The speech of structural patterns

- application of linguistic word-association methodologies onto binary fingerprints and structure keys

Ivan Čmelo[1*], Daniel Svozil[1]
[1]*Laboratory of Informatics and Chemistry, UCT Prague, Technická 5, CZ-166 28 Prague 6, Czech Republic (*cmeloi@vscht.cz)*

ZINC15

285 732 863 compounds
157 914 301 uniq. InChIKeys

PubChem

91 221 617 compounds
69 081 967 uniq. InChIKeys

ChEMBL

1 666 863 compounds
1 512 302 uniq. InChIKeys

DRUGBANK

6 768 compounds
6 496 uniq. InChIKeys

378 628 111 compounds
213 777 358 uniq. InChIKeys

Corpus of chemical structures

Joined structures from all sets by their inChIKeys

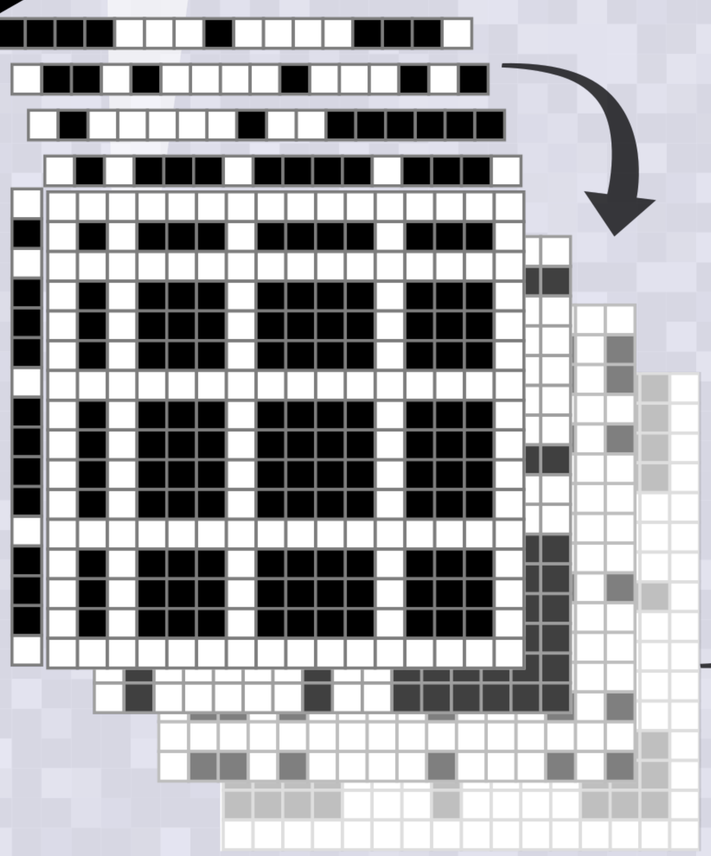
Fingerprinting
Morgan [512|1024] bit, [2|3] diameter

PubChem Substructure Fingerprint

Set of named SMARTS structures "FGFP"

MACCS keys

Fingerprints



Own outer products (k⊗k)

How many times do the (words/patterns) occur together within the same (sentence/molecule) from a given (text corpus/chemical database)?

