PubChem

Based on:
https://en.wikipedia.org/wiki/Catch_Me_If_You_Can

On generating cluster hierarchies based on Mol2Vec embeddings of compounds from large databases

Practical Course – Summer Term 2021 – Daniyal Kazempour

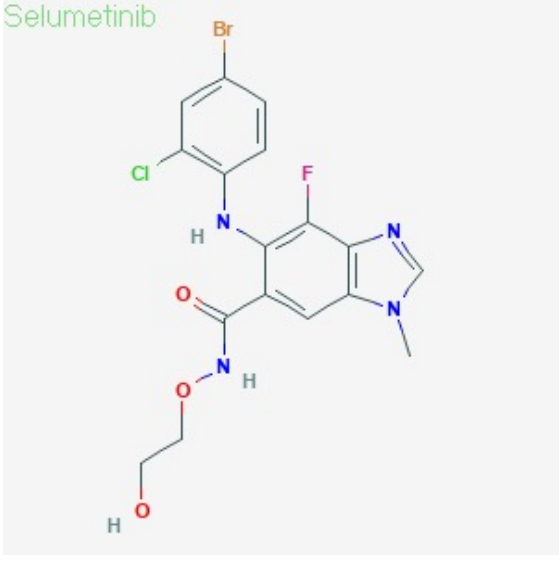
The PubChem Database

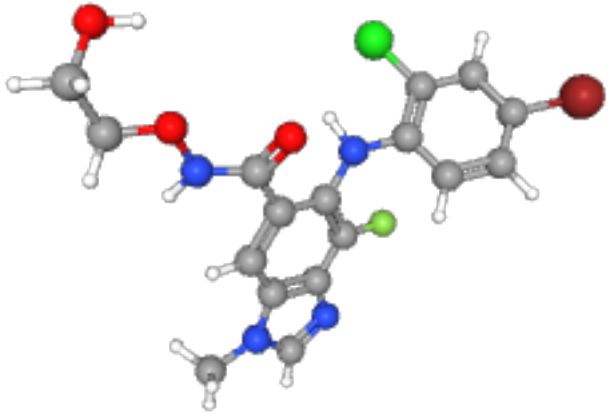
- Open database of chemical content
- Managed by the National Center for Biotechnology Information (NCBI)
- Around 111 Mio. validated compounds

Structure-data Files - SDF

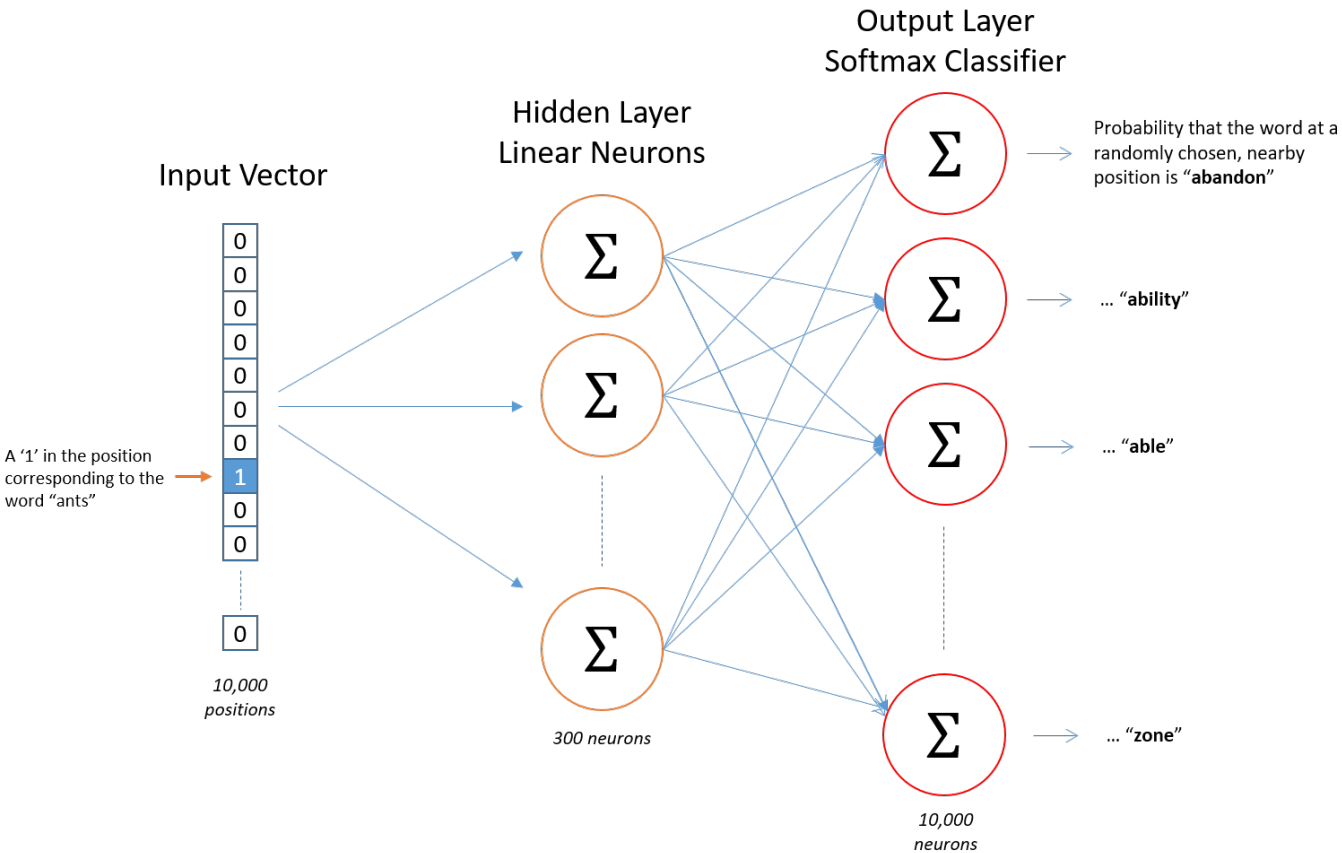
```
> <PUBCHEM_EXACT_MASS>
456.00001

> <PUBCHEM_MOLECULAR_FORMULA>
C17H15BrClFN4O3

> <PUBCHEM_MOLECULAR_WEIGHT>
457.7

> <PUBCHEM_OPENEYE_CAN_SMILES>
CN1C=NC2=C1C=C(C(=C2F)NC3=C(C=C(C=C3)Br)Cl)C(=O)NOCCO

> <PUBCHEM_OPENEYE_ISO_SMILES>
CN1C=NC2=C1C=C(C(=C2F)NC3=C(C=C(C=C3)Br)Cl)C(=O)NOCCO

> <PUBCHEM_CACTVS_TPSA>
88.4

> <PUBCHEM_MONOISOTOPIC_WEIGHT>
456.00001

> <PUBCHEM_TOTAL_CHARGE>
0

> <PUBCHEM_HEAVY_ATOM_COUNT>
27
```

- Contains information including
  - atoms
  - bonds
  - Properties like charge, mass…
  - Connectivity and coordinates of a molecule





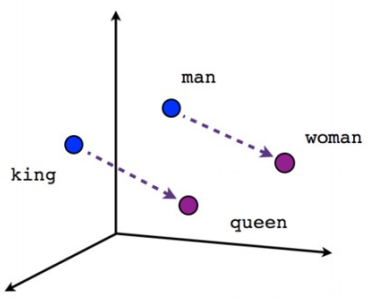
word2Vec Embeddings…*embedded* in a nutshell



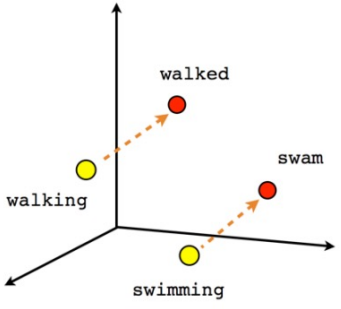
- Neural network with one hidden layer
- Use hidden weights as word embeddings
- Architecture similar to autoencoder → large input vector compressed to smaller dense vector
- But we have no decoding to original input vector
- Instead output is: probabilities of target words

Source: https://israelg99.github.io/2017-03-23-Word2Vec-Explained/

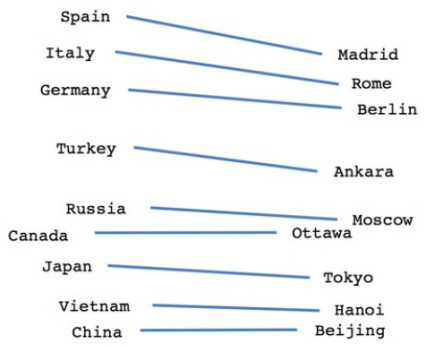
word2Vec Embeddings…*embedded* in a nutshell



Male-Female



Verb tense

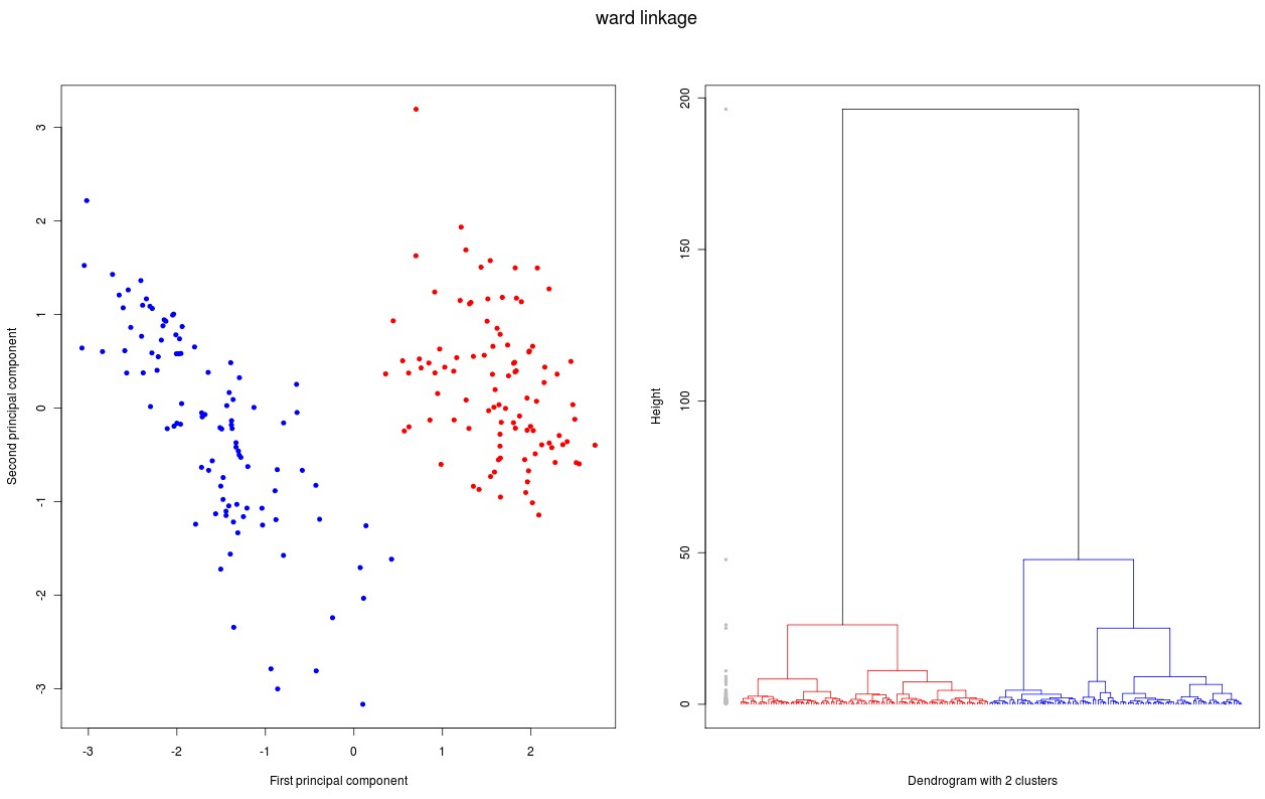


Country-Capital

- Underlying idea: Word2Vec learns similar word vectors for words that occur in similar context
- Can capture different ‚views' of similarity
- Quality can be controlled with different practices like:
  - Sub-sampling
  - Dimensionality
  - Context window size
- Many more *2Vec methods emerged like: Node2Vec, Gene2Vec, Prot2Vec, **Mol2Vec**, Job2Vec, Process2Vec…even Emoji2Vec (no kidding…)
- More thorough elaborations will be provided on a „on-demand" basis in this practical or in the "Advanced Data Mining and Machine Learning" module ☺

Source: https://israelg99.github.io/2017-03-23-Word2Vec-Explained/

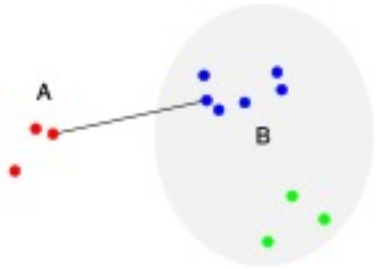
RECAP: Hierarchical Clustering



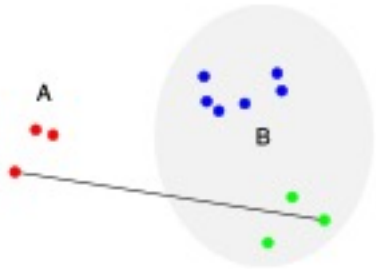
- Aims to yield **hierarchy** of clusters
- Two general types w.r.t. their operational level:
  - **Agglomerative** aka ‚bottom-up‘
  - **Divisive** aka ‚top-down‘
- Height of dendrogram represents distances
- A cut in the dendrogram at a specific distance yields a hard cluster assignment
- In non-optimized version it requires
  - $\mathcal{O}(n^3)$ in runtime and
  - $\mathcal{O}(n^2)$ in memory

https://de.wikipedia.org/wiki/Hierarchische_Clusteranalyse

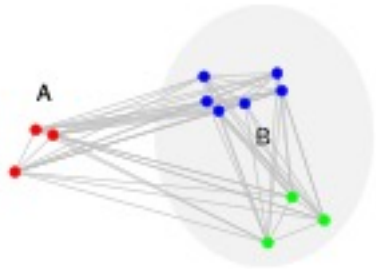
RECAP: Hierarchical Clustering: Distance functions



- Single-link → min distance



- Complete-link → max distance



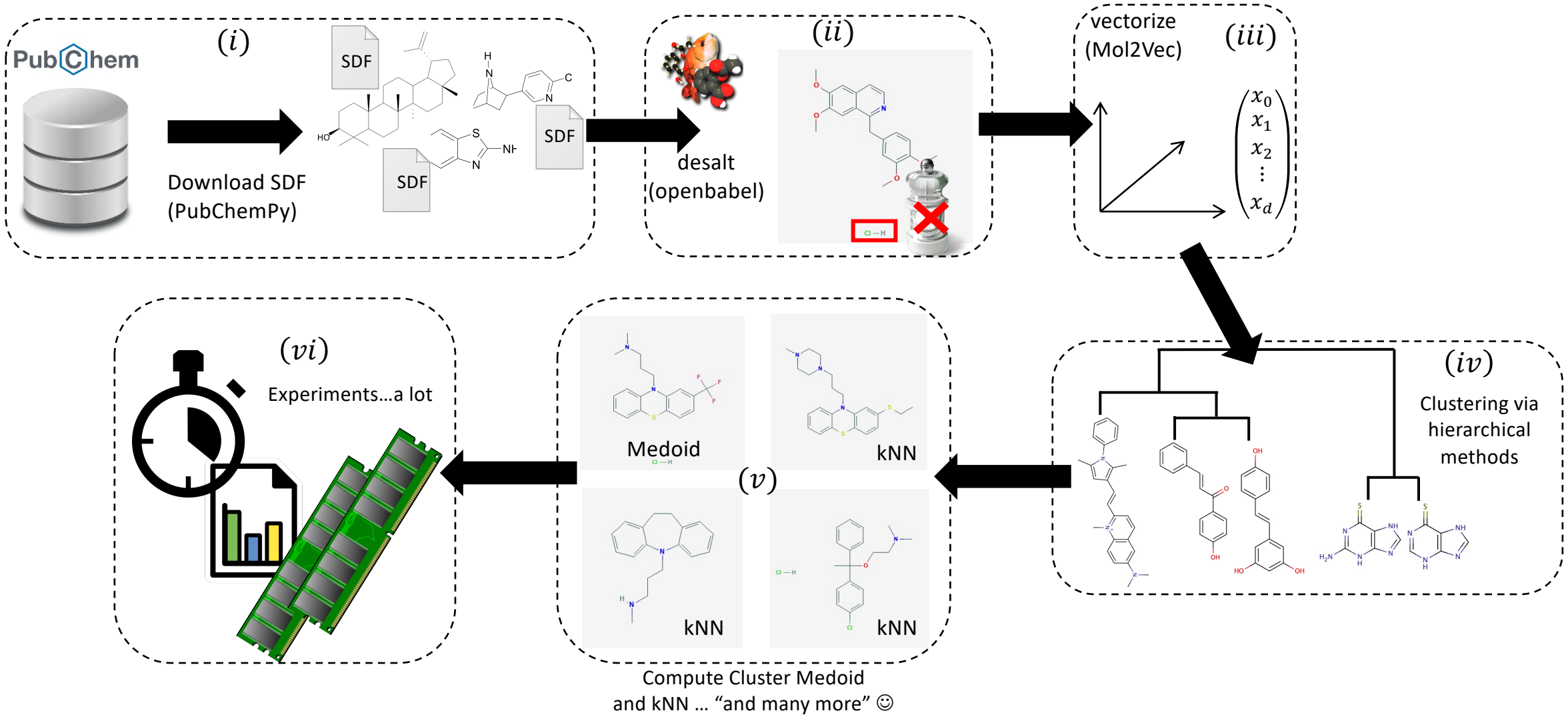
- Average-link → average of distances

https://de.wikipedia.org/wiki/Hierarchische_Clusteranalyse

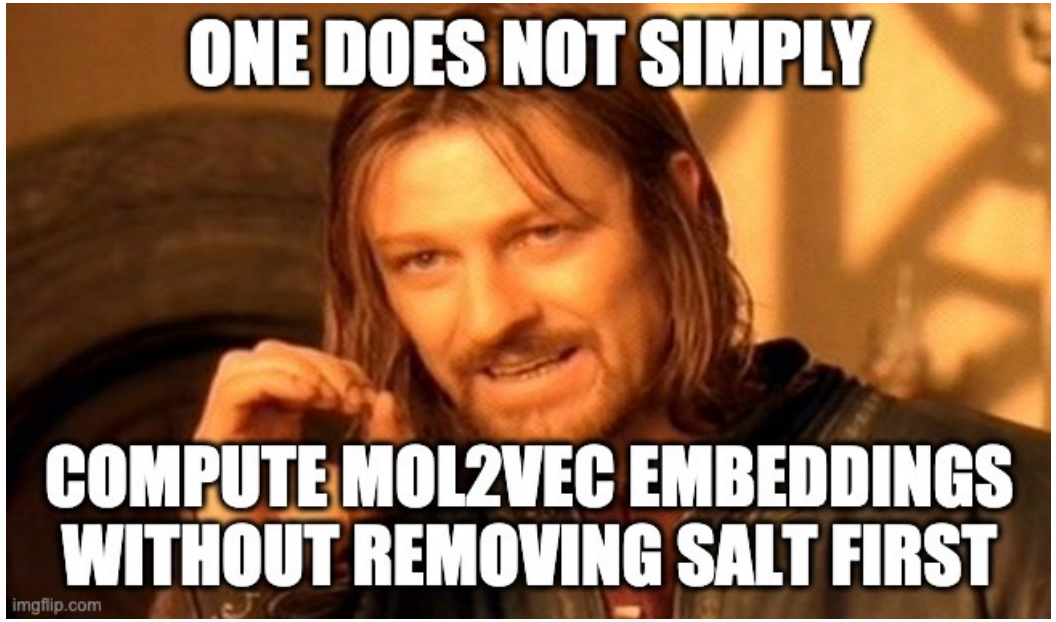
Why not using already existing methods?

- Existing clustering methods on compounds rely on a-priori knowledge of domain experts
- This knowledge may be (a) faulty or (b) biased and is (c) labour intense
- The *learning of the embeddings* and the *clustering* itself is **unsupervised**

Roadmap for project **E3CH**: **E**mbedding-based **C**omputation of **C**ompound **C**luster **H**ierarchies



What is ‚desalting' and why should we care?



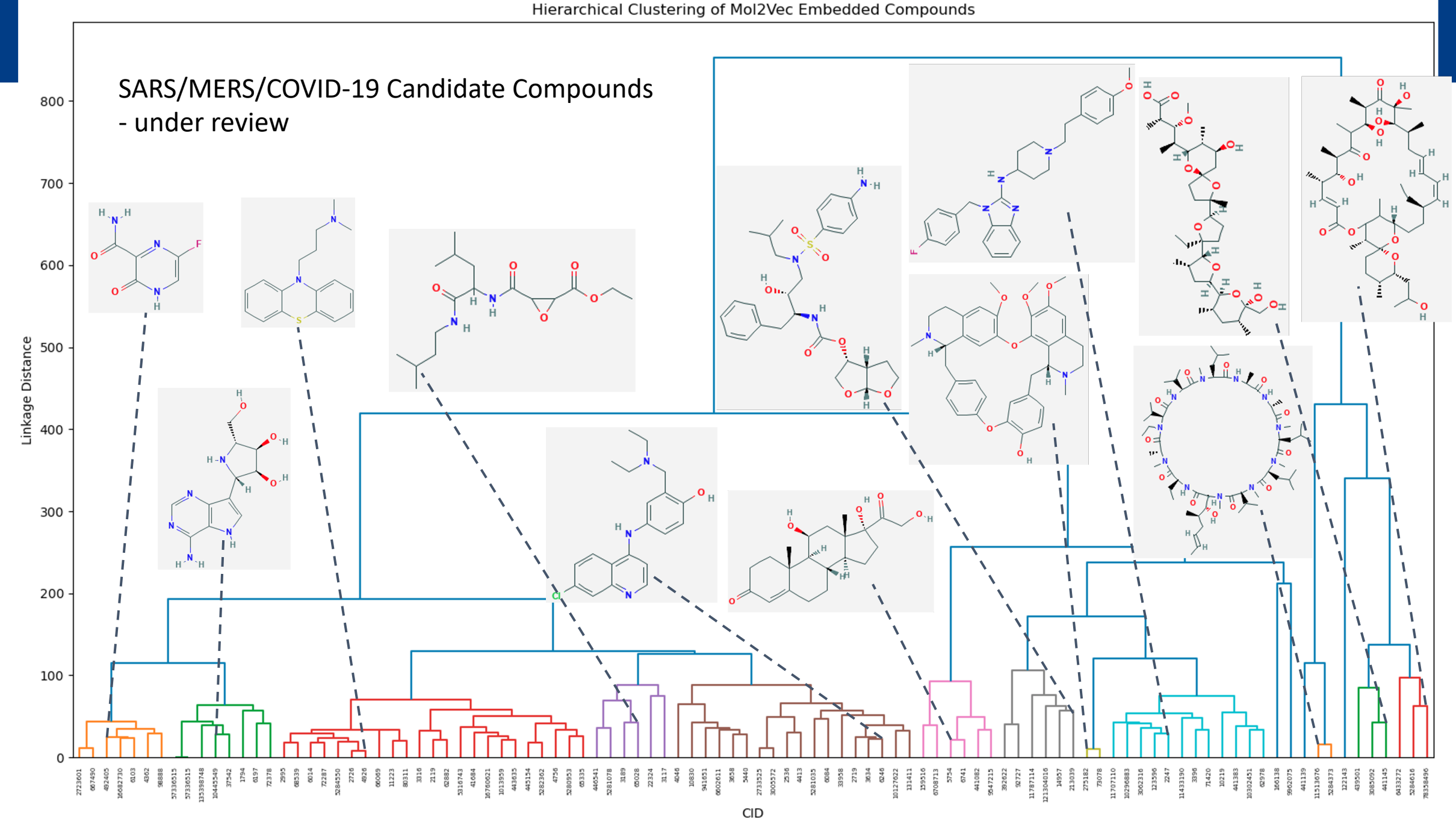
- Many compound entries in PubChem contain counter-ions or acidic or base compounds which are not relevant for the properties of the compound
- Mol2Vec could learn embeddings, based on these ‚add-ons' if not removed → „**bias"**
- Openbabel library provides functions to eliminate them

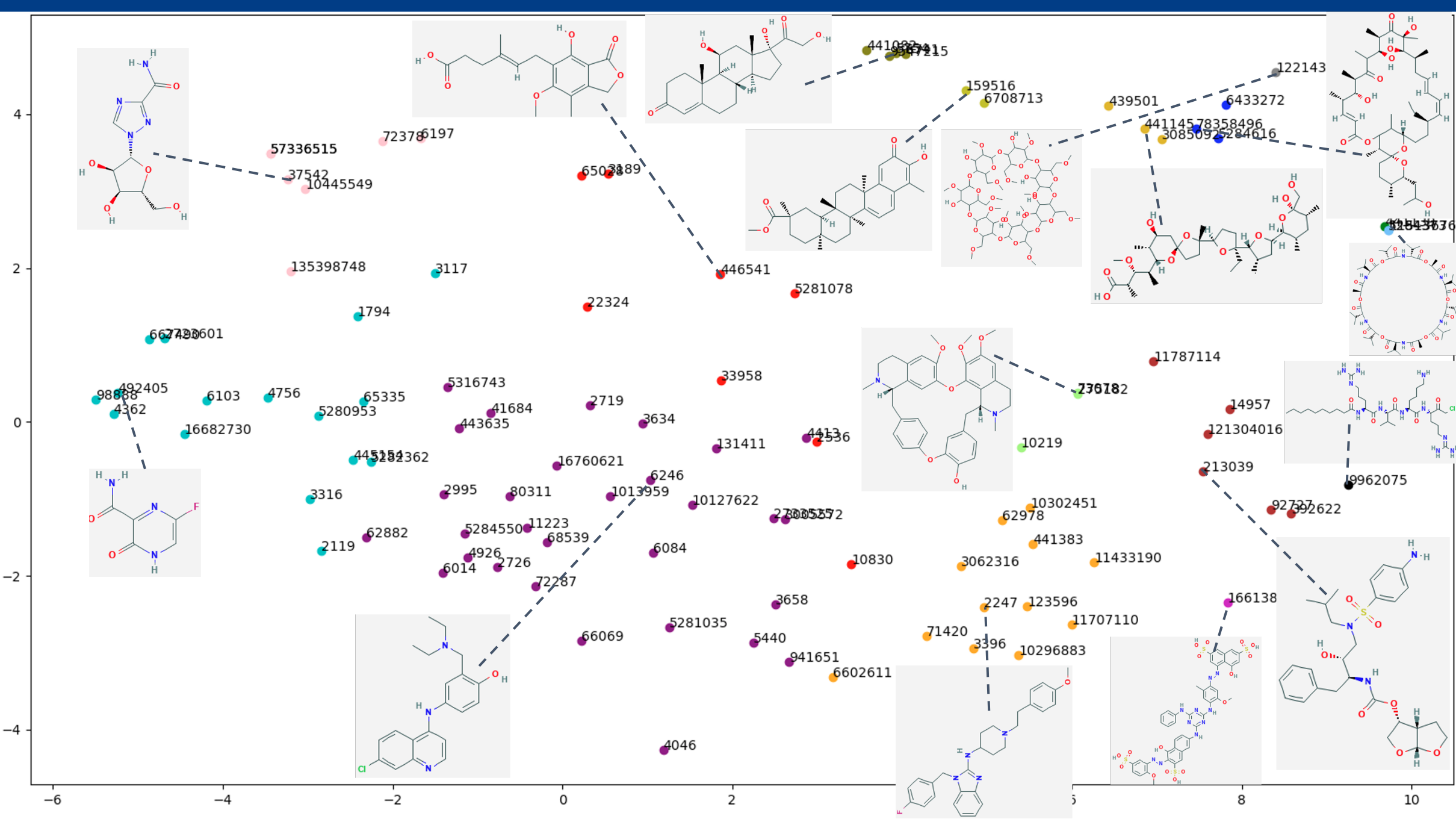
What the heck is that good for what we are doing?

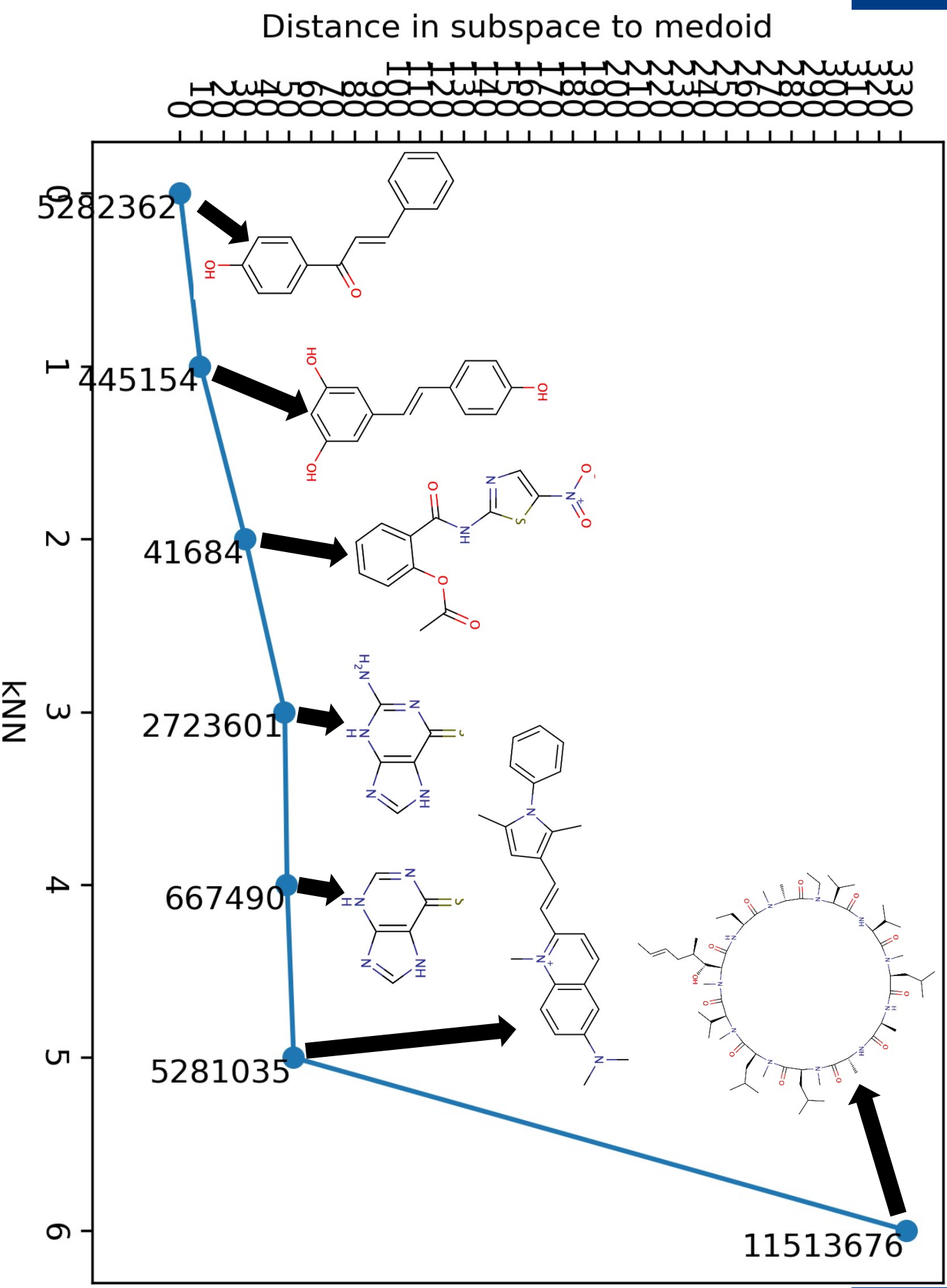
- During virtual screening it is of paramount interest to:
  - achieve a *high coverage of the chemical space*
  - while keeping *number of compounds to be screened to a minimum*

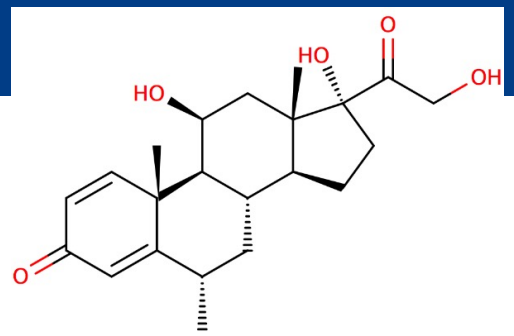
- Having the representative compounds (i.e. Medoids) reduces the computational costly screening efforts

- Finding *non-obvious yet still similar compounds* to a target compound → Clustering hierarchies

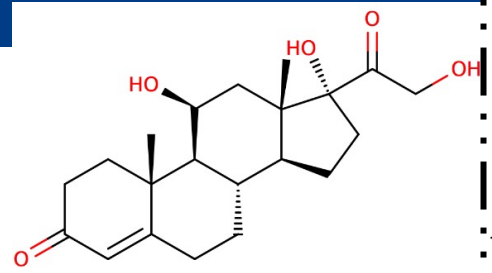

Hierarchical Clustering of Mol2Vec Embedded Compounds
SARS/MERS/COVID-19 Candidate Compounds
- under review
Linkage Distance
CID
0
100
200
300
400
500
600
700
800


4
2
0
−2
−4
−6
−4
−2
0
2
8
10
57336515
72378
6197
37542
10445549
650288
3389
135398748
3117
446541
5281078
22324
1794
6627428
27433601
492405
98868
6103
4362
4756
65335
5280953
16682730
5316743
41684
443635
2719
3634
33958
131411
4413
2536
16760621
6246
3316
2995
80311
1013959
10127622
5284550
11223
62882
2119
68539
4926
6014
2726
6084
72287
10830
3062316
11433190
3658
5281035
66069
5440
941651
6602611
4046
159516
6708713
439501
6433272
122143
441145
78358496
308509
5284616
11787114
14957
121304016
10219
213039
9962075
10302451
62978
441383
11707110
71420
3396
10296883
166138
2247
123596

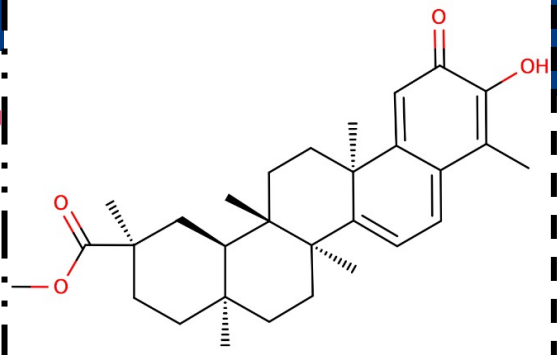

Distance in subspace to medoid
kNN
0
1
2
3
4
5
6
5282362
445154
41684
2723601
667490
5281035
11513676



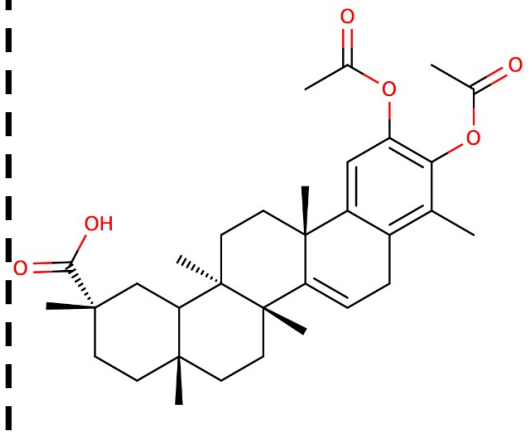
Hydrocortisone



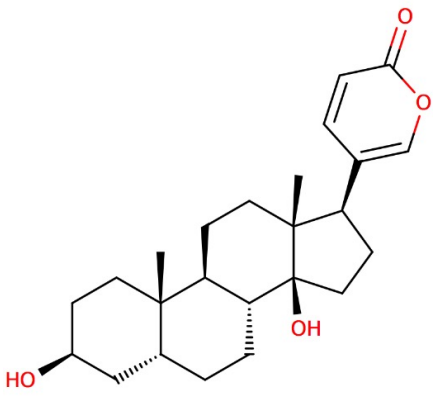
Methylprednisolone



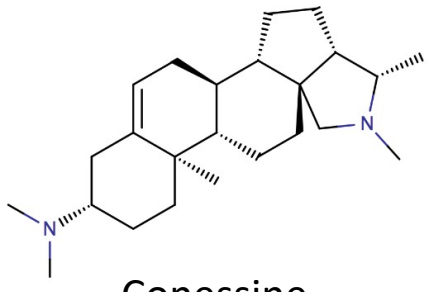
Pristimerin



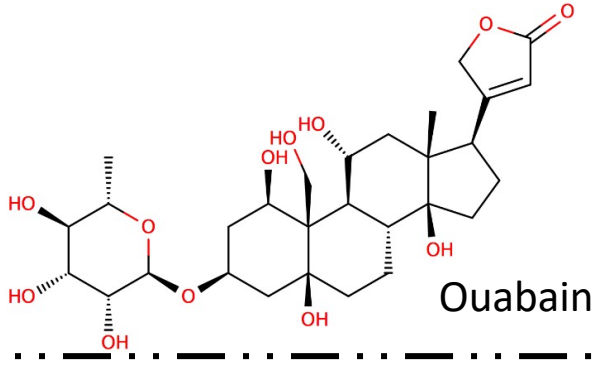
Dihydrocelastryl



Bufalin



Conessine



Ouabain

k-means,
Hier. Clustering

ORCLUS

Code, data and results available on Github:

SCAN ME

Stipulated Subtasks

1) Perform **embedding** of all about 111 million compounds from PubChem database
2) Computer **dendrogram** from hierarchical clustering
3) Compute for each (sub)cluster the **medoid compound**
4) Compute for each (sub)cluster the **variance** (increase...)
5) Runtime experiments:
    1) Run one **experiment** for a query compound q brute-force against all compounds in sequential order
    2) Run one **experiment** for q using the hierarchy with medoids and variance

Overall challange:
Embedding and clustering of the entire 111 Mio compounds of PubChem and studying the applicability for compound search acceleration.

United we research!
current and *upcoming* collaborators

Dept. of Chemistry
Domain expert:
Structural Biochemistry

Dept. of Computer Science
Domain „expert“:
High Dimensional Clustering

Dept. of Computer Science
Domain expert:
Relational Databases and Index Structures

Dept. of Statistics
Domain expert:
Manifold Learning

Epilogue

Compounds, compounds in the database
Let me tell you how much I am amazed
By how you differ from others in the set
Showing us your friends and foes we haven‘t met

Know yourself, know others, be known

- What are your strengths, meaning:
  things you need low efforts and a maximum motivation doing it?

- In what are you currently not as proficient as you would be,
  but want to rise/improve?

- What do you expect from yourself? (Punctuality, Communication,…)

- What do you expect from others?

Know yourself, know others, be known

- What do I expect from you?
  - That you learn from this practical as much as you can
  - That you work together as one team (this takes time and sometimes also disputes → you will master them, I got faith in you ;-))
  - That you blend in your own ideas. If there is something you think may be cool: tell us your idea (I‘m excited when people bring in their own minds) and start hacking ☺
  - That you take care of yourself! Take a break; If you need councelling, approach me!

What remains?

- Need a „jour fix“ per week or bi-weekly to synchronize about the current status → propose vote for a day and time

- Which questions and expectations do you have?