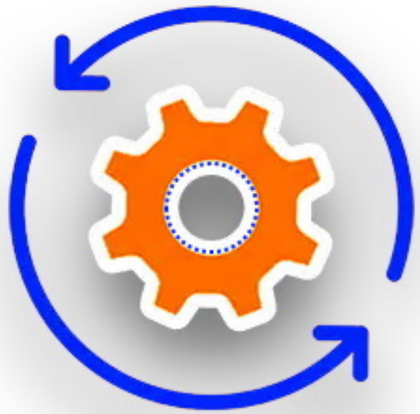
'elephant'

The image that shows an elephant.

The predictive system.

The output (prediction).

Representations



'elephant'

The image that shows an elephant.

The predictive system.

The output (prediction).

Representations

We see this:

source: Wikipedia

The computer sees this:

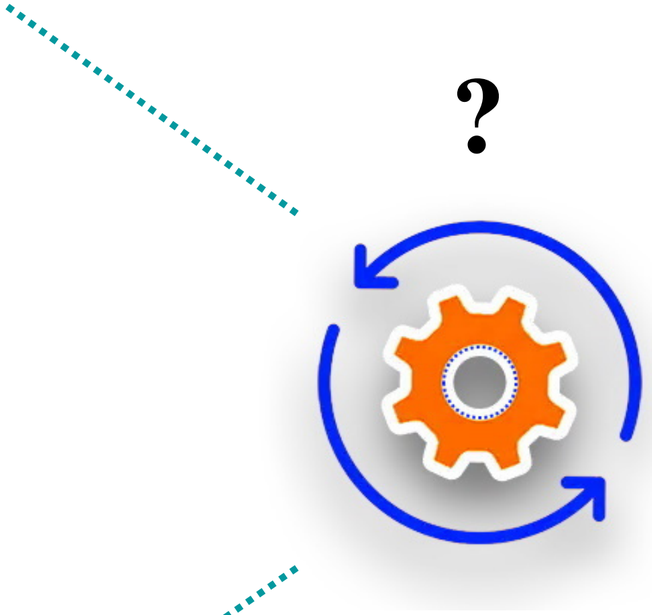
```
49 49 99 40 17 81 18 57 60 87 17 40 98 43 69 48 04 56 62 00
81 49 31 73 55 79 14 29 93 71 40 67 53 88 30 03 49 13 36 65
08 02 22 97 38 15 00 40 00 75 04 05 07 78 52 12 50 77 91 08
22 31 16 71 51 67 63 89 41 92 36 54 22 40 40 28 66 33 13 80
24 47 32 60 99 03 45 02 44 75 33 53 78 36 84 20 35 17 12 50
32 98 81 28 64 23 67 10 26 38 40 67 59 54 70 66 18 38 64 70
67 26 20 68 02 62 12 20 95 63 94 39 63 08 40 91 66 49 94 21
24 55 58 05 66 73 99 26 97 17 78 78 96 83 14 88 34 89 63 72
21 36 23 09 75 00 76 44 20 45 35 14 00 61 33 97 34 31 33 95
78 17 53 28 22 75 31 67 15 94 03 80 04 62 16 14 09 53 56 92
16 39 05 42 96 35 31 47 55 58 88 24 00 17 54 24 36 29 85 57
86 56 00 48 35 71 89 07 05 44 44 37 44 60 21 58 51 54 17 58
19 80 81 68 05 94 47 69 28 73 92 13 86 52 17 77 04 89 55 40
04 52 08 83 97 35 99 16 07 97 57 32 16 26 26 79 33 27 98 66
88 36 68 87 57 62 20 72 03 46 33 67 46 55 12 32 63 93 53 69
04 42 16 73 38 25 39 11 24 94 72 18 08 46 29 32 40 62 76 36
20 69 36 41 72 30 23 88 34 62 99 69 82 67 59 85 74 04 36 16
20 73 35 29 78 31 90 01 74 31 49 71 48 86 81 16 23 57 05 54
```

Machine Learning problems: classification

ML systems return predictions from examples.

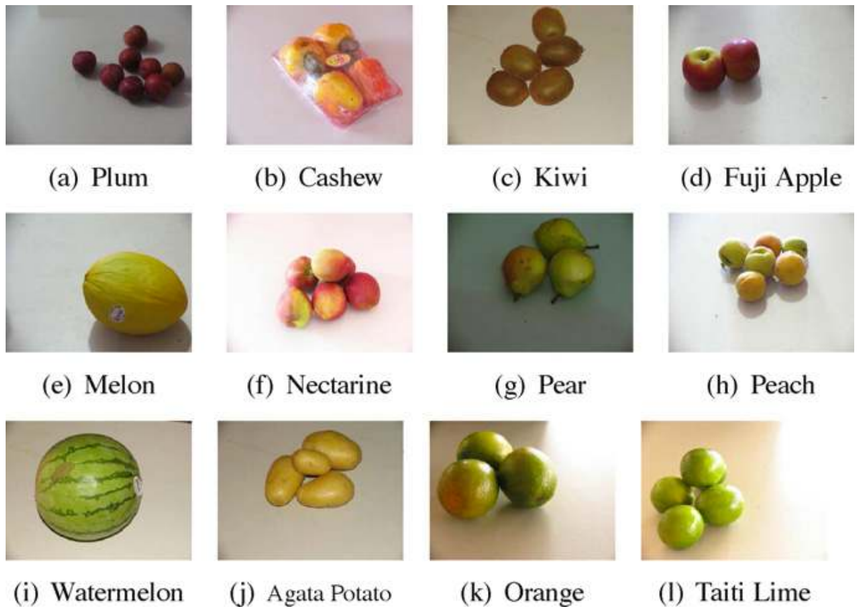
source: Wikipedia

source: Wikipedia



Machine Learning problems: classification

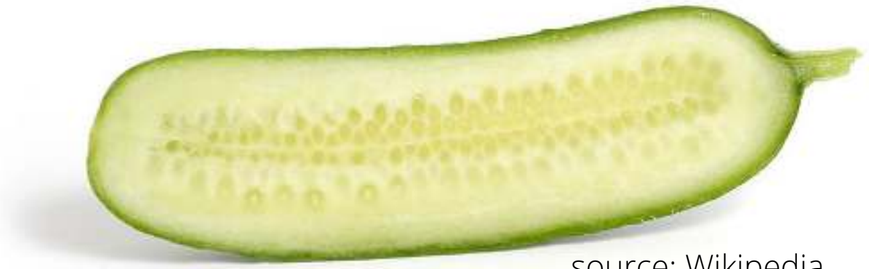
- Point out the species of a particular fruit (i.e., apple, banana, pear)
- Point out not only the species of a particular fruit but also its variety (i.e., Golden Delicious, Jonagold, Fuji)
- ...



Rocha et al., 2010

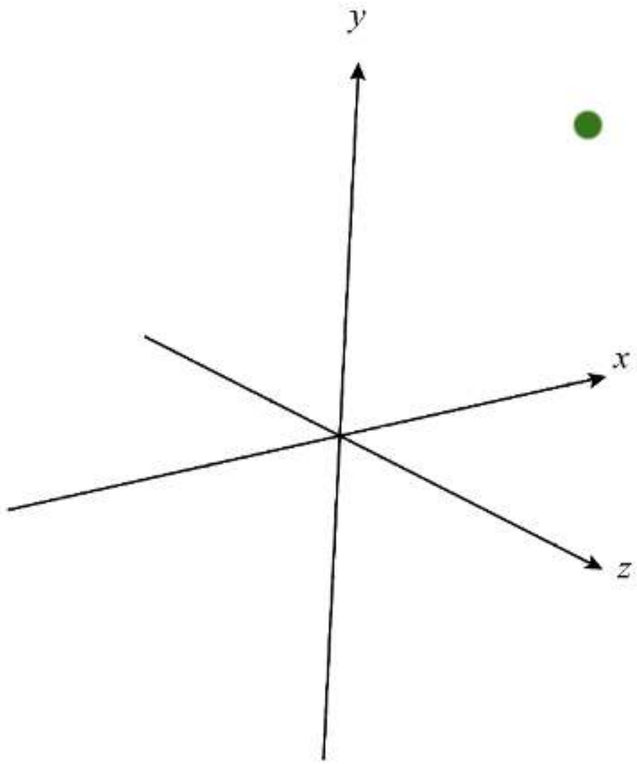
Machine Learning problems: classification

- We want to classify products as fruits or vegetables.
- Let's represent a product as a list of numbers.
  - What colour is it?
  - Does it contain seeds?
  - Does it have leaves?
  - ...

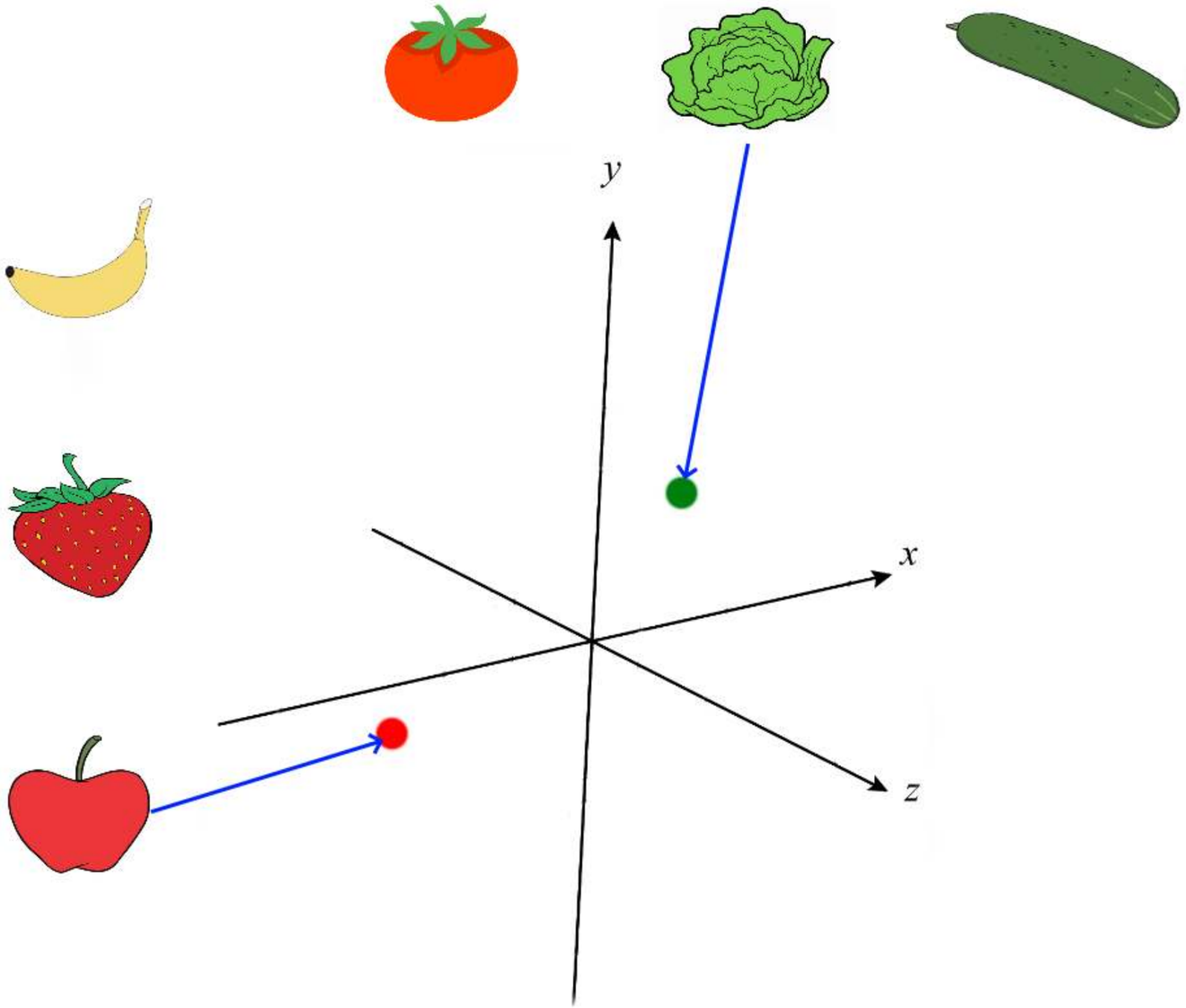


source: Wikipedia

representation = (24, 1, 0)

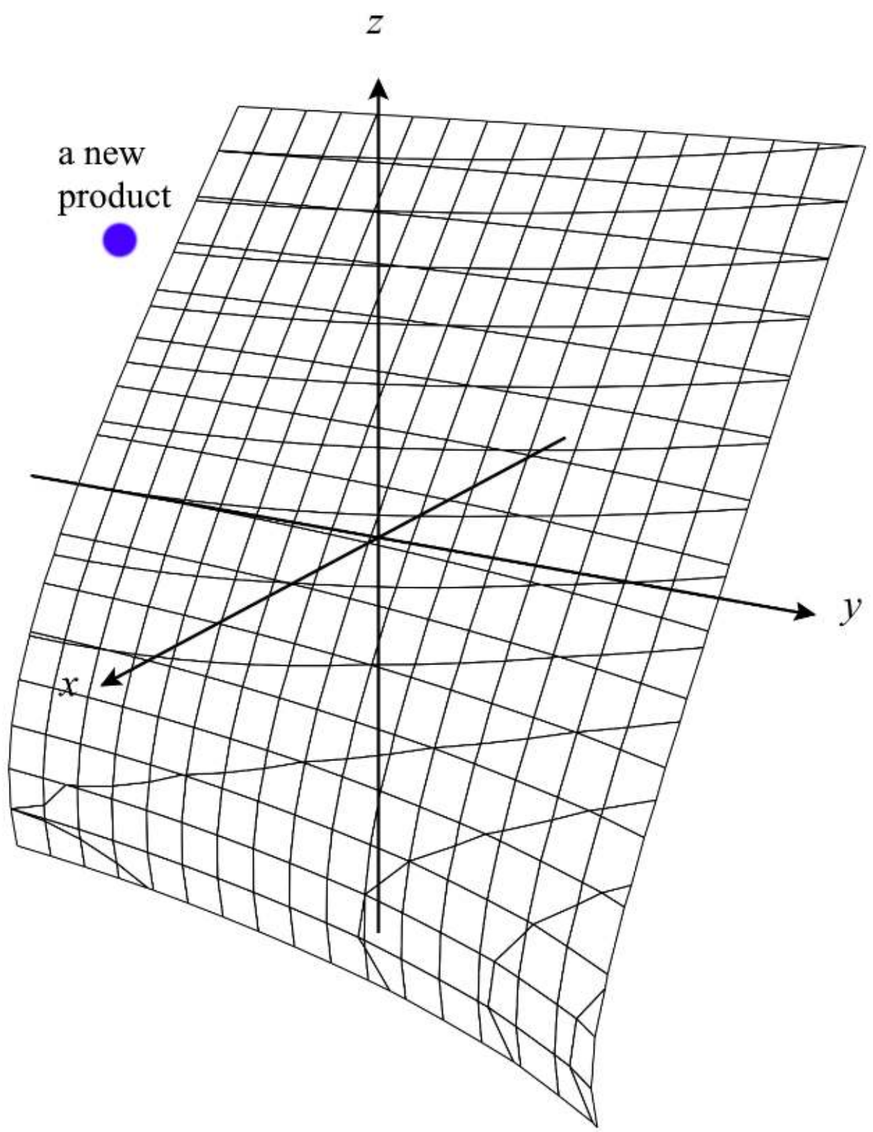


Machine Learning problems: classification



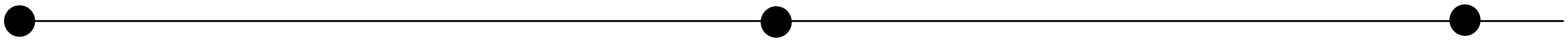
- Each product is a point in 3D space.

Machine Learning problems: classification



- The goal is to find a surface that separates the products.
- For a new product, one has to calculate a representation and point out the side of the surface.

Representations - history



feature engineering

- features do not scale well
- limited expressivity

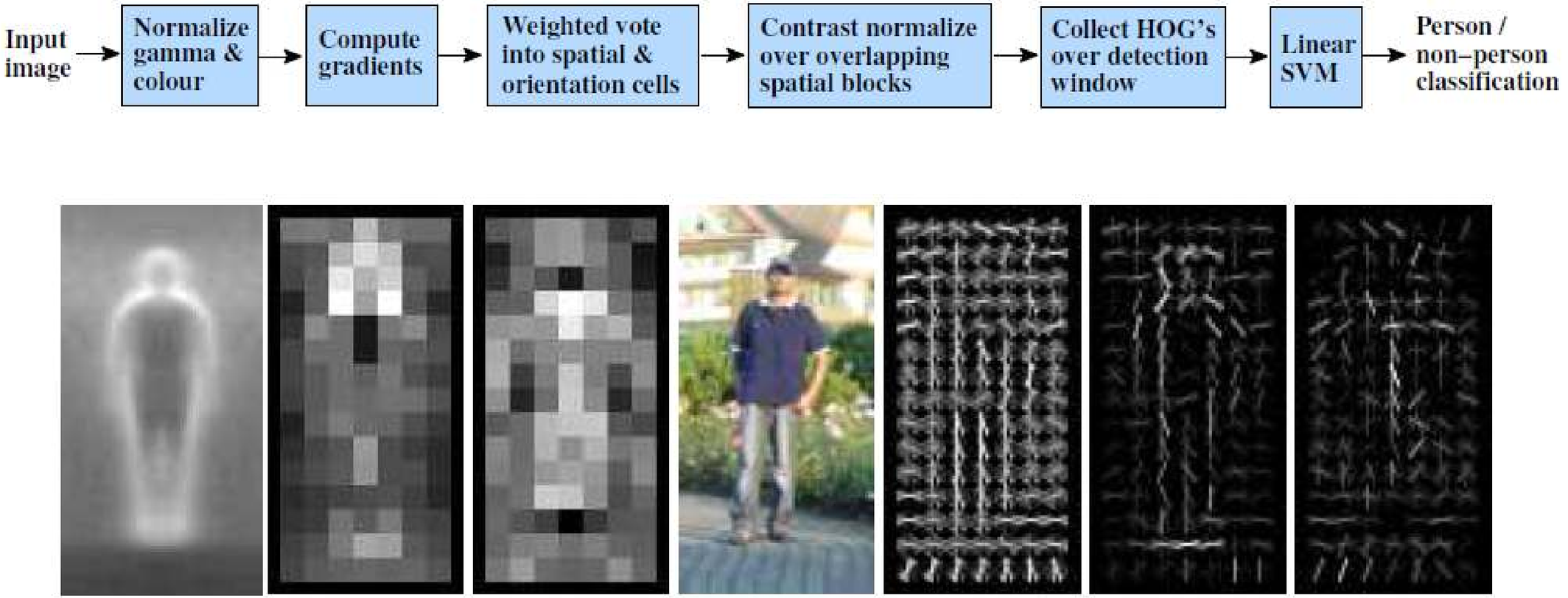
deep learning with representation learning

- expressivity

representation learning

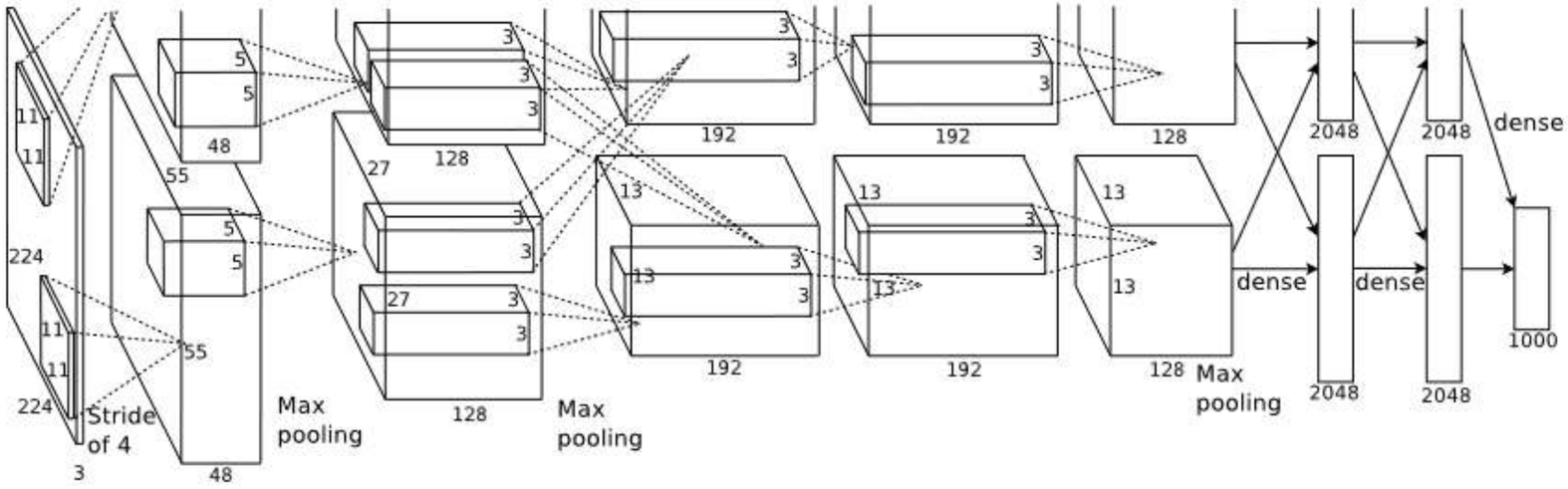
- usability

Representations - feature engineering



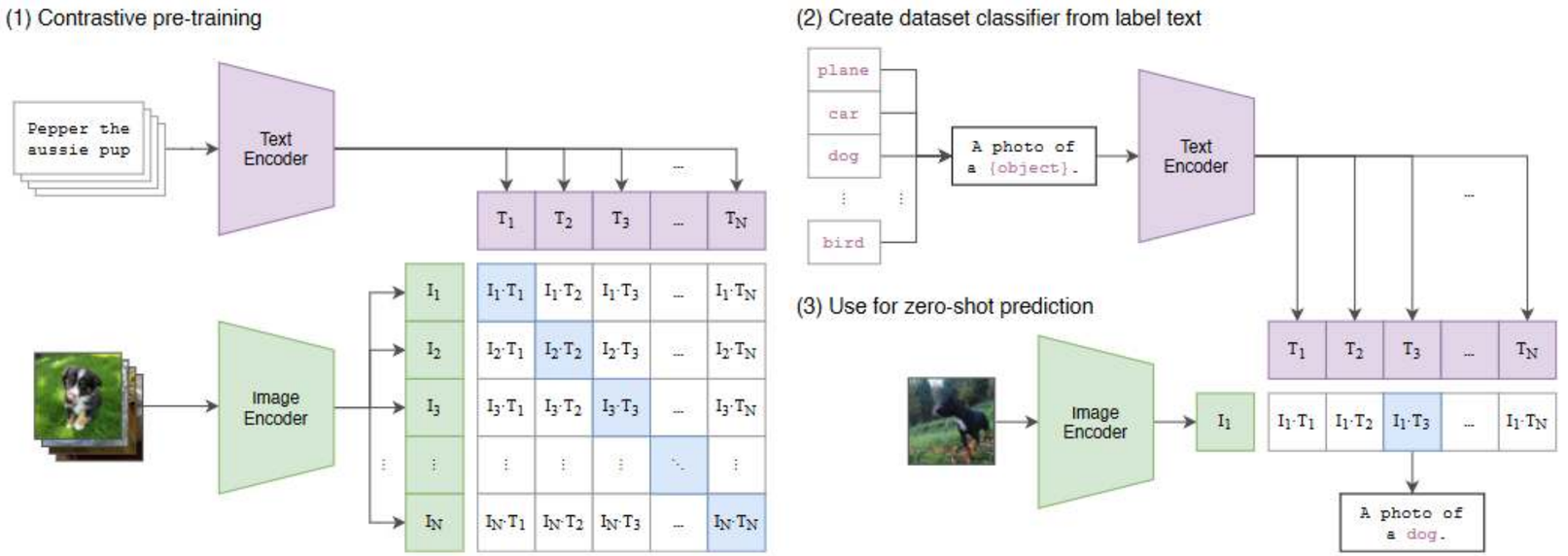
Dalal and Triggs, 2005

Representations - deep learning with representation learning



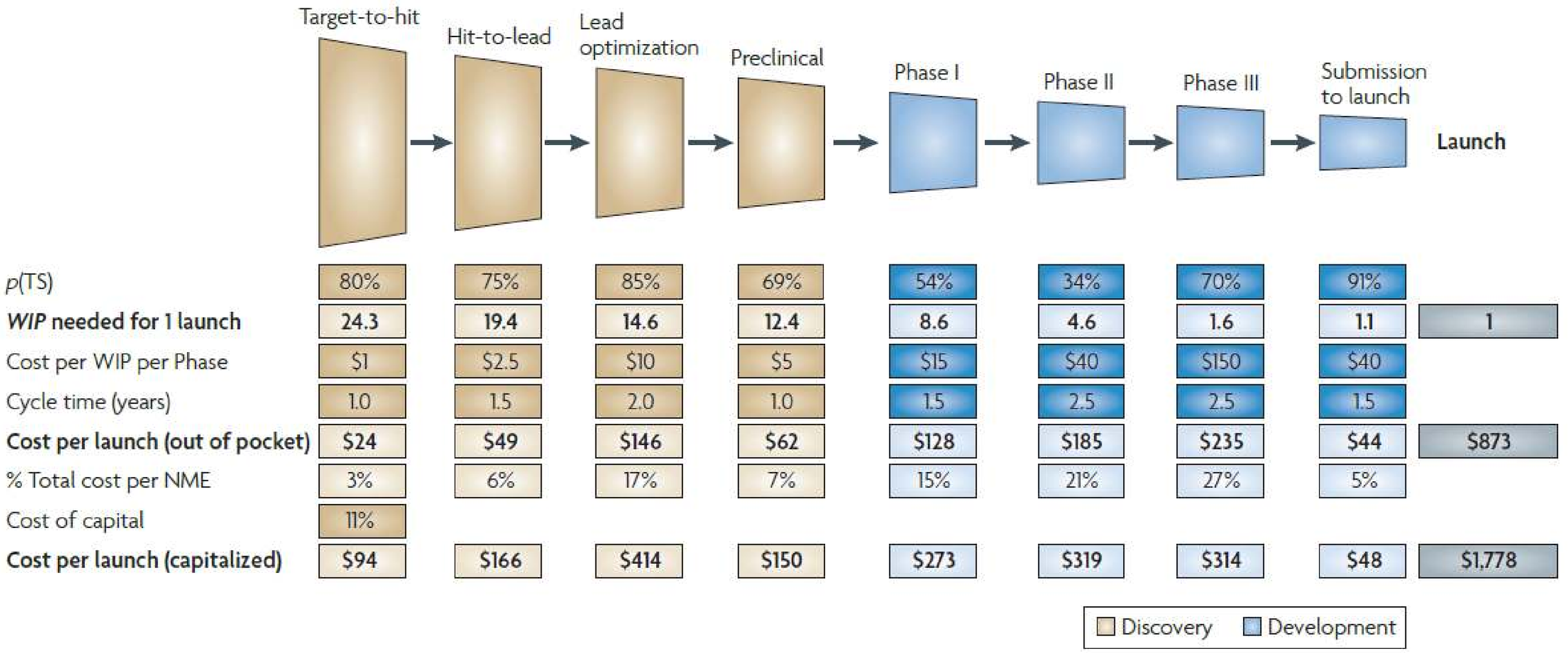
Krizhevsky et al., 2012

Representations - representation learning



Radford et al., 2021

Drug discovery



Paul et al., 2010

Drug discovery

Drug discovery

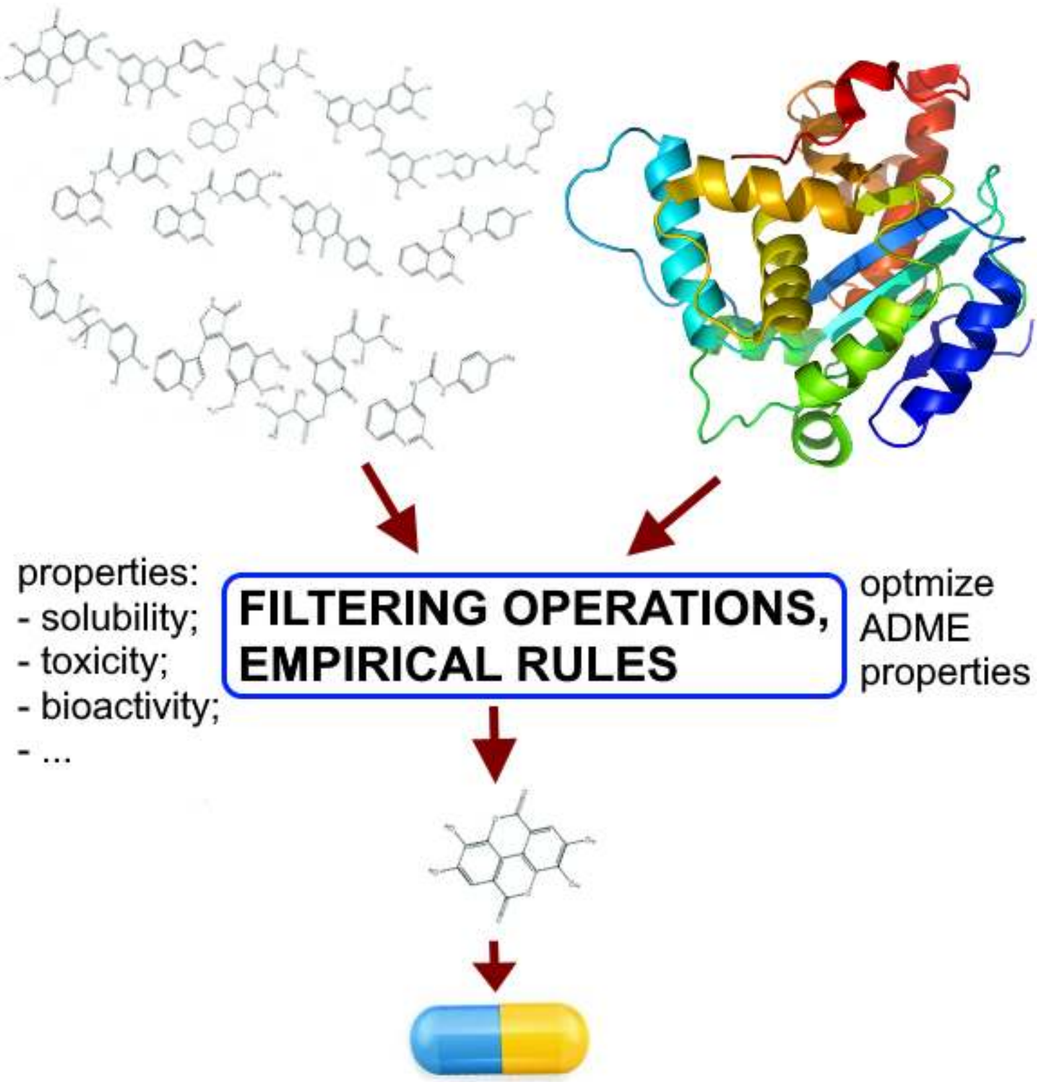
Virtual screening (with machine learning)

Enables to prioritize compounds from compound libraries which have a high potential to bind to a target of interest.

- **faster and cheaper than wet lab experiments**

However:

- restricted to the avaliable compounds;
- uses hand-crafted features.



Drug discovery

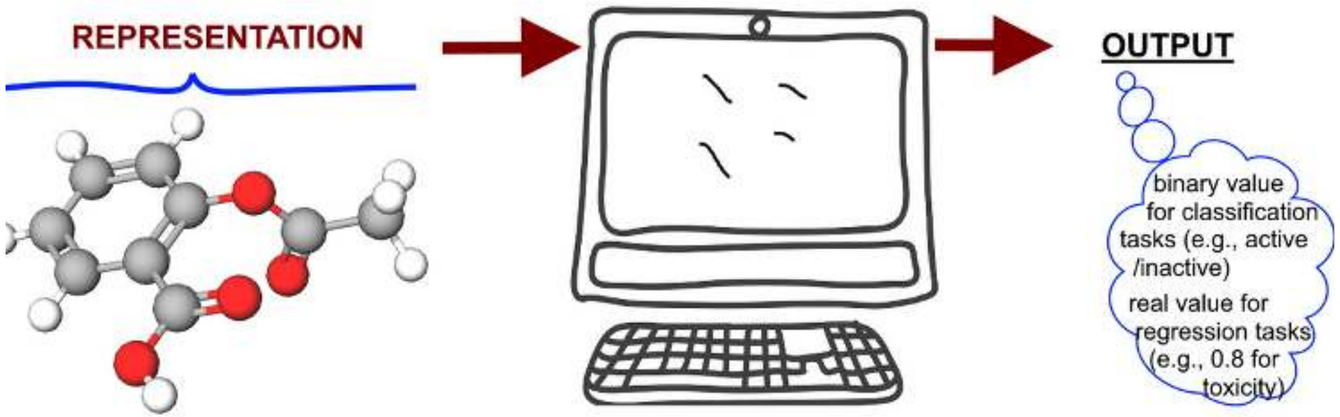
Virtual screening (with machine learning)

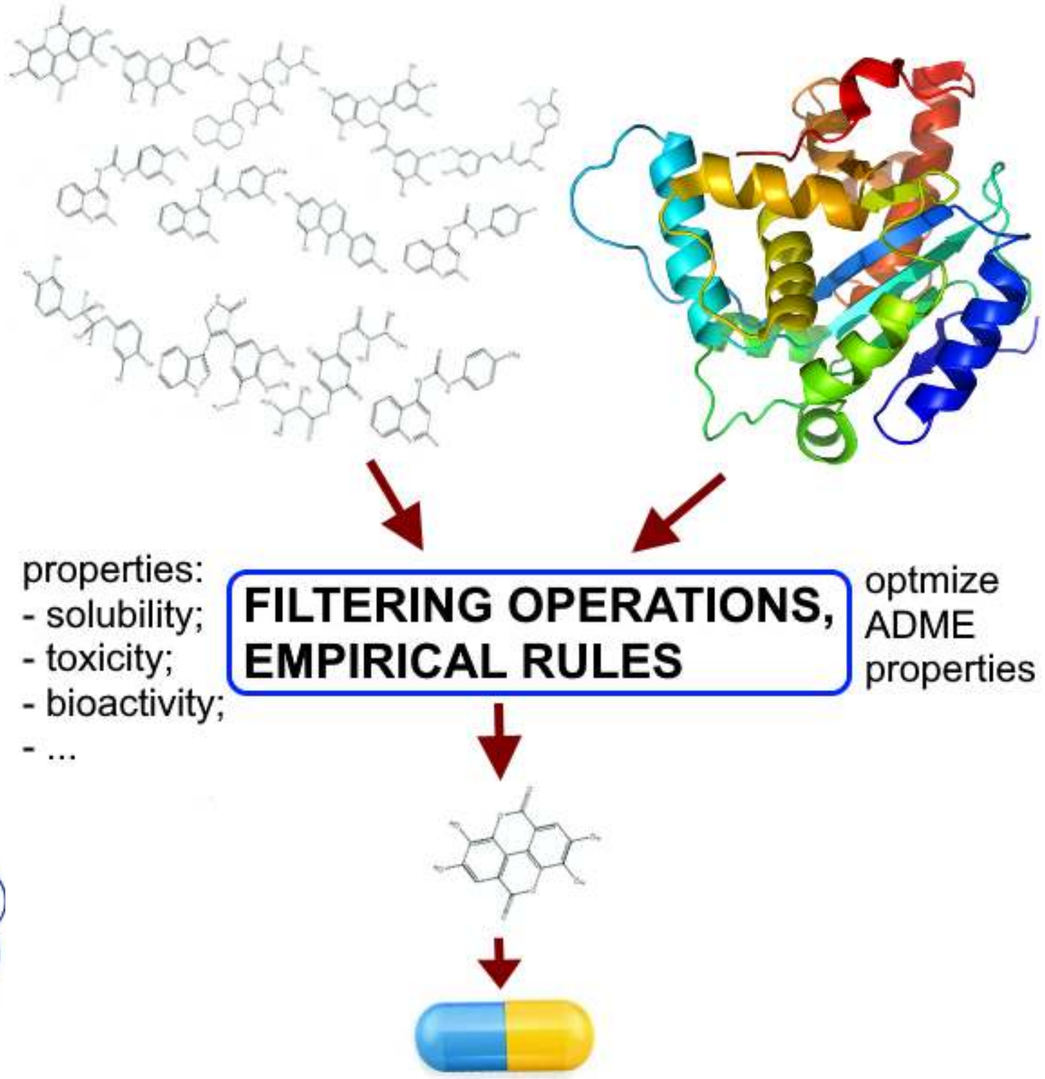
Enables to prioritize compounds from compound libraries which have a high potential to bind to a target of interest.

- **faster and cheaper than wet lab experiments**

However:

- restricted to the avaliable compounds;
- uses hand-crafted features.





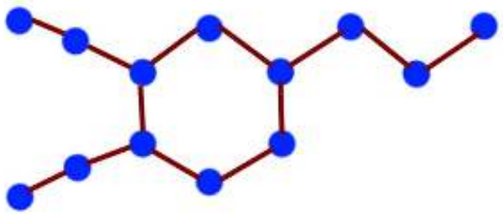
Drug discovery

Representations

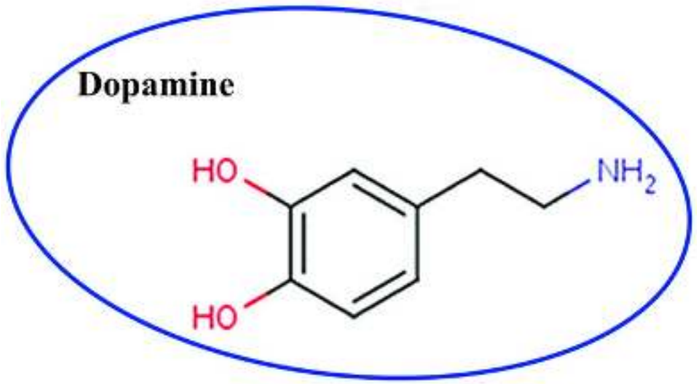
SMILES

C1=CC(=C(C=C1CCN)O)O

Molecular graph

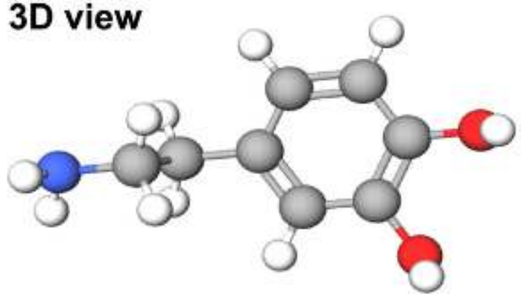


Dopamine



Fingerprint

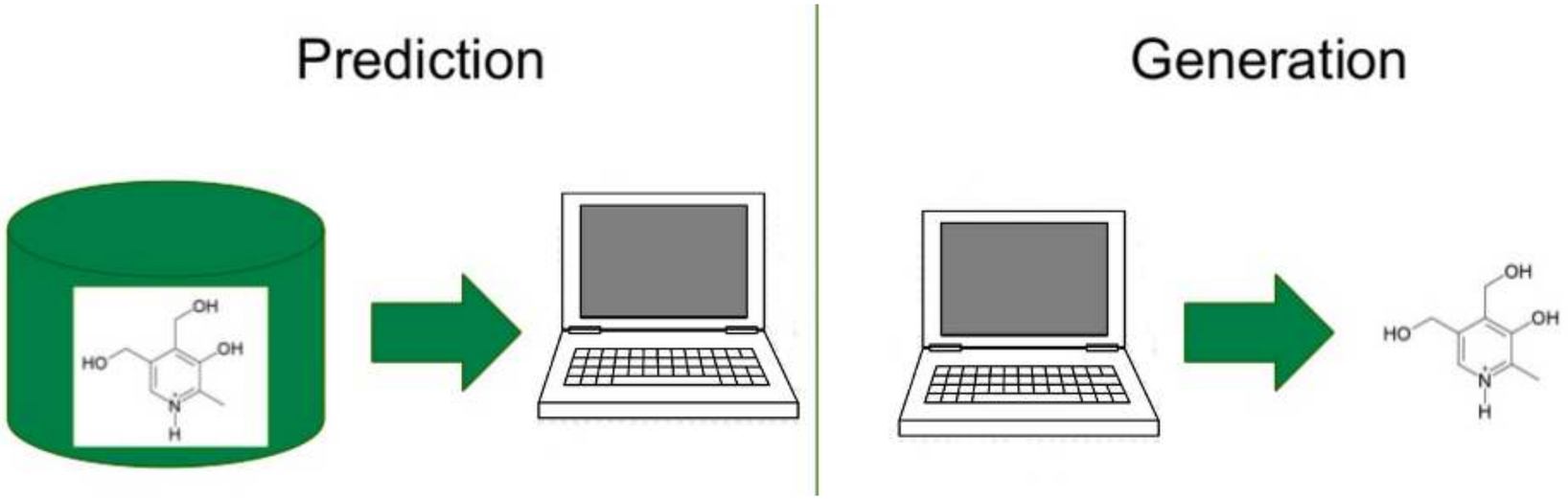
3D view



Drug discovery

Representations

It shows the accurate modeling and prediction of molecular properties is strictly connected with the choice of molecular representation (*Cano et al., 2017; Wiercioch, 2018*; *Wiercioch, 2019*; *Chuang et al., 2020*).



- Searching for molecules with desired properties from given compound libraries.
- Produce molecules that have desired properties.

Methods

M. Wiercioch, On Modeling Objects Using Sequence of Moment Invariants, in *Proceedings of the 17th International Conference CISIM 2018*, Olomouc, Czech Republic, 2018

- This paper explores the problem of rotational invariance of objects.
- A lot of compounds representations and metrics are available but none reflects the activity satisfactory.

Theorem

Let us consider complex moments up to the order $r \geq 2$. Let a set of rotation invariants B be constructed as follows:

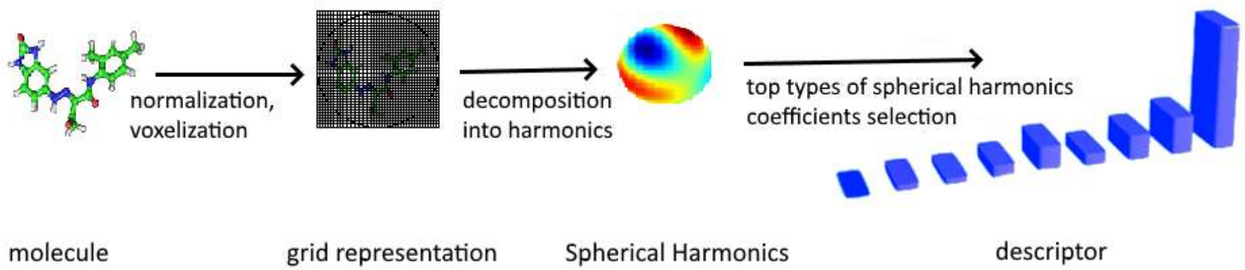
$$B = \{\phi(p,q) \equiv c_{p,q} c_{q_0,p_0}^{p-q} | p \geq q \wedge p + q \leq r\}$$

where p_0 and q_0 are arbitrary indices such that $p_0 + q_0 \leq r$, $p_0 - q_0 = 1$ and $c_{p_0 q_0} \neq 0$ for all admissible one dimensional objects. Then B is a basis of all rotation invariants created from the moments of any kind up to the order r.

Methods

M. Wiercioch, Exploring the Potential of Spherical Harmonics and PCVM for Compounds Activity Prediction, in *International Journal of Molecular Sciences*, 23 pages, 2019

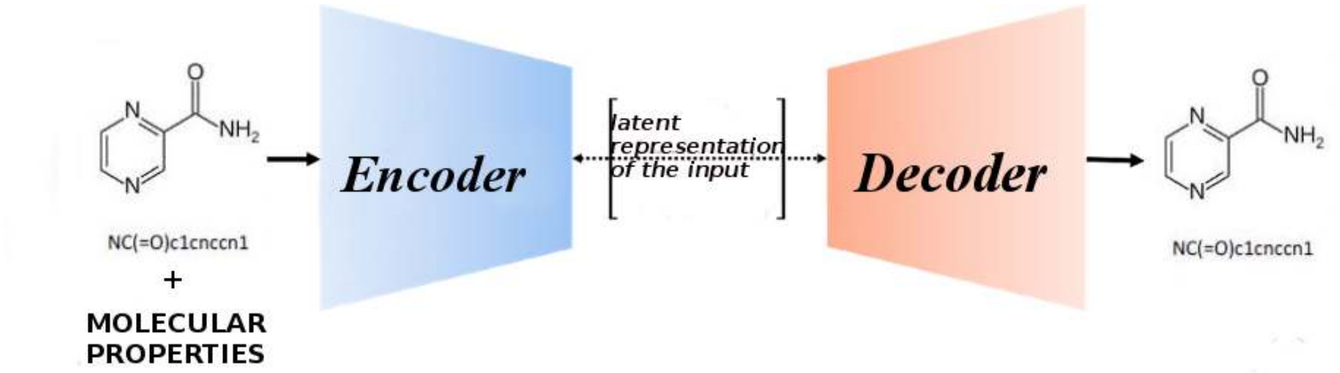
- A methodology that involves Probabilistic Classification Vector Machines (PCVM) and Spherical Harmonics-based descriptor.
- Experimental results for G protein-coupled receptors (GPCRs) demonstrate SHPCVM produces the best performance ranging from 0.742 Accuracy to 0.862, and from 0.691 to 0.794 in terms of Matthew Correlation Coefficient. Although the goal was to find out a tradeoff between the descriptive capabilities and computational costs of the descriptor, our approach may pave the way for more interpretability oriented research on molecule's computational model.



Methods

M. Wiercioch, S. Podlewska, Automated de-novo molecule design based on Deep Neural Networks, in *14th German Conference on Chemoinformatics*, Mainz, Germany, 2018

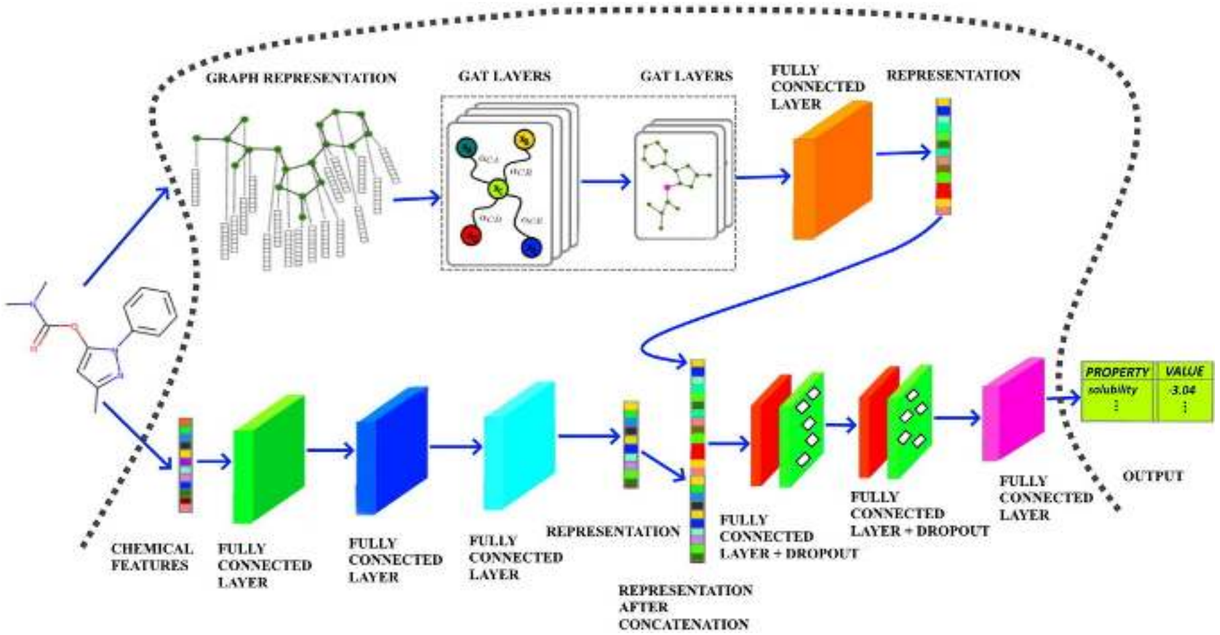
- We propose a molecular generative model called FGVAE that uses the grammar variational autoencoder (GVAE) (Kusner et al., 2017).
- In our model, the molecular properties we want to consider were added as the extra production rules that can be used for constructing a molecule.



Methods

M. Wiercioch, J. Kirchmair, Deep Neural Network Approach to Predict Properties of Drugs and Drug-Like Molecules, in *ML for Molecules Workshop at NeurIPS 2020*, Vancouver, Canada, 2020

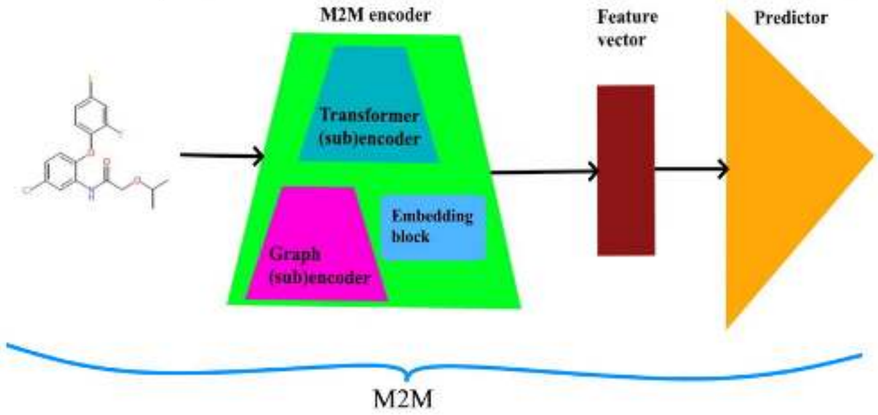
- We propose a deep neural network-based architecture that learns molecular representation to enhance the process of molecular properties prediction.
- The performance of our method is evaluated on the ESOL, FreeSolv, Lipophilicity, ClinTox, BBBP, and BACE datasets from MoleculeNet.

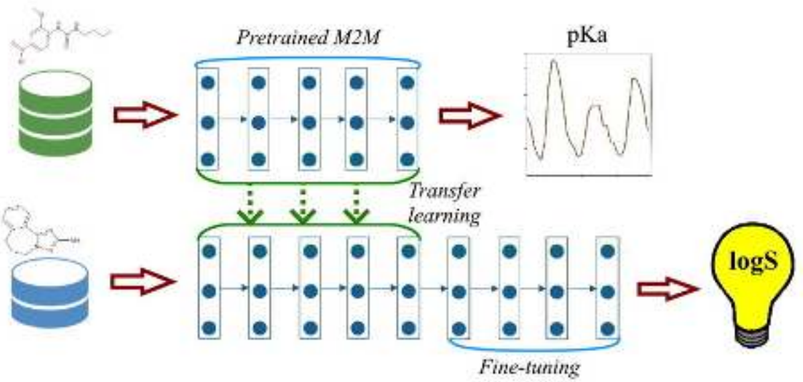


Methods

M. Wiercioch, J. Kirchmair, Dealing with a Data-limited Regime: Combining Transfer Learning And Transformer Attention Mechanism to Increase Aqueous Solubility Prediction Performance, in *Artificial Intelligence in the Life Sciences*, 2021

- We treat aqueous solubility prediction as a translation problem.





Thank you!