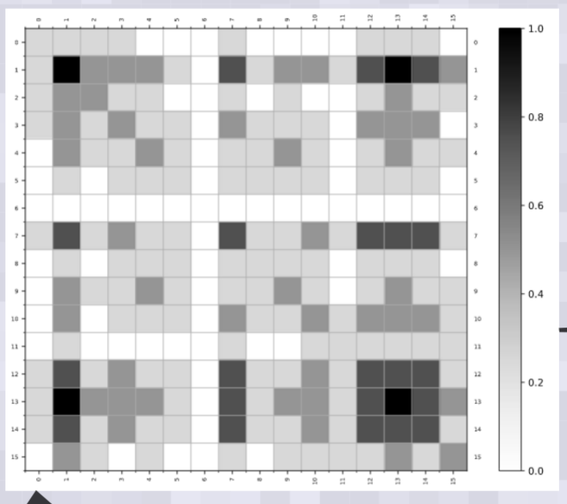
Co-occurrence

1	1	1	1	0	0	0	1	0	0	0	0	1	1	1	0
1	4	2	2	2	1	0	3	1	2	2	1	3	4	3	2
1	2	2	1	1	0	0	1	0	1	0	0	1	2	1	1
1	2	1	2	1	1	0	2	1	1	1	0	2	2	2	0
0	2	1	1	2	1	0	1	1	2	1	0	1	2	1	1
0	1	0	1	1	1	0	1	1	1	1	0	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	3	1	2	1	1	0	3	1	1	2	1	3	3	3	1
0	1	0	1	1	1	0	1	1	1	1	0	1	1	1	0
0	2	1	1	2	1	0	1	1	2	1	0	1	2	1	1
0	2	0	1	1	1	0	2	1	1	2	1	2	2	2	1
0	1	0	0	0	0	0	1	0	0	1	1	1	1	1	1
1	3	1	2	1	1	0	3	1	1	2	1	3	3	3	1
1	4	2	2	2	1	0	3	1	2	2	1	3	4	3	2
1	3	1	2	1	1	0	3	1	1	2	1	3	3	3	1
0	2	1	0	1	0	0	1	0	1	1	1	1	2	1	2

$$COMX(S) = \sum_{o=1}^{[S]} k_o . k_o^T$$

What is the **probability** of two (words/patterns) occurring together in a (sentence/molecule) randomly selected from a given (text corpus/chemical database)?

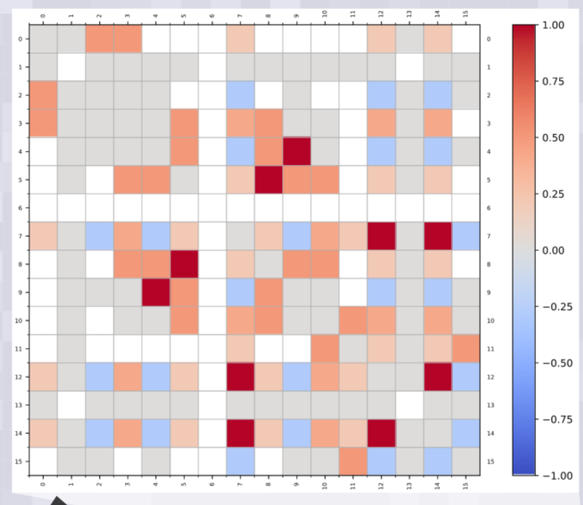
Probability



$$PMX(S) = \frac{COMX(S)}{[S]}$$

What is the amount of **interassociation** between individual (words/patterns) within a given (text corpus/chemical database?)

