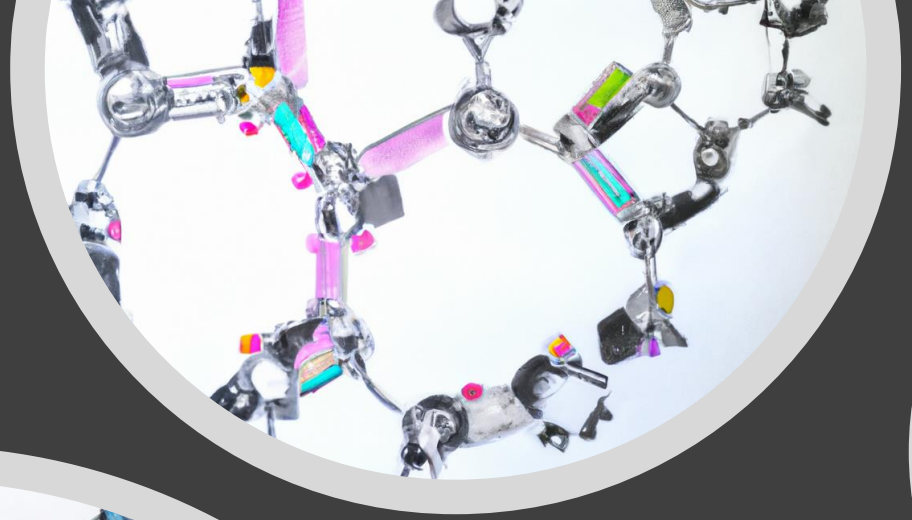
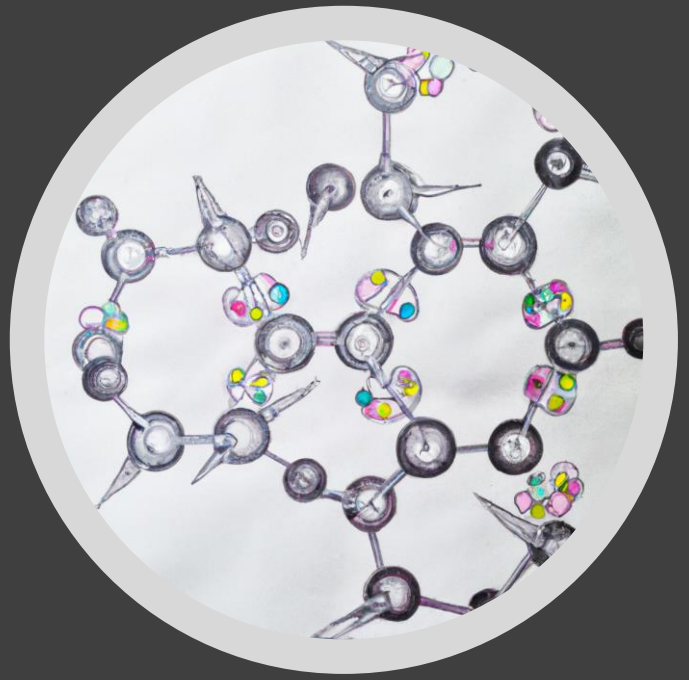
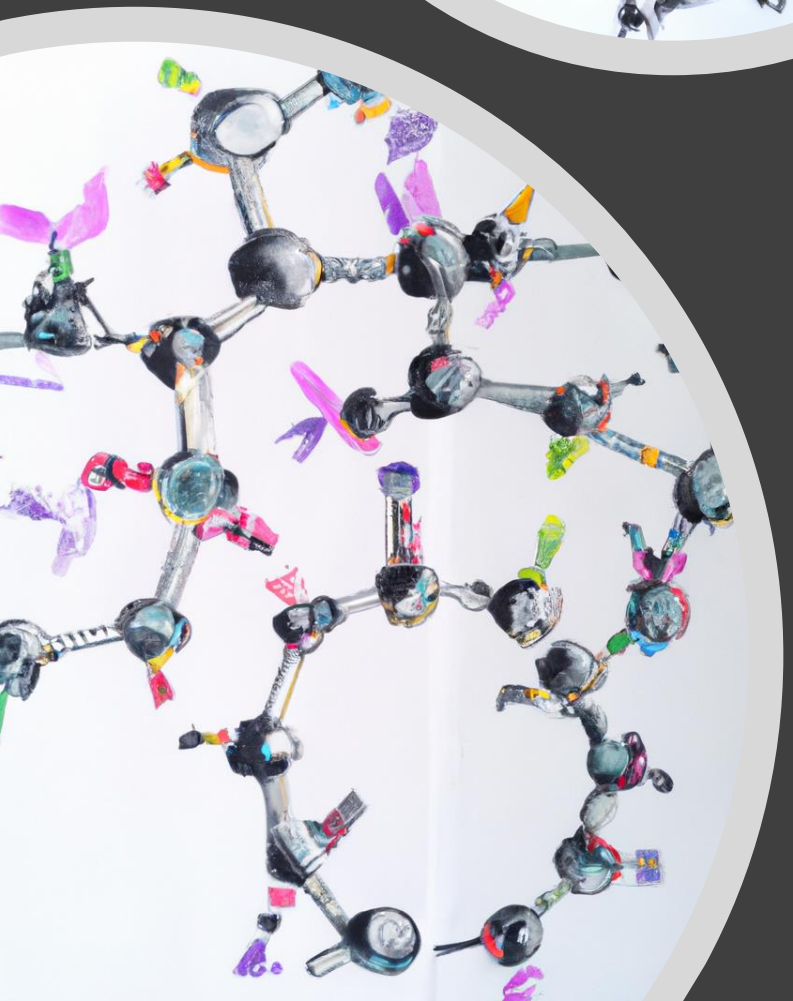






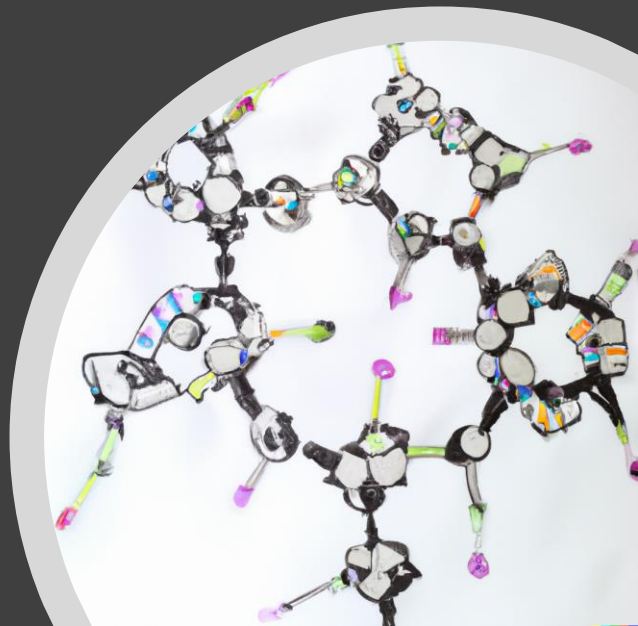
Molecular similarity and optimization

Wim Dehaen, Mgr., PhD

31.1.2023

6th Advanced in silico Drug Design workshop/challenge

Olomouc



Part 1: Molecular similarity

What is similarity?

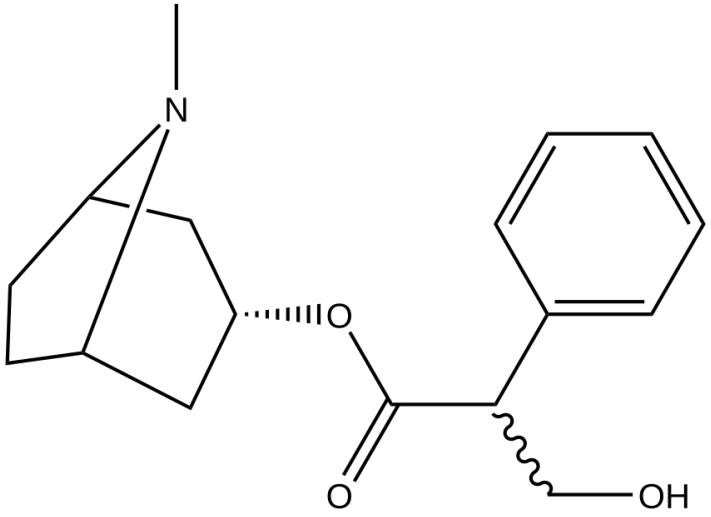
- "Everything is like everything, and in endless ways" - Donald Davidson, What Metaphors Mean

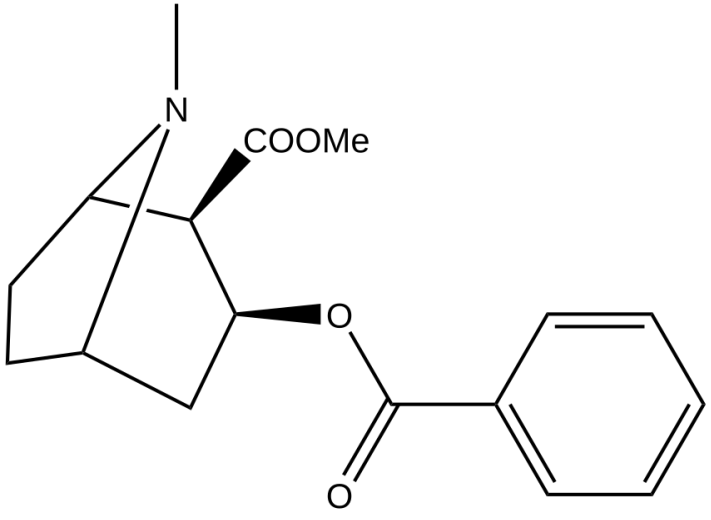
What is similarity?

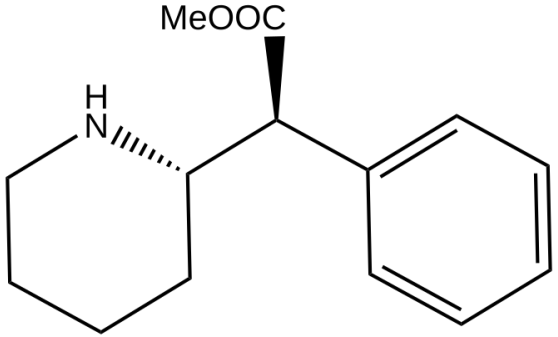
- Similarity is a degree of sameness for different things
- Similarity is a measure of shared features between non-identical things

When and why are molecules similar?

Which of these 3 molecules are most similar to each other?

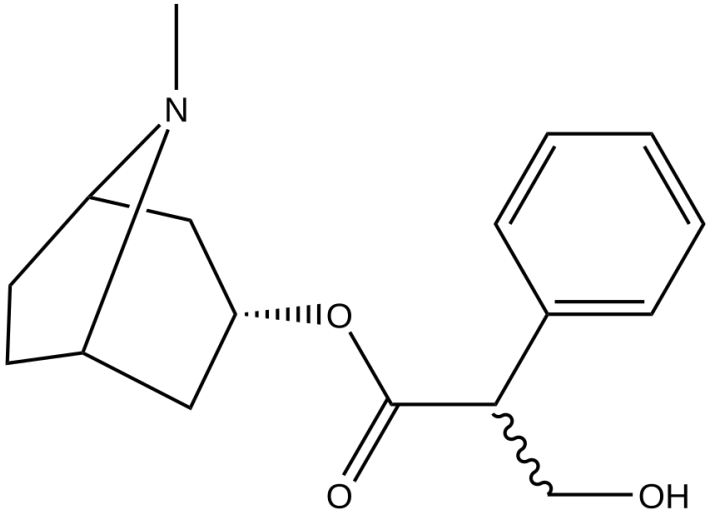




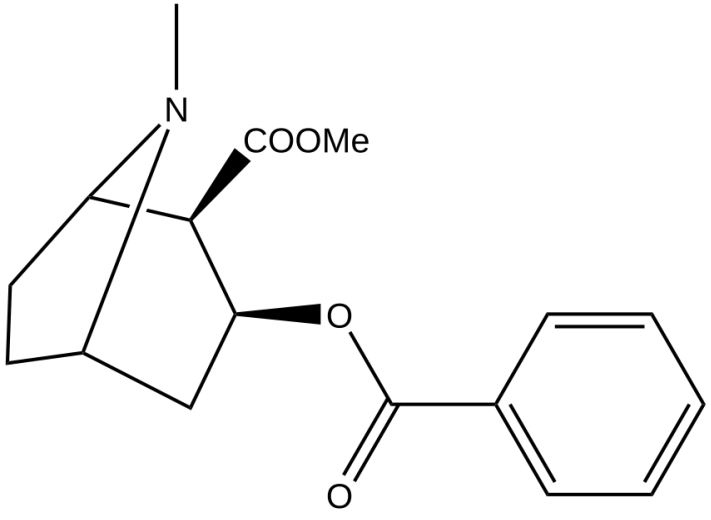


When and why are molecules similar?

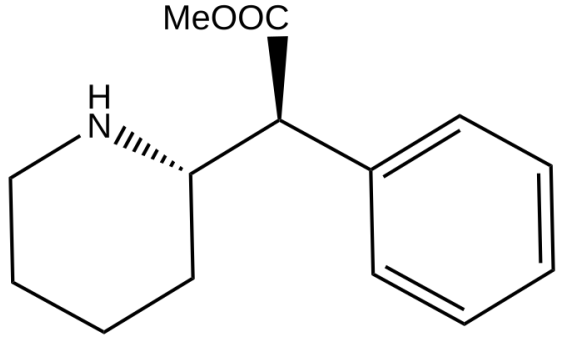
Which of these 3 molecules are most similar to each other?



Atropine
(anticholinergic)



Cocaine
(stimulant drug via DARI)



Methylphenidate
(stimulant drug via DARI)

Why similarity?

- The similarity principle (neighborhood behavior):
  - Similar structures have similar properties, including biological
  - "TS of 0.85 corresponds to same biological activity"
- Applicability Domain problem:
  - More confidence in prediction similar to training data of models
  - Similarity to judge what are the things we know about
- Ligand based drug design/virtual screening:
  - Based on finding important common features in molecules
  - No explicit structural information needed (as in SBDD)

When and why are molecules similar?

- Molecules can be similar in more than one way
- Choosing meaningful features to compare is crucial

In which ways can molecules be similar?

- Topologically: based on atom connectivity
  - Local: presence or non-presence of substructures
  - Global: topological distance of substructures
- Geometrically: based on molecule geometry
  - Euclidean distance of substructures
  - Shape similarity
  - Electrostatic similarity
  - Pharmacophore matches (3d feature distribution)
- Physicochemically: based on physical and chemical properties
  - Can be estimated by models
  - Can be measured

In which ways can molecules be similar?

- Biologically:
  - Can be predicted (e.g. QSAR, pharmacophores)
  - Can be measured
  - In general this the property we want as an endpoint!

Descriptors

- **"[T]he set of all descriptors for a particular compound [can be considered] as being akin to keywords used in a (computer) search of a library of books"** - Stuart Rosenfeld & Nalini Bhushan, *Chemical Synthesis: Complexity, Similarity, Natural Kinds, and the Evolution of a "Logic"*

Descriptors

- Number corresponding to a calculated,predicted or measured property of the molecule
- Presence or non-presence of substructures
- Polarity
- Predicted toxicity
- Graph invariants
- HTS measurement
- 3D features
- Substituent contributions
- …

Fingerprints

- Efficient and standardized representation of chemical features
- Typical form:binary vector of fixed length
- Extended connectivity fingerprint (ECFP/morgan)
- Structural keys (e.g. MACCS)
- Atom pairs
- Pharmacophore fingerprint
- ...
- Use: building models, efficient searching, similarity estimation

Fingerprint example: MACCS vs ECFP

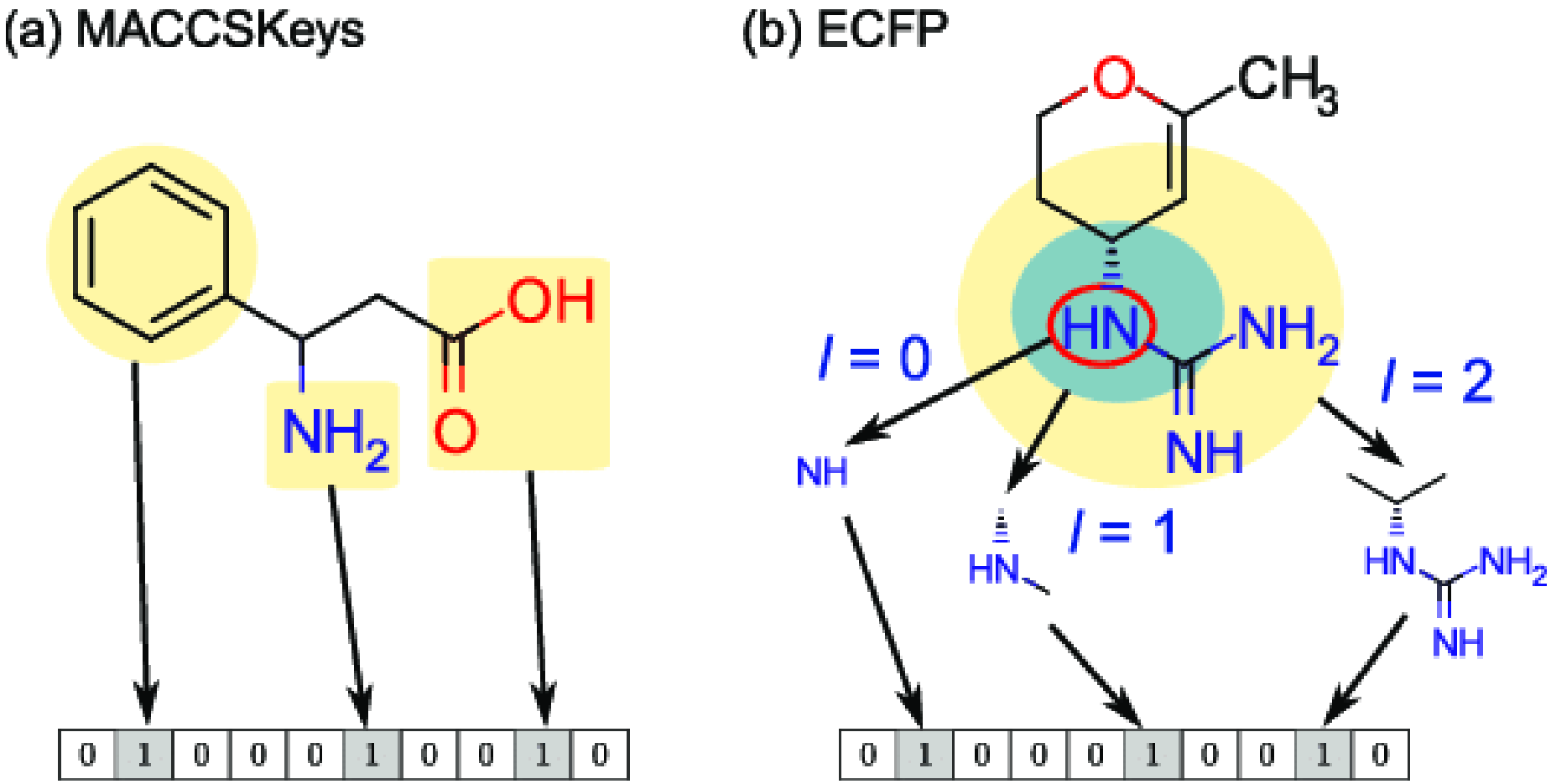
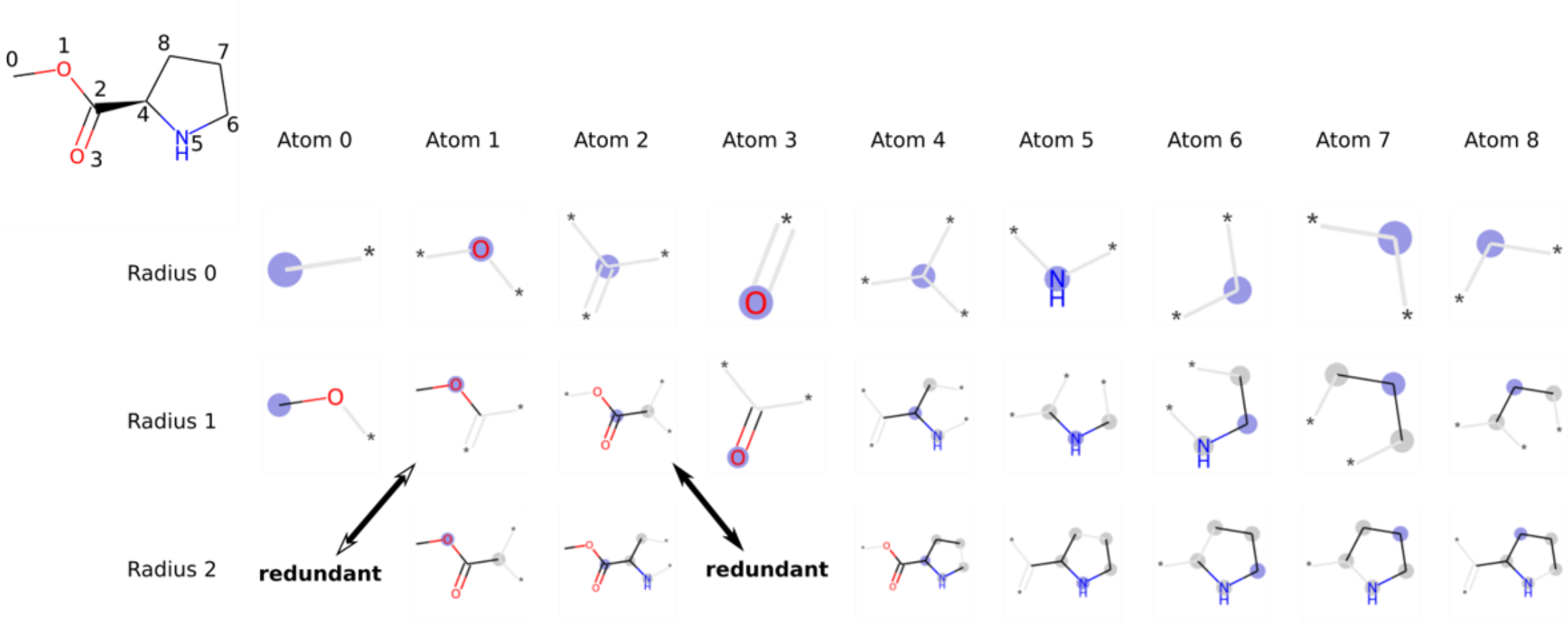
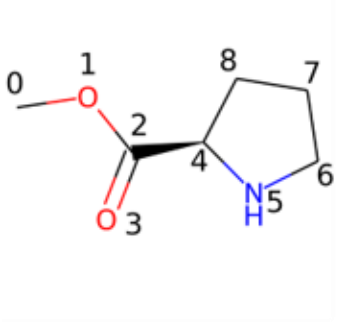


Figure reference: *ACS Omega* 2022, 7, 22, 19030–19039

Fingerprints example: construction of ECFP

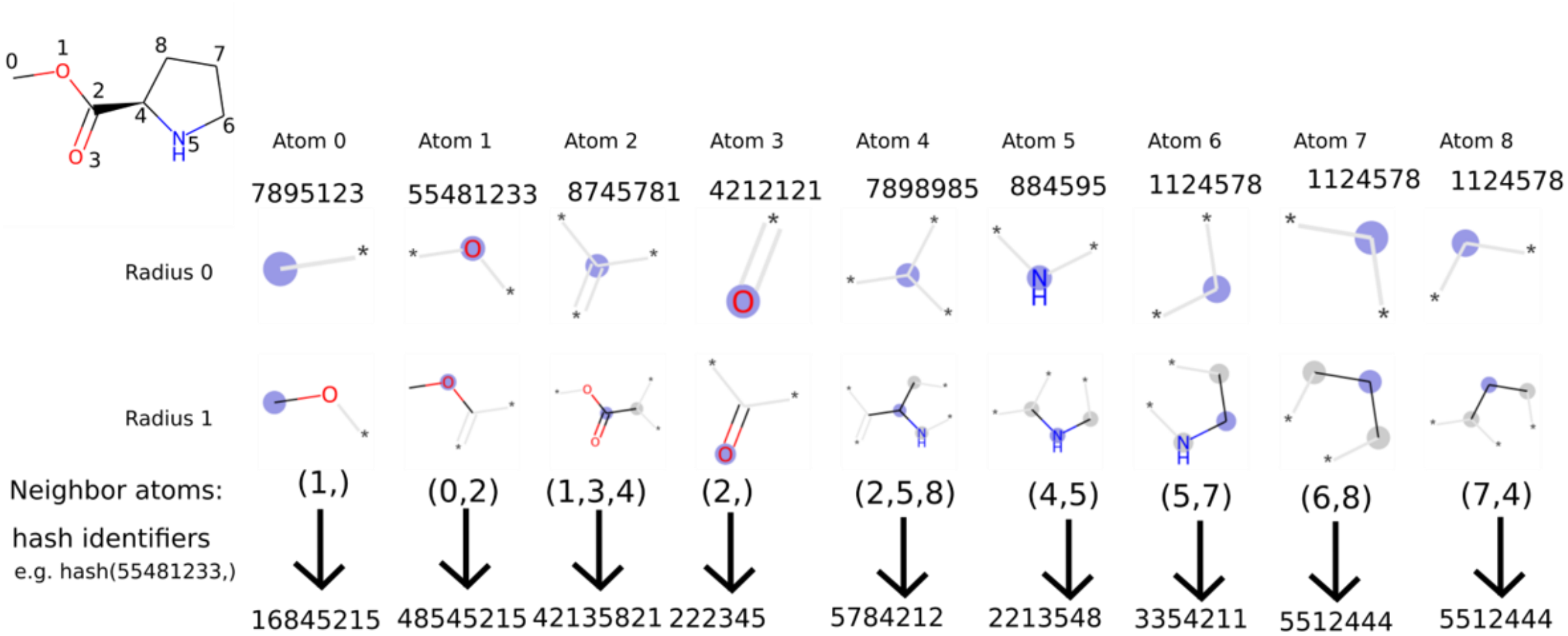


Fingerprints example: construction of ECFP

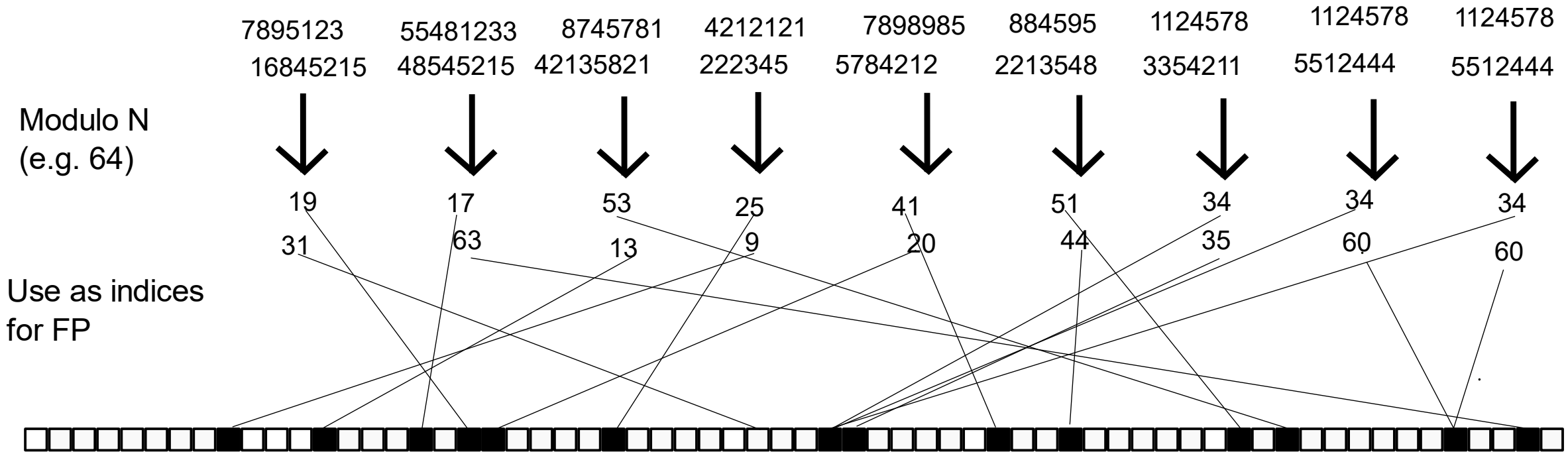


	Atom 0	Atom 1	Atom 2	Atom 3	Atom 4	Atom 5	Atom 6	Atom 7	Atom 8
Radius 0									
Atom properties:	(6,1,3)	(8,2,0)	(6,3,0)	(8,1,0)	(6,3,1)	(7,2,1)	(6,2,2)	(6,2,2)	(6,2,2)
Apply hash function	↓	↓	↓	↓	↓	↓	↓	↓	↓
32-bit integer	7895123	55481233	8745781	4212121	7898985	884595	1124578	1124578	1124578

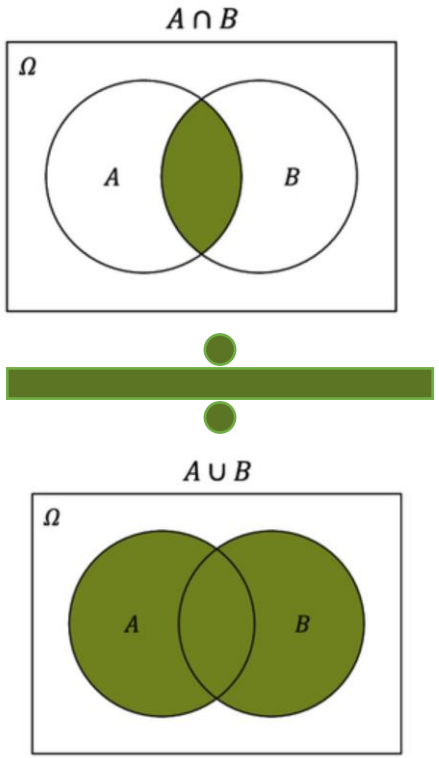
Fingerprints example: construction of ECFP



Fingerprints example: construction of ECFP



Quantitative similarity



- Tanimoto similarity
  - "features in common divided by total features"
- Euclidean distance
  - "distance in Euclidean space"
- Cosine distance
  - "their dot product divided by the product of their magnitudes"

Comparing apples to oranges using Tanimoto Similarity

	Color	Edible	Fruit	Grow climate	Shape	Main export country	Skin
	red	yes	yes	moderate	Round	China	Smooth
	green	yes	yes	moderate	Non-round	China	Smooth
	orange	yes	yes	hot	Round	Egypt	Rough

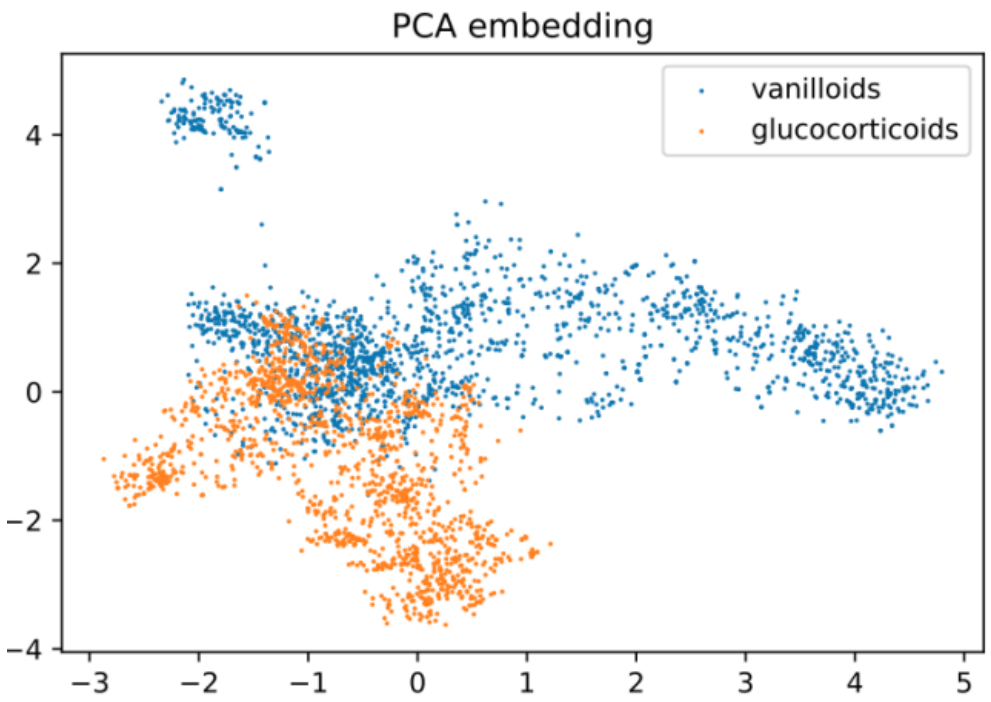
	1	5/9	3/11
	0.56	1	2/12
	0.27	**0.17**	1

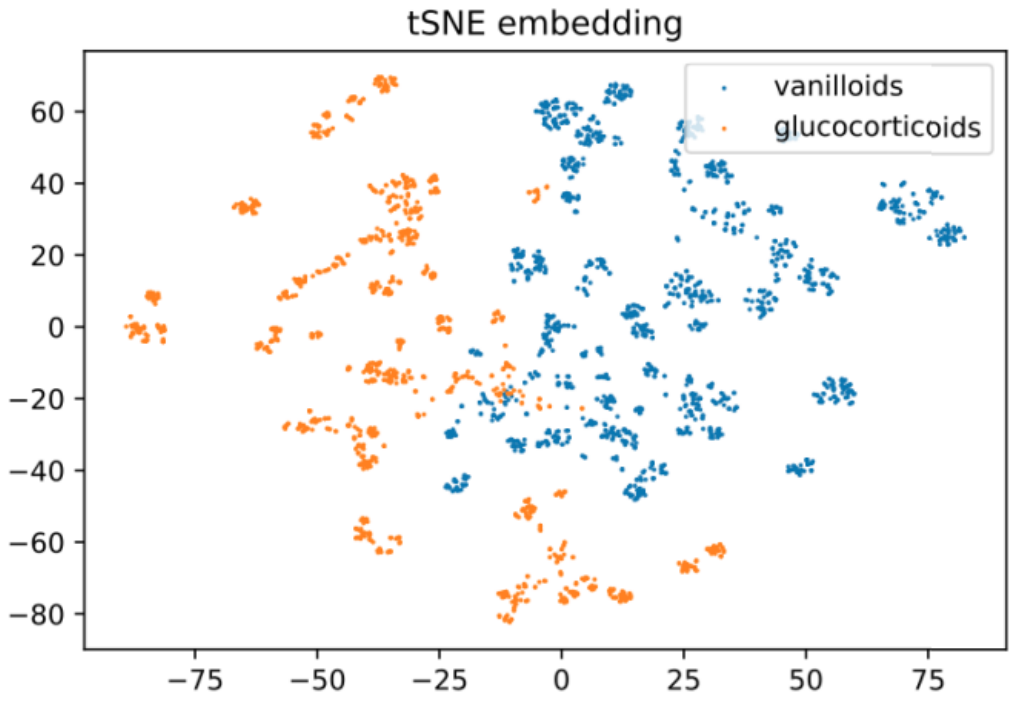
Apples are more similar to pears (0.56) than they are to oranges (0.27)
Oranges are more similar to apples (0.27) than they are to pears (0.17)
Apples and pears form a cluster!

Chemical space(s)

- All possible structures existing under given criteria (heavy atoms, druglikeness, synthesizability,...)
- Very vast (10^20 to 10^60)
- Visualization: PCA, MCA (multiple correspondence analysis),tSNE, various other unsupervised learning based techniques
- Exploration of chemical space: molecular optimization!

Chemical space representation



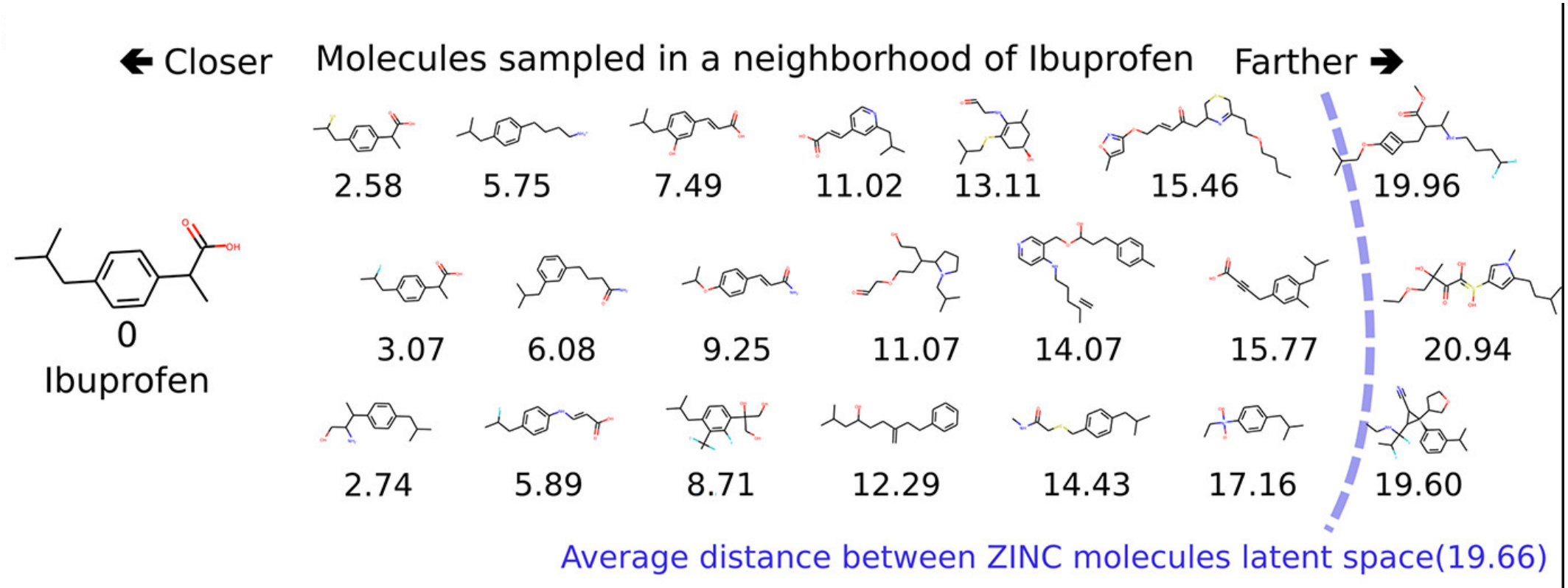


PCA vs tSNE (ECFP6_2048)

Exploring Chemical Space

- Molecular interpolation:
  - MOLPHER
  - MolVAE

Figure reference: *ACS Cent. Sci.* 2018, 4, 2, 268–276



Part 2: Molecular optimization

Molecular optimization

- Enhance the desired properties, and diminish the undesired properties of a molecules by directed exploration of similar molecules

Properties to optimize

- "ADME(T)" - pharmacokinetics
  - Absorption
  - Distribution
  - Metabolism
  - Elimination
  - (T)oxicity (incl. of metabolites)

Properties to optimize

- Activity - Pharmacodynamics
  - Binding energy
  - Assay activity
  - Ki, Kd, EC50, IC50
  - Host-guest affinity more generally

Properties to optimize

- Synthesizability and cost of production
  - Expert assessment
  - SAScore
  - Price prediction (QS$R)

Properties to optimize

- General physicochemical properties
  - Molecular weight
  - Lipophilicity

Properties to optimize

- Steric/spatial properties
  - Space complementarity to binding site
  - Also cavities in MOFs, Zeolites etc

Molecular optimization is a multiparameter optimization

- Lipinski Rule of 5
- QED: quantitative estimation of drug-likeness
- LogP and activity are correlated
- Descriptors (incl. those in QED) are often correlated

Molecular optimization strategies

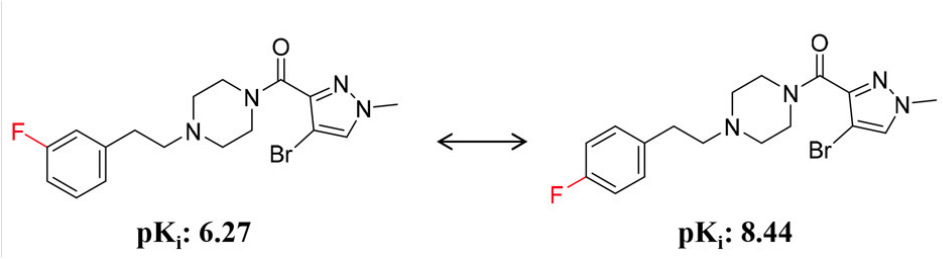
- Molecular optimization is a movement through chemical space
- Directed by feedback (models, measurements)
- Assumption of smooth path
- Activity landscapes:
  - Continuous
  - Discontinuous
  - Heterogenous

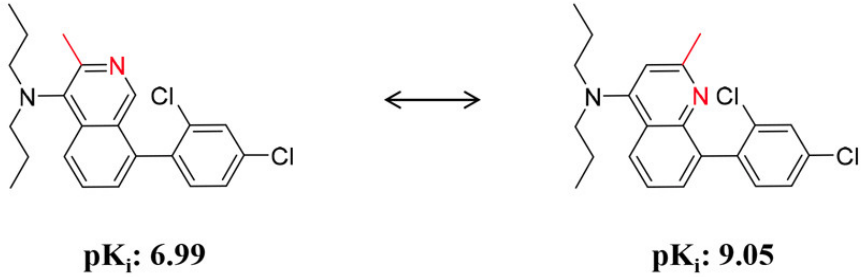
Virtual screening and optimization

- QSAR models:
  - Predict properties such as activity, toxicity, solubility, …
- Docking:
  - Validate structure-based theories
  - Ranking (unreliable but with enrichment)
- FEP, MM-GBSA:
  - Ranking (more reliable than docking)

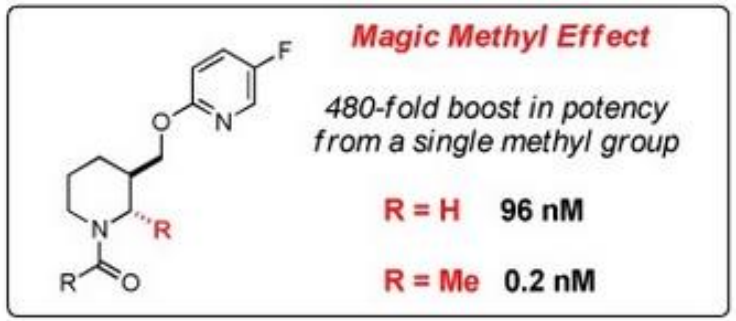
Optimization discontinuities

- Activity cliffs:





ACS Omega 2019, 4, 11, 14360–14368



Angewandte Chemie, Volume52, Issue47,
November 18, 2013

Bio-isosteric replacement

- Chemical substructures that can (sometimes) be substituted for each other while retaining the same biological activity
- Underlying reason is often steric and electronic

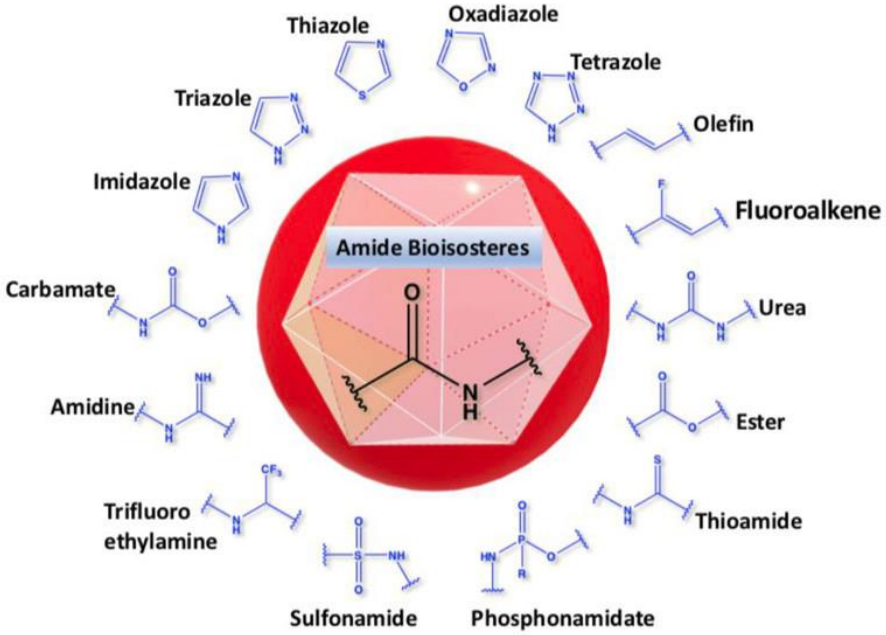
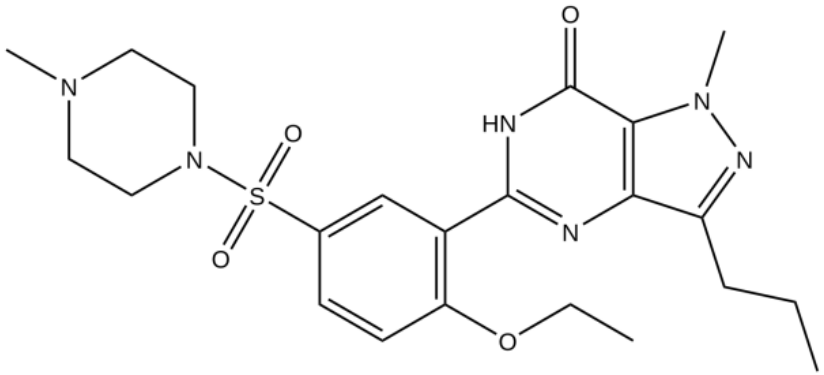


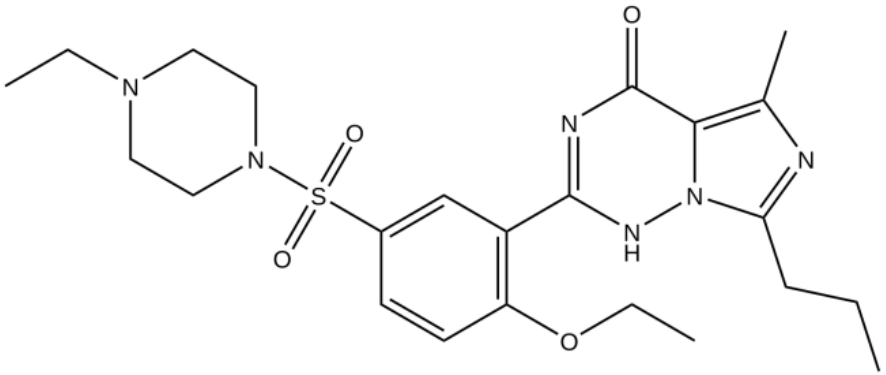
Figure taken from *J. Med. Chem.* 2020, 63, 21, 12290-12358

Scaffolds

- Molecular core structure that gets decorated with substituents
- Defined at various levels of coarseness
- Scaffold hopping:
  - Replace the core structure but retain activity
  - Find "dissimilar" actives



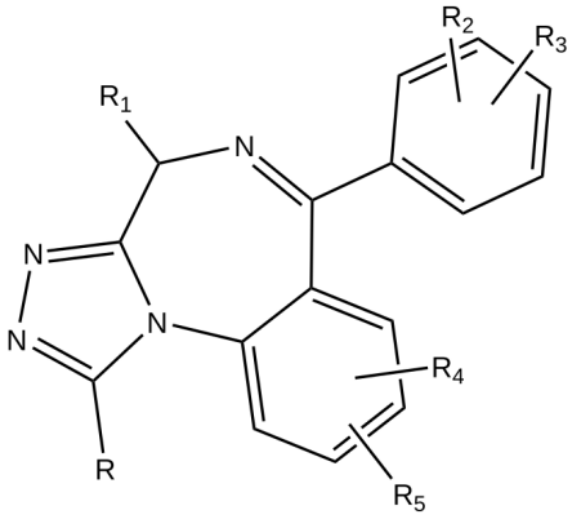
Sildenafil



Vardenafil

Patentability

- Markush structures
- Bio-isosteric replacements and scaffold hopping make it possible to explore non-patented chemical space
- They also allow search more dissimilar, more novel chemical space in an efficient way



R is selected from the group consisting of hydrogen, alkyl of 1 to 3 carbon atoms, inclusive, phenyl, benzyl and -COOR' in which R' is alkyl of 1 to 4 carbon atoms, inclusive;
R_1 is selected from the group consisting of hydrogen and alkyl of 1 to 3 carbon atoms, inclusive;
R_2, R_3, R_4 and **R_5** are selected from the group consisting of hydrogen, alkyl of 1 to 3 carbon atoms, inclusive, halogen, nitro, cyano, trifluoromethyl, and alkoxy, alkylthio, alkylsulfinyl, alkylsulfonyl, alkanoylamino and dialkylamino in which the carbon chain moieties are of 1 to 3 carbon atoms, inclusive;

Conclusion

- To optimize a molecule in the direction we want, we need good, quantitative similarity metrics
- To have a good similarity metric we need to pick meaningful features
- These features can form the basis of more advanced modelling