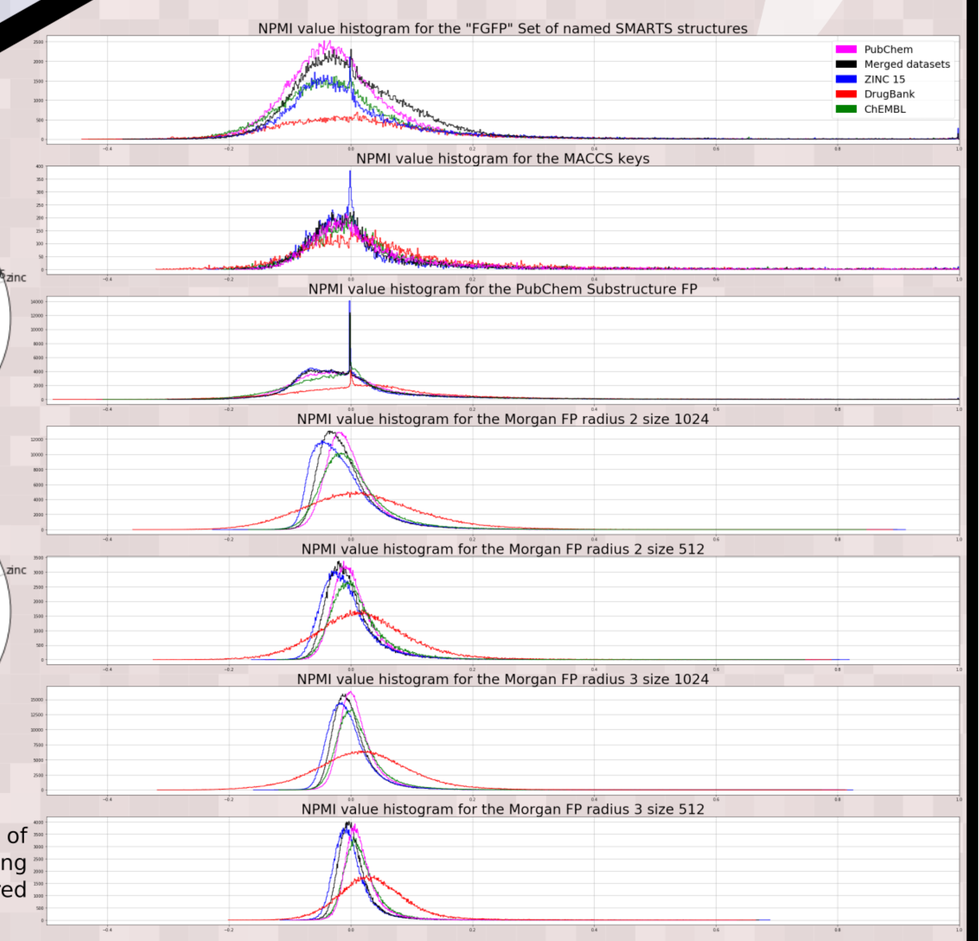
Normalized Pointwise Mutual Information



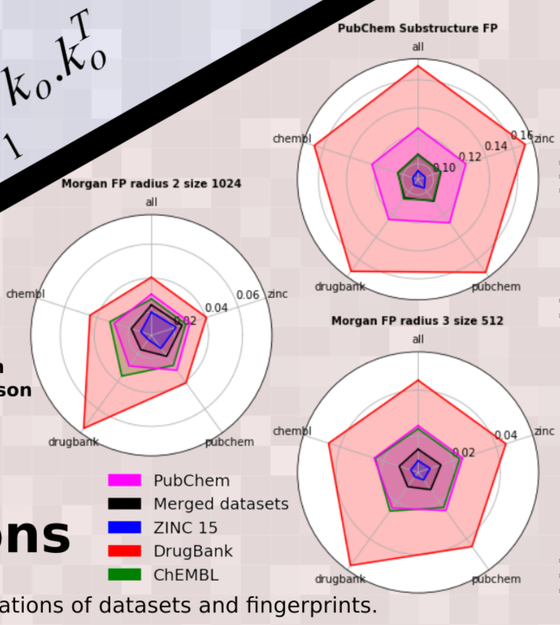
$$NPMI = \left(\log_2 \frac{p(a,b)}{p(a)p(b)} \right) \Big/ - \log_2 p(a,b)$$

NPMI value histograms

for all used sets and fingerprints



'Relative pattern tightness' comparison



Current conclusions

There are noticeable interrelations between bits in all combinations of datasets and fingerprints.

The bit interrelation values generally follow a normal-like distribution, just like word interrelations in linguistics.

Flor et al. used the NPMI-based word interrelation profile from a very large text corpus to quantify the 'lexical tightness' of word pairs within an arbitrary text. They used this 'lexical tightness' metric to build a working model to estimate reading difficulty of texts. The larger the average NPMI value was, the more conventional word pairings were present in the measured text. More conventional word pairings are associated with texts that are easier to read (lower grade level).

The closest thing we have to the linguistic text corpus are cheminformatic databases. We merged four of them to create a baseline set of chemical structures. Unlike the linguistic text corpus that covers a significant portion of current literature, the chemical databases contain only a small fraction of the chemical space. Therefore, we can't tell how globally conventional the pattern pairs within molecules are. We can, however, calculate the NPMI-based pattern interassociation profiles for arbitrary sets of molecules. These profiles characterize the strength of individual pattern relations within the given set, and can be compared to profiles of other sets or used to quantify the overall strength of interrelations within itself ('**inner pattern tightness**'). Also, any set of molecules can be evaluated using any pattern profile ('**relative pattern tightness**'), akin to word-pair evaluation of a text sample by a corpus NPMI profile.

UCT PRAGUE

Linguistic workflow source
M. Flor, B. B. Klebanov, and K. M. Sheehan, “Lexical Tightness and Text Complexity,”
Proc. 2th Work. Nat. Lang. Process. Improv. Textual Access., pp. 29–38, 2013.
... and its references

Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures"