Extended Similarity Indices:

Benefits of Comparing more than Two Objects Simultaneously. Theory, Speed, Consistency, and Diversity Selection

R. A. Miranda-Quintana[1], D. Bajusz[2], A. Rácz[2], K. Héberger[2]

[1]Univeristy of Florida

[2] HUN-REN Research Centre for Natural Sciences, Institute of Excellence, Hungarian Academy of Sciences

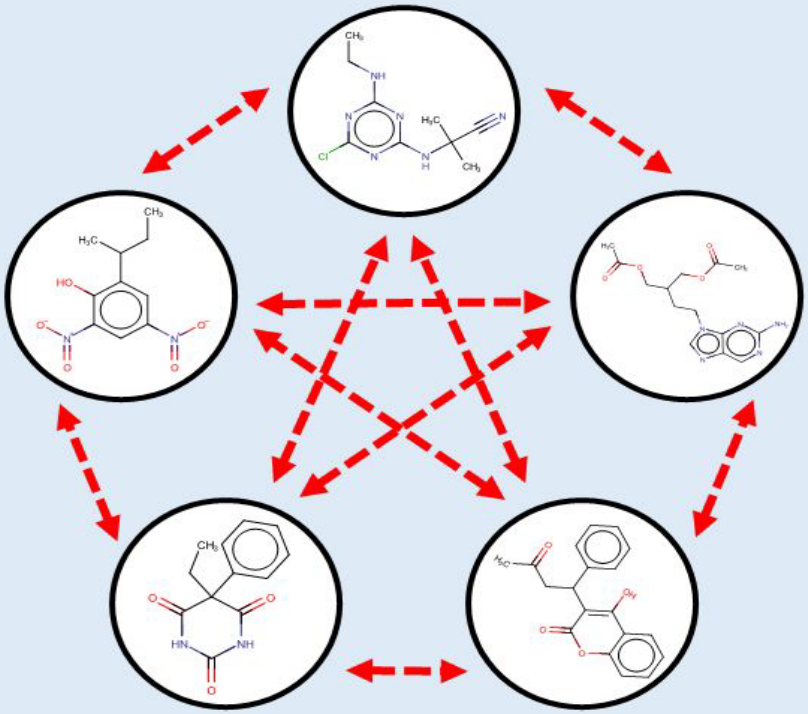
The quote

"None of the axioms employed by great generals [scientist] is difficult. Indeed, once they have been employed successfully, they reveal their innate simplicity and appear to be the obvious and sometimes only logical solution. Yet all great ideas are simple. The trick is to see them before others."

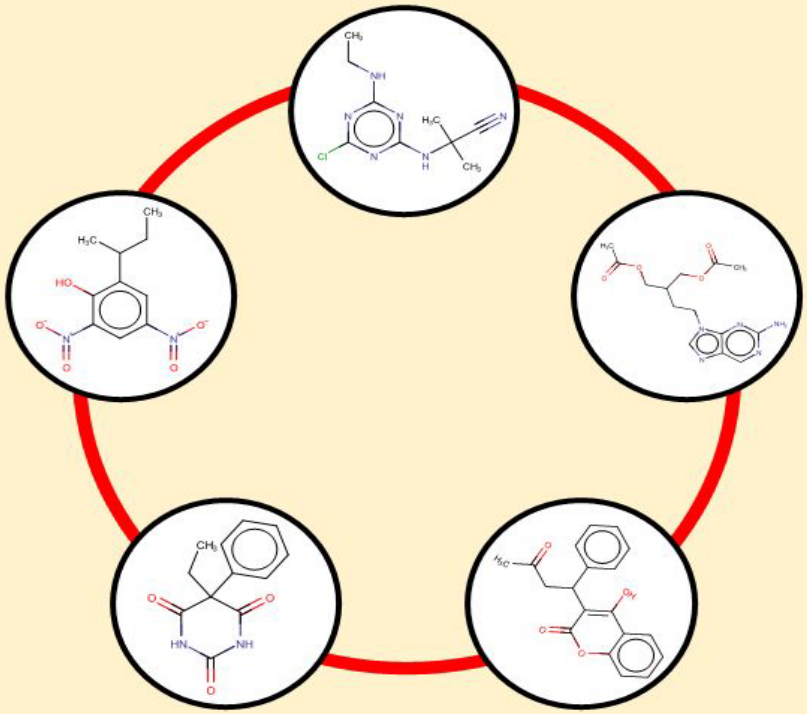
Bevin Alexander, How Great Generals Win, 1993

COMPUTE TIME

Binary similarity



Extended (*n*-ary) similarity



DIVERSITY

Definitions, terms

- Nominal scale — binary.
- Fingerprint: a bit string consisting of 0-s and 1-s.
- Sum of rank differences (SRD): a city block (Manhattan) distance of ranks.
- ANOVA (a classification procedure, comparison of means)
- Similarity of molecules (pairwise comparison) a number:

  S [0,1] r[-1,1] és $S=1-d$
- Distance (Dissimilarity, d): D[0,+∞];

  $$S(x,y) = 1/(1 + D(x,y))$$
- Maximum likelihood principle & central limit theorem

Bajusz *et al. Journal of Cheminformatics* (2015) 7:20
DOI 10.1186/s13321-015-0069-3

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?

Dávid Bajusz[1], Anita Rácz[2,3] and Károly Héberger[2*]

Rácz *et al. J Cheminform* (2018) 10:48
https://doi.org/10.1186/s13321-018-0302-y

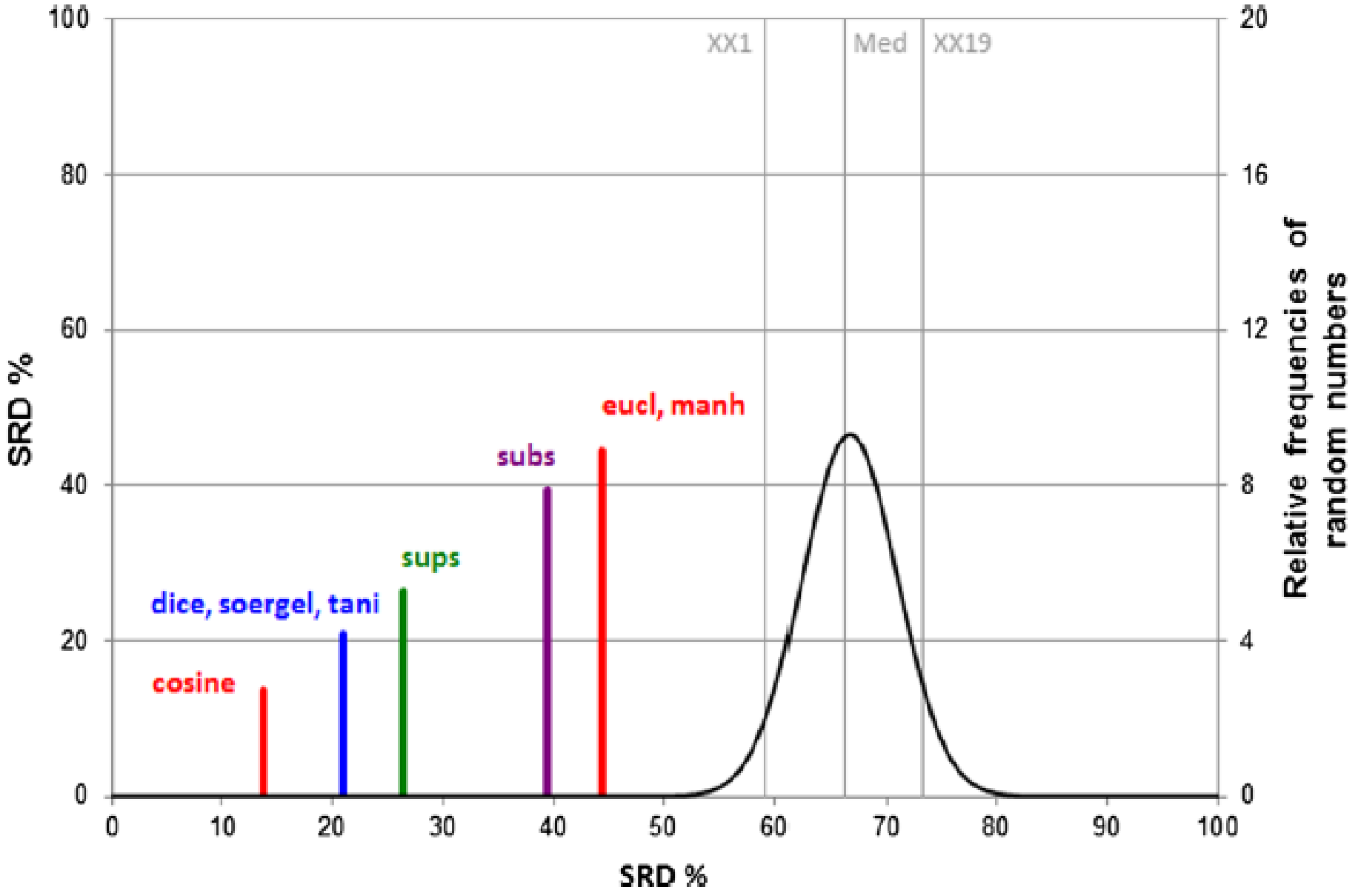
Journal of Cheminformatics

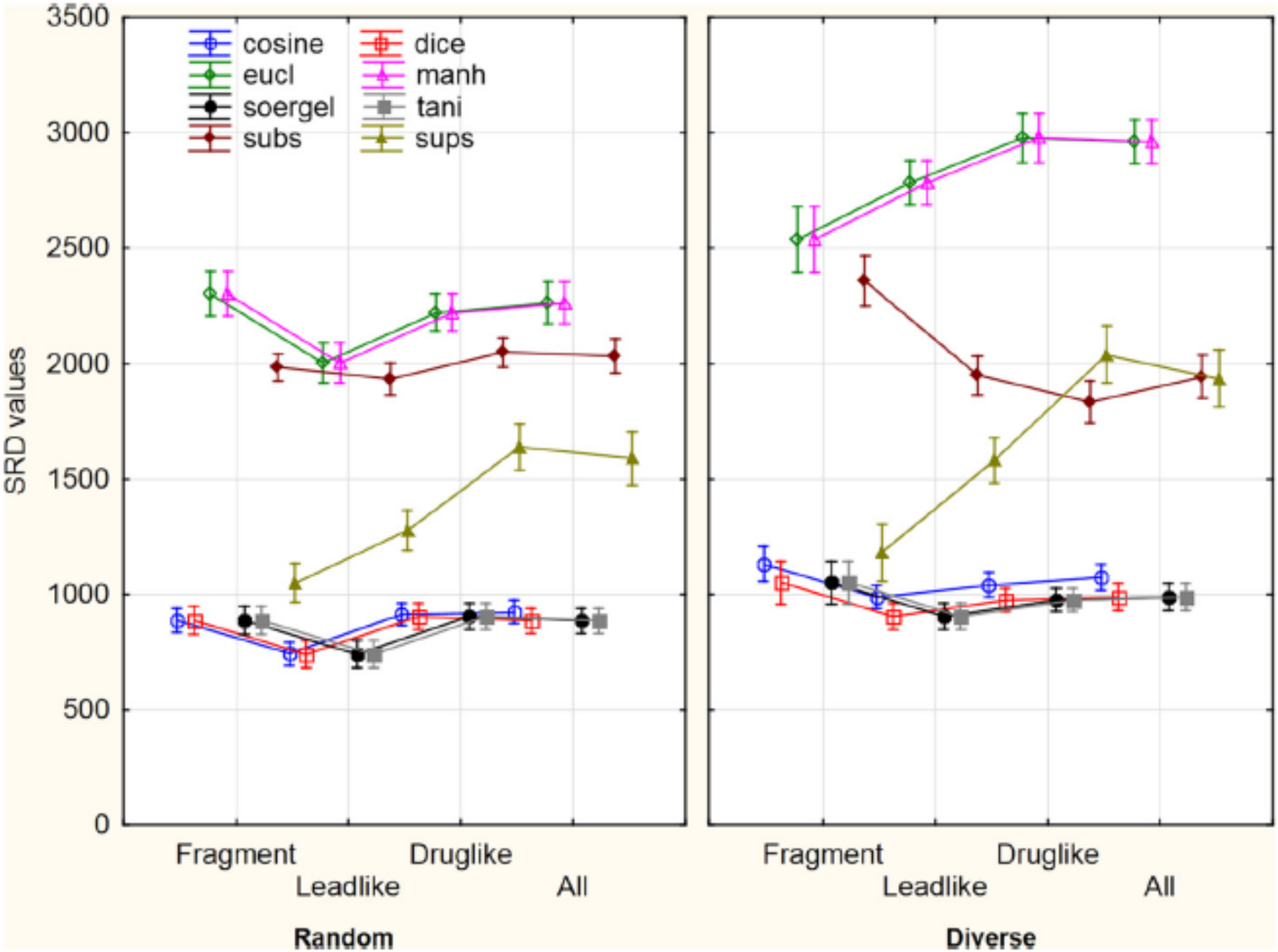
RESEARCH ARTICLE Open Access

Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints

Anita Rácz[1], Dávid Bajusz[2*] and Károly Héberger[1]

Ranking of similarity measures




cosine
dice
eucl
manh
soergel
tani
subs
sups
SRD values
3500
3000
2500
2000
1500
1000
500
0
Fragment
Druglike
Leadlike
All
Random
Fragment
Druglike
Leadlike
All
Diverse

Bajusz *et al. Journal of Cheminformatics* (2015) 7:20
DOI 10.1186/s13321-015-0069-3

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?

Dávid Bajusz[1], Anita Rácz[2,3] and Károly Héberger[2*]

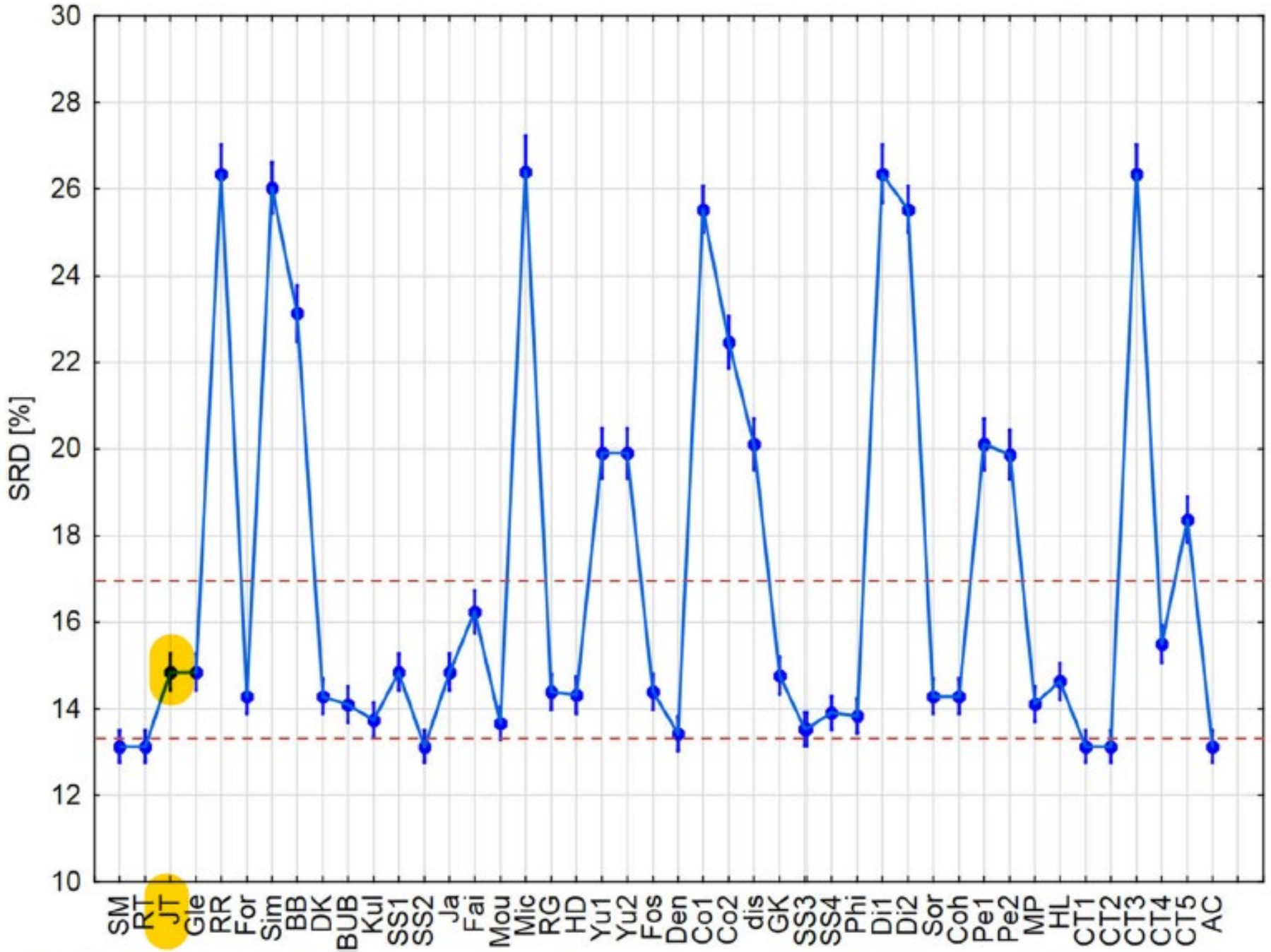
Rácz *et al. J Cheminform* (2018) 10:48
https://doi.org/10.1186/s13321-018-0302-y

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints

Anita Rácz[1], Dávid Bajusz[2*] and Károly Héberger[1]


SRD [%]
30
28
26
24
22
20
18
16
14
12
10
SM
RT
JT
Gle
RR
For
Sim
BB
DK
BUB
Kul
SS1
SS2
Ja
Fai
Mou
Mic
RG
HD
Yu1
Yu2
Fos
Den
Co1
Co2
dis
GK
SS3
SS4
Phi
Di1
Di2
Sor
Coh
Pe1
Pe2
MP
HL
CT1
CT2
CT3
CT4
CT5
AC

What is the basic question of research?

- Is it possible to express the similarity of molecules better?
- To create faster algorithms?
- Well, and if so, how?

Representation of molecules

- It is not unambiguous how to express the **similarity** of two molecules.
- **Bitwise representation** (encoding of representative information) has proved **successful**.
- The bit string of a molecule is called a **fingerprint**
- The similarity of the two bands shows the similarity of the molecules, *e.g.* **Tanimoto coefficient**

A	1	0	1	1	0	1	$\lvert A\rvert = 4$
B	1	1	0	1	0	0	$\lvert B\rvert = 3$
$A \wedge B$	1	0	0	1	0	0	$\lvert A \wedge B\rvert = 2$
$A \vee B$	1	1	1	1	0	1	$\lvert A \vee B\rvert = 5$

$$S_T(A, B) = \frac{2}{5}$$

BV1	BV2		Freq.
1	1	*a*	3
0	1	*b*	4
1	0	*c*	2
0	0	*d*	1
1	0		
0	1		
0	1		
1	1		
…	…		
0	1		
1	1		

$$Sokal - Mitchener$$

$$S = \frac{a + d}{a + b + c + d}$$

$$Jaccard - Tanimoto$$

$$S = \frac{a}{a + b + c}$$

Contingency table

$p = a + b + c + d$		2. molecule	
		1 (substructure present)	0 (no substructure)
1. molecule	1 (substructure present)	*a*	*b*
	0 (no substructure)	*c*	*d*

Extended Sokal-Michener index

$F_1 = (1\ 0\ 1\ 1\ 0\ 1\ 0\ 0)$

$F_2 = (0\ 0\ 1\ 0\ 0\ 1\ 0\ 1)$

- $F_3 = (1\ 0\ 1\ 1\ 1\ 0\ 0\ 1)$

$F_4 = (0\ 0\ 1\ 1\ 0\ 1\ 0\ 0)$

$F_5 = (0\ 0\ 1\ 1\ 0\ 1\ 1\ 0)$

$$S = \frac{a+d}{a+b+c+d}$$

- $C_{5(5)} = 1; C_{5(4)} = 2; C_{5(3)} = 0; C_{5(2)} = 2;$

$$C_{5(1)} = 2; C_{5(0)} = 1$$

$$\Delta_{5(k)} = |2k - 5|$$

Similarity=1 (if $2k - n > \gamma$), Similarity=0 (if $n - 2k > \gamma$), and dissimilarity (if $\Delta_{n(k)} \leq \gamma$) counter

Extended Sokal-Michener index

- $S_{eSM(1s\_wd)} = \frac{\sum_s f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_s f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)}}$
- $S_{eSM(1s\_d)} = \frac{\sum_s f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_s C_{n(k)} + \sum_d C_{n(k)}}$
- $f_s(\Delta_{n(k)}) = \frac{\Delta_{n(k)}}{n}$ és $f_d(\Delta_{n(k)}) = 1 - \frac{\Delta_{n(k)} - n \bmod 2}{n}$
- $\gamma = n \bmod 2$

Factorial ANOVA

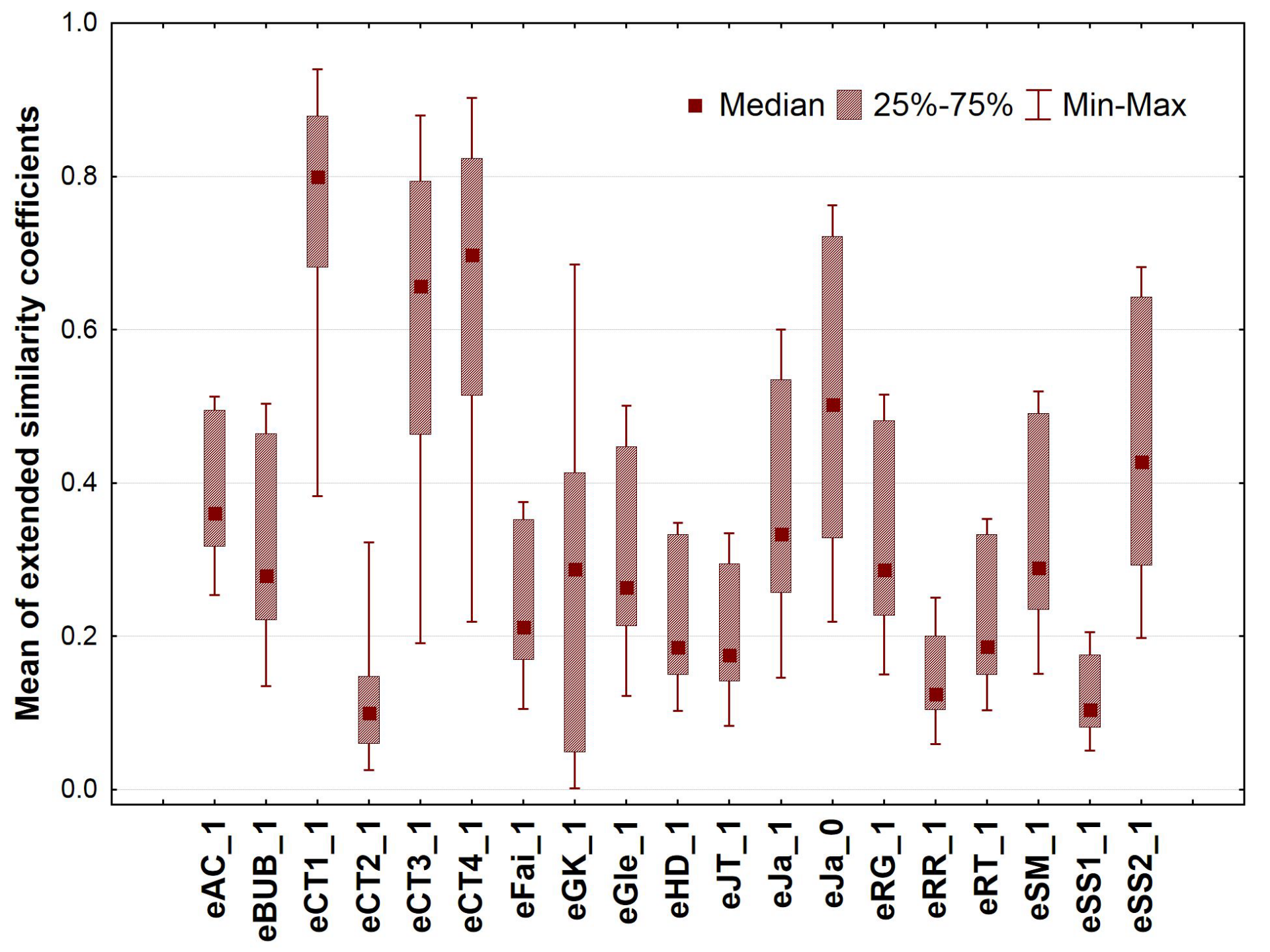
Factors taken into account:

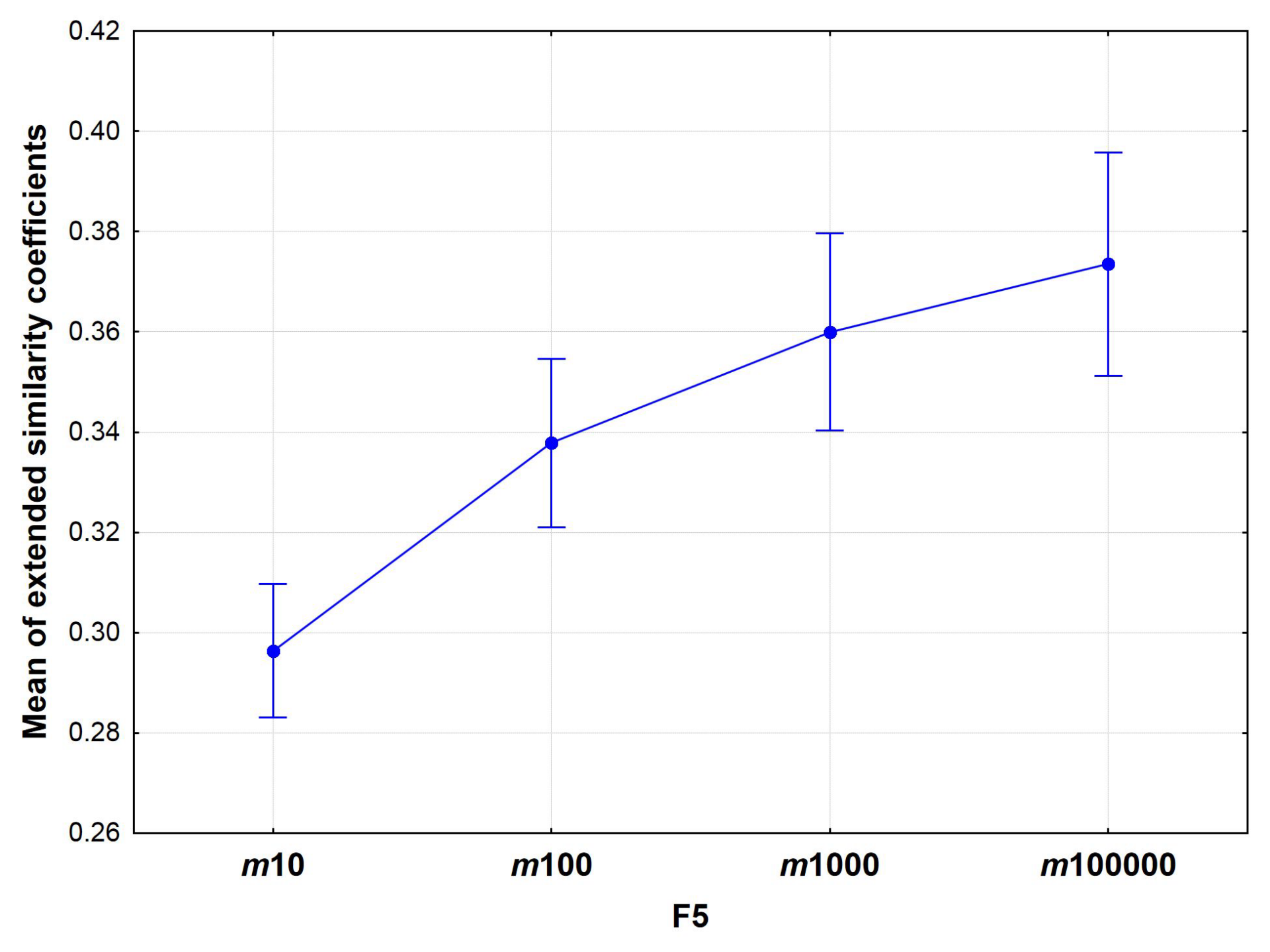
- F2: Number of comparisons. Number (*n*) of bitstreams (*e.g.* fingerprints) compared, 14 levels:

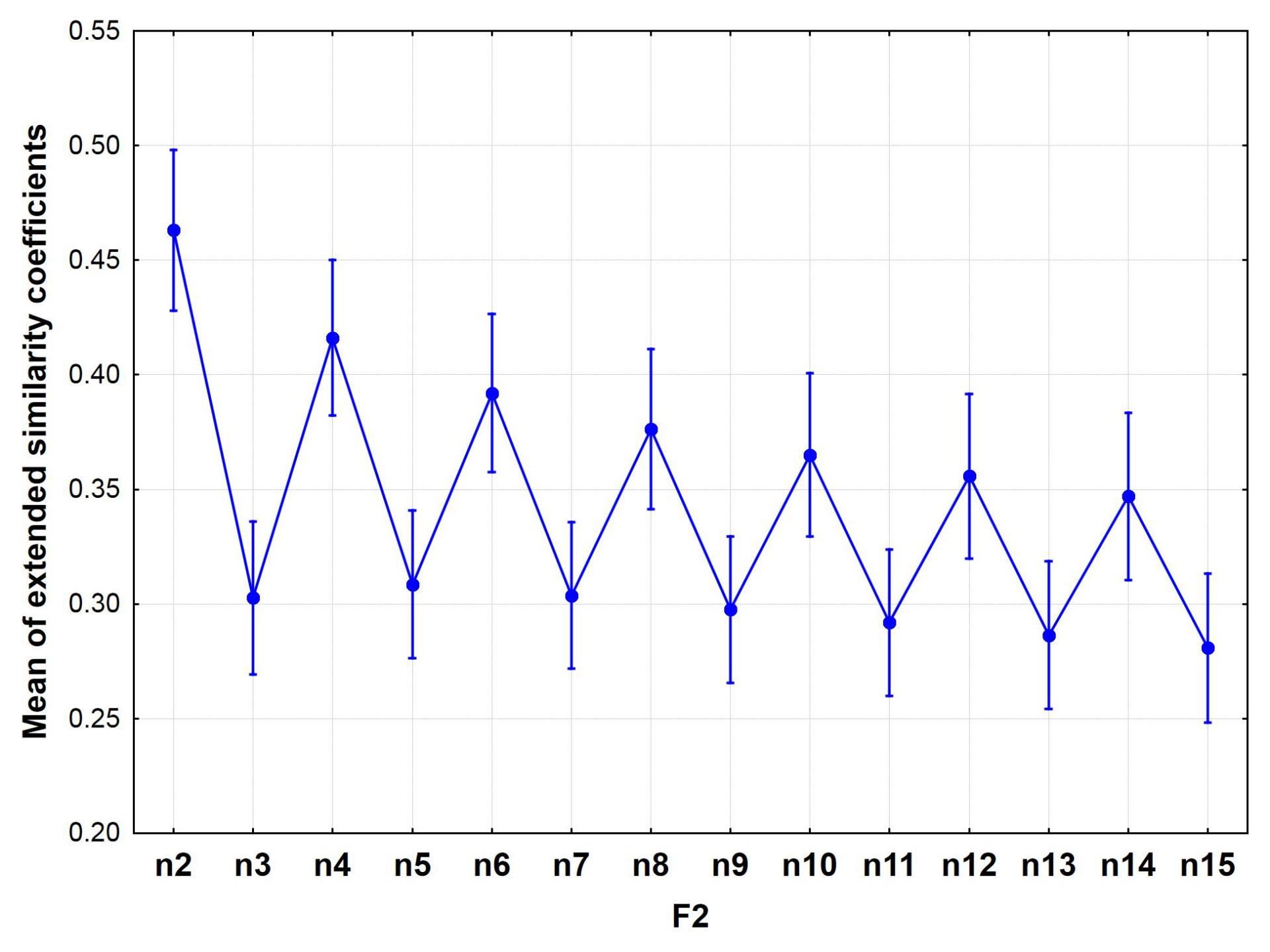
  n = 2, 3, ... 15 *n*-ary;

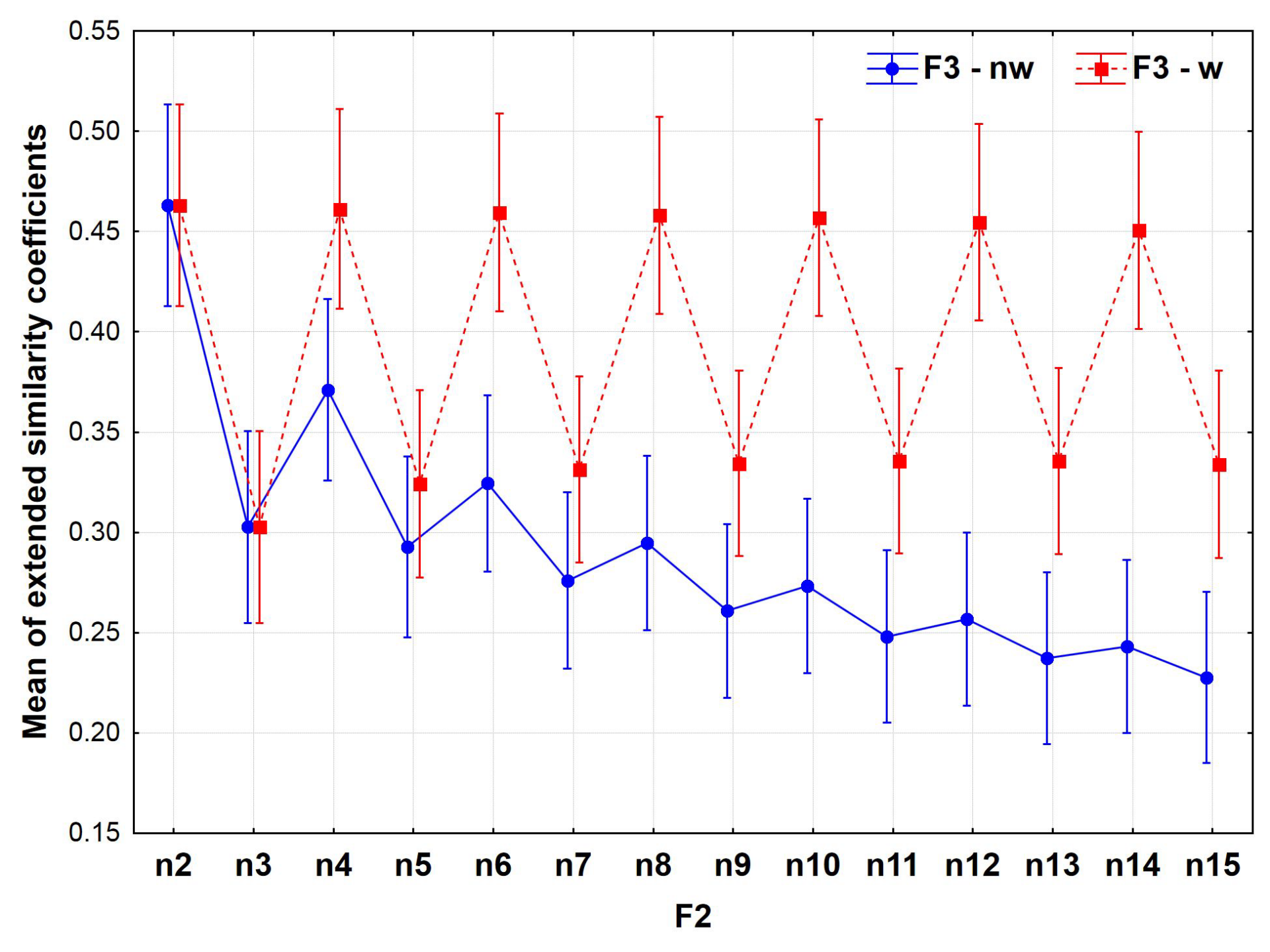
- F3: role of weighting, two levels: weighted and unweighted versions of the new similarity coefficients and
- F4: the similarity coefficients themselves, 19 levels
- F5: *m* - length of fingerprints, four levels:

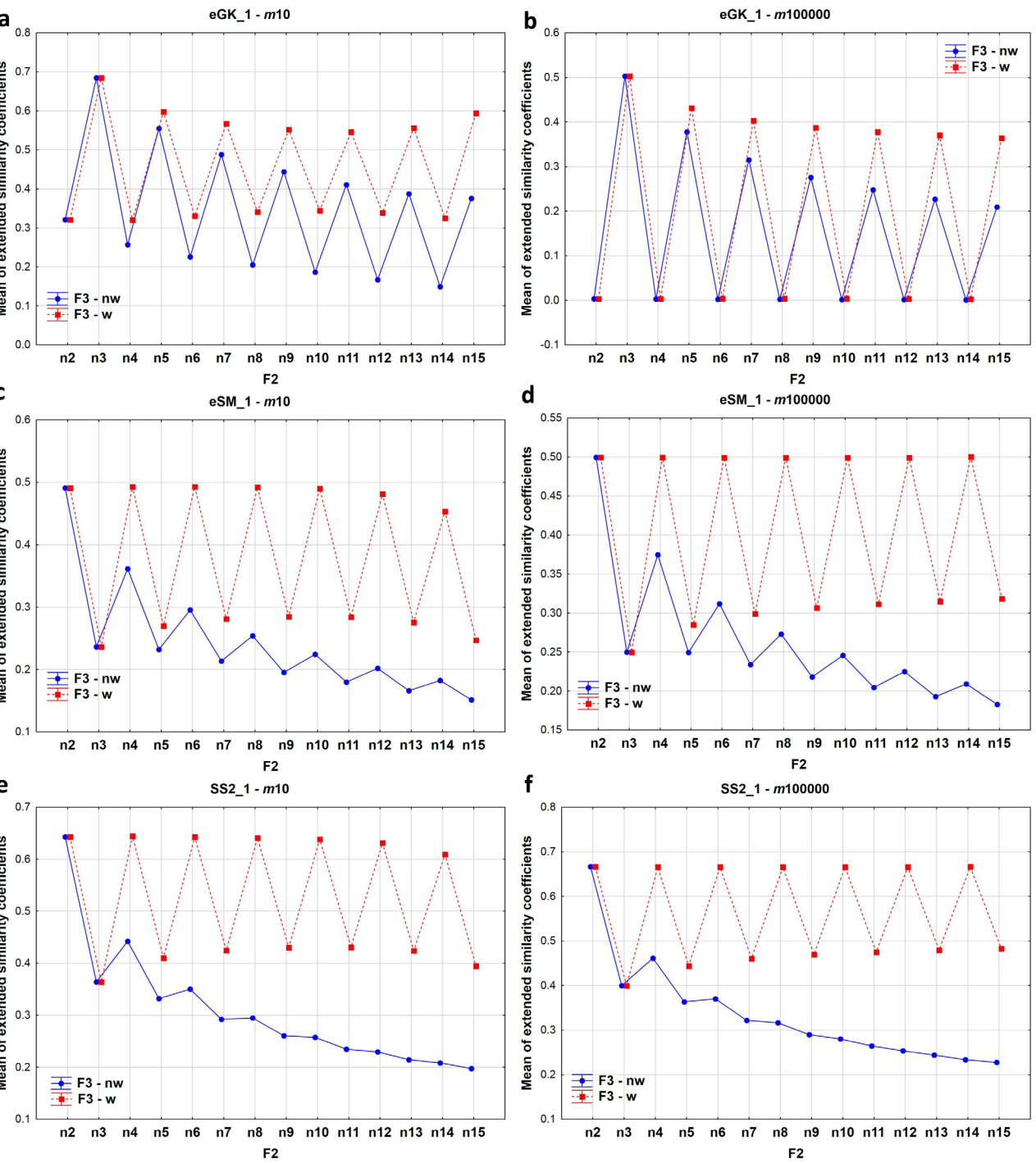
m = 10, 100, 1000, 100 000 (fingerprints are random dichotomous vectors of length *m*);


Median
25%-75%
Min-Max
Mean of extended similarity coefficients
1.0
0.8
0.6
0.4
0.2
0.0
eAC_1
eBUB_1
eCT1_1
eCT2_1
eCT3_1
eCT4_1
eFai_1
eGK_1
eGle_1
eHD_1
eJT_1
eJa_1
eJa_0
eRG_1
eRR_1
eRT_1
eSM_1
eSS1_1
eSS2_1


Mean of extended similarity coefficients
0.42
0.40
0.38
0.36
0.34
0.32
0.30
0.28
0.26
m10
m100
m1000
m100000
F5


Mean of extended similarity coefficients
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
n2
n3
n4
n5
n6
n7
n8
n9
n10
n11
n12
n13
n14
n15
F2


F3 - nw
F3 - w
Mean of extended similarity coefficients
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
n2
n3
n4
n5
n6
n7
n8
n9
n10
n11
n12
n13
n14
n15
F2



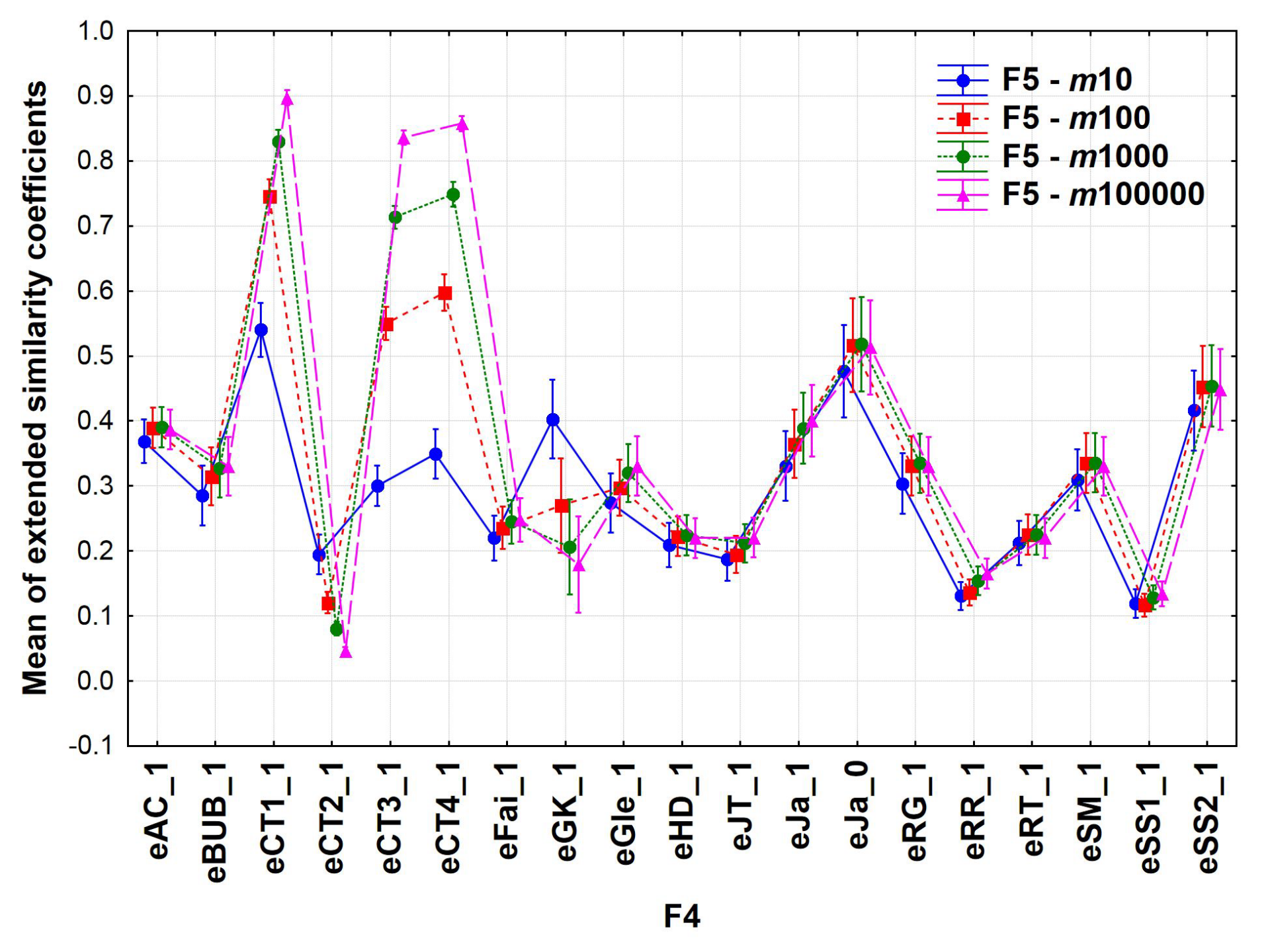
Lefthand side *m*=10

Righthand side *m*=100 000

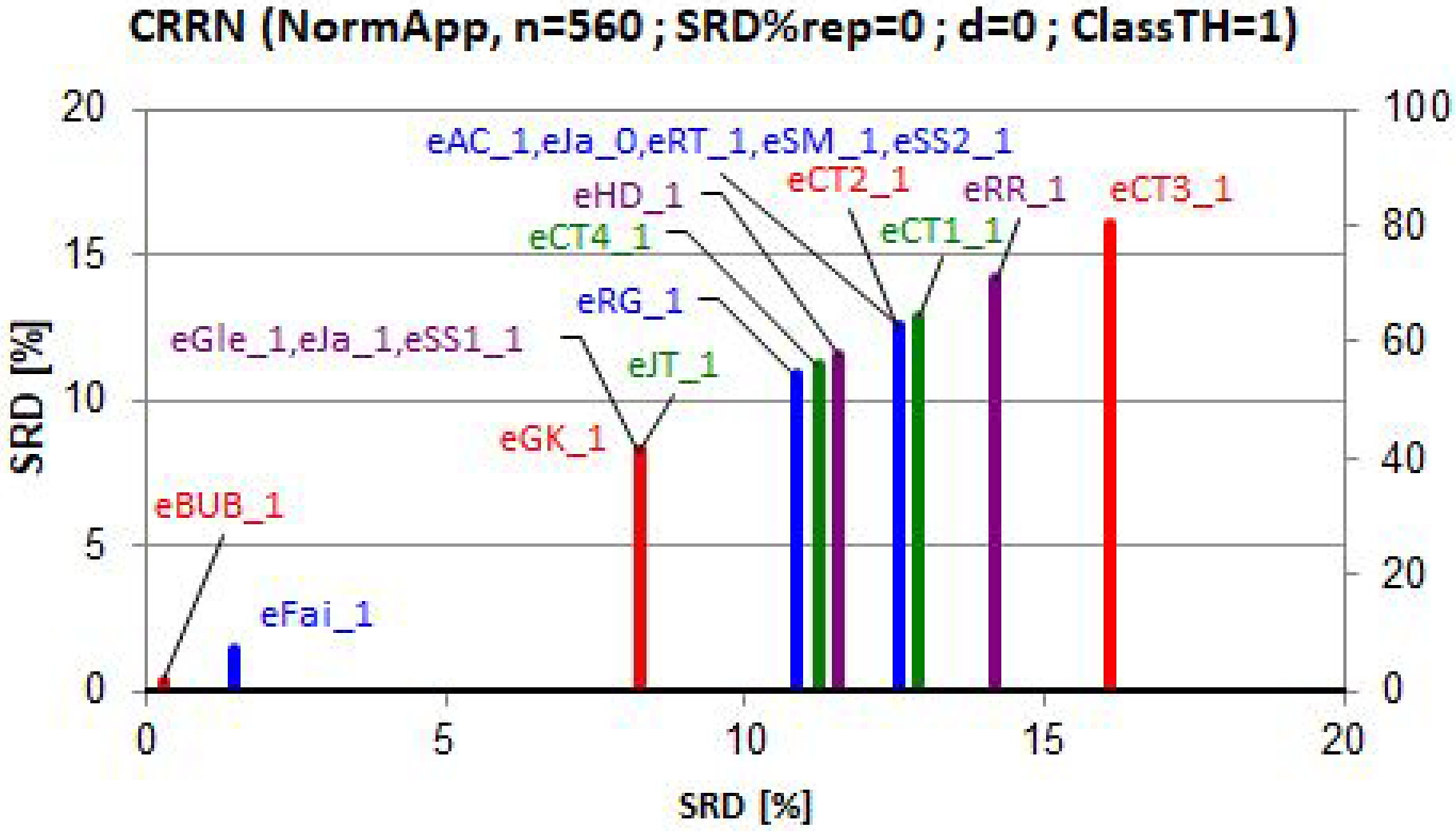
eGK_1:
extended Goodman-Kruskal,

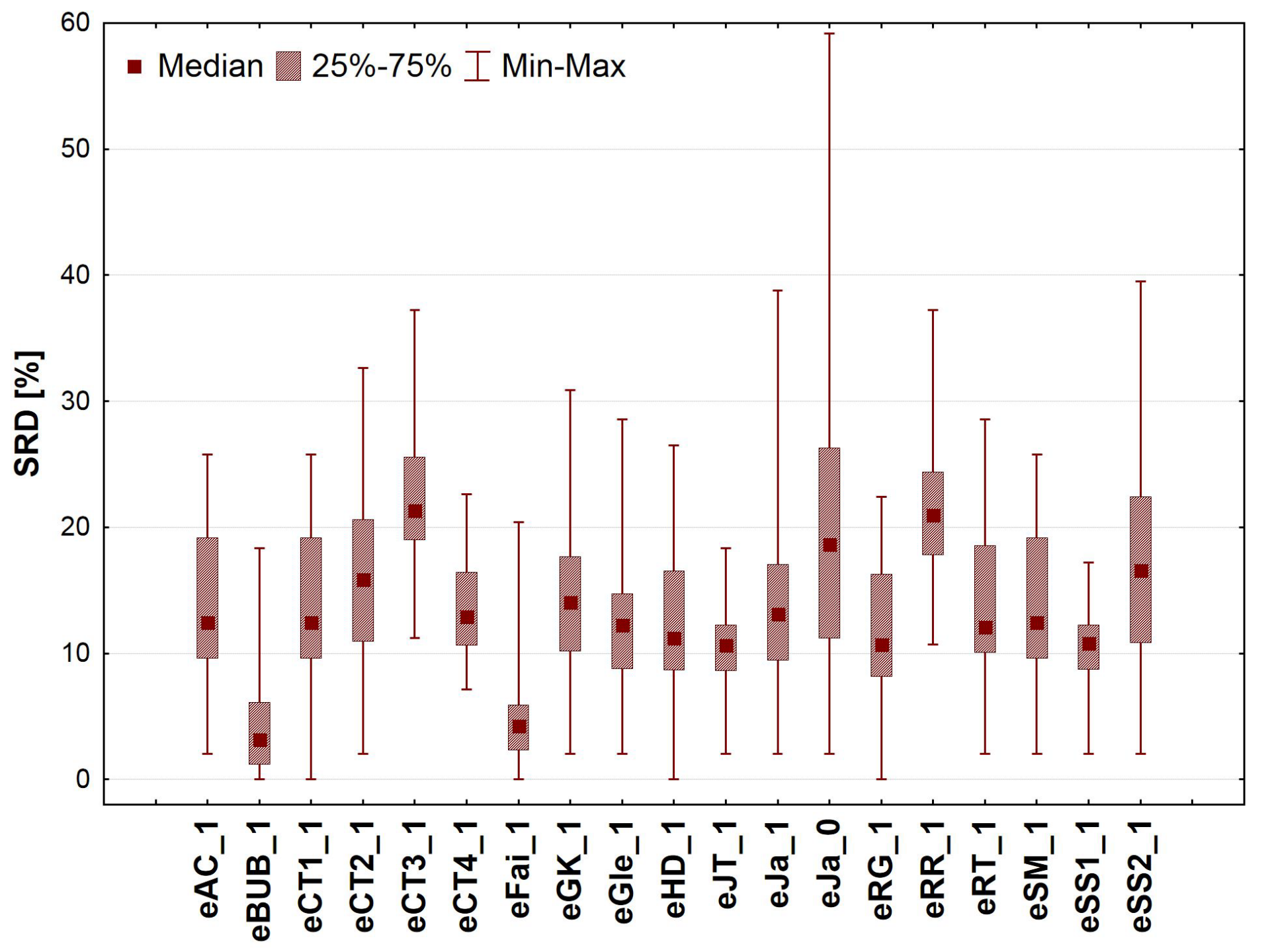
eSM_1:
extended Sokal-Michener

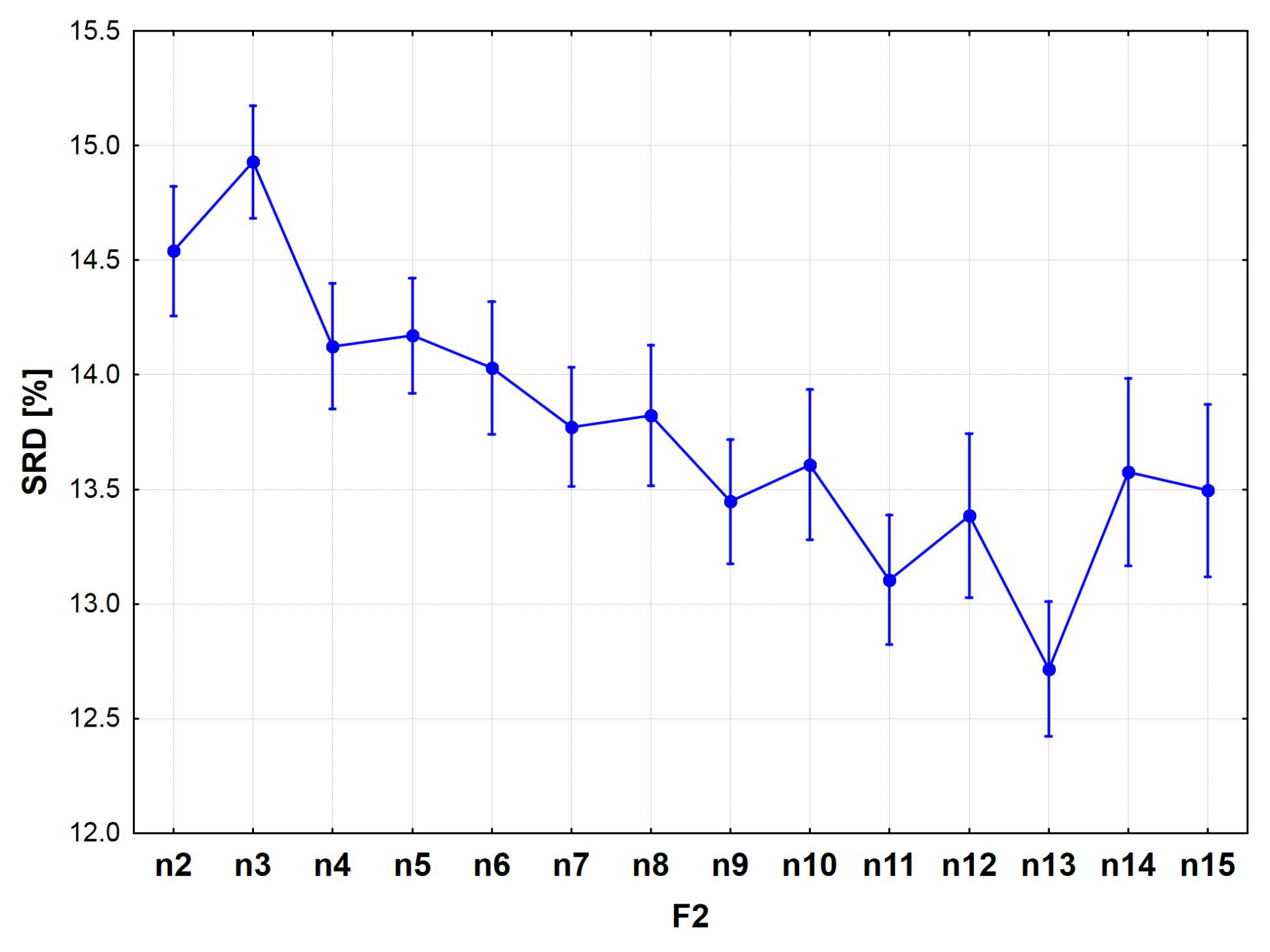
eSS_2:
extended Sokal-Sneath 2

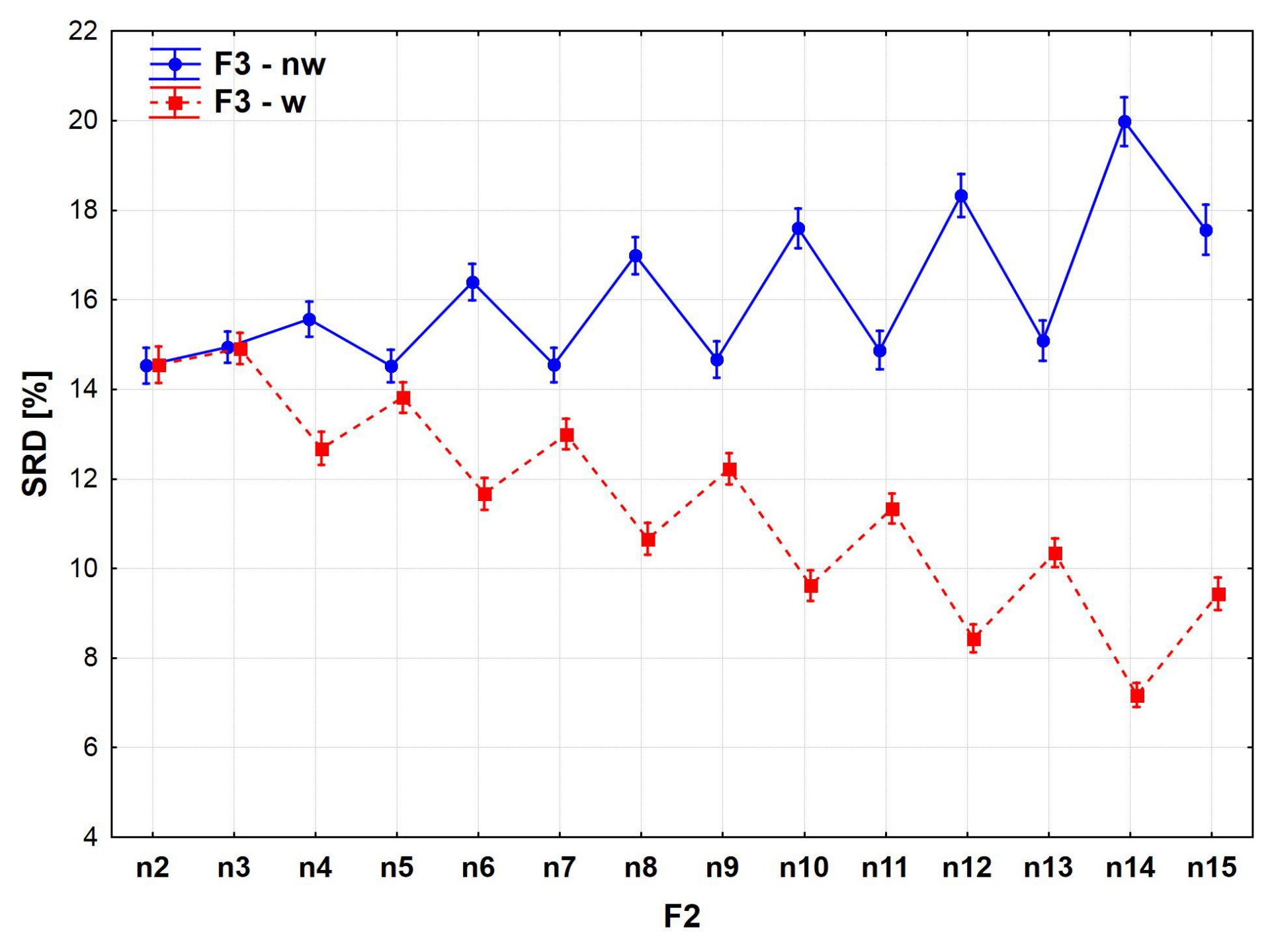

Mean of extended similarity coefficients
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
F5 - m10
F5 - m100
F5 - m1000
F5 - m100000
eAC_1
eBUB_1
eCT1_1
eCT2_1
eCT3_1
eCT4_1
eFai_1
eGK_1
eGle_1
eHD_1
eJT_1
eJa_1
eJa_0
eRG_1
eRR_1
eRT_1
eSM_1
eSS1_1
eSS2_1
F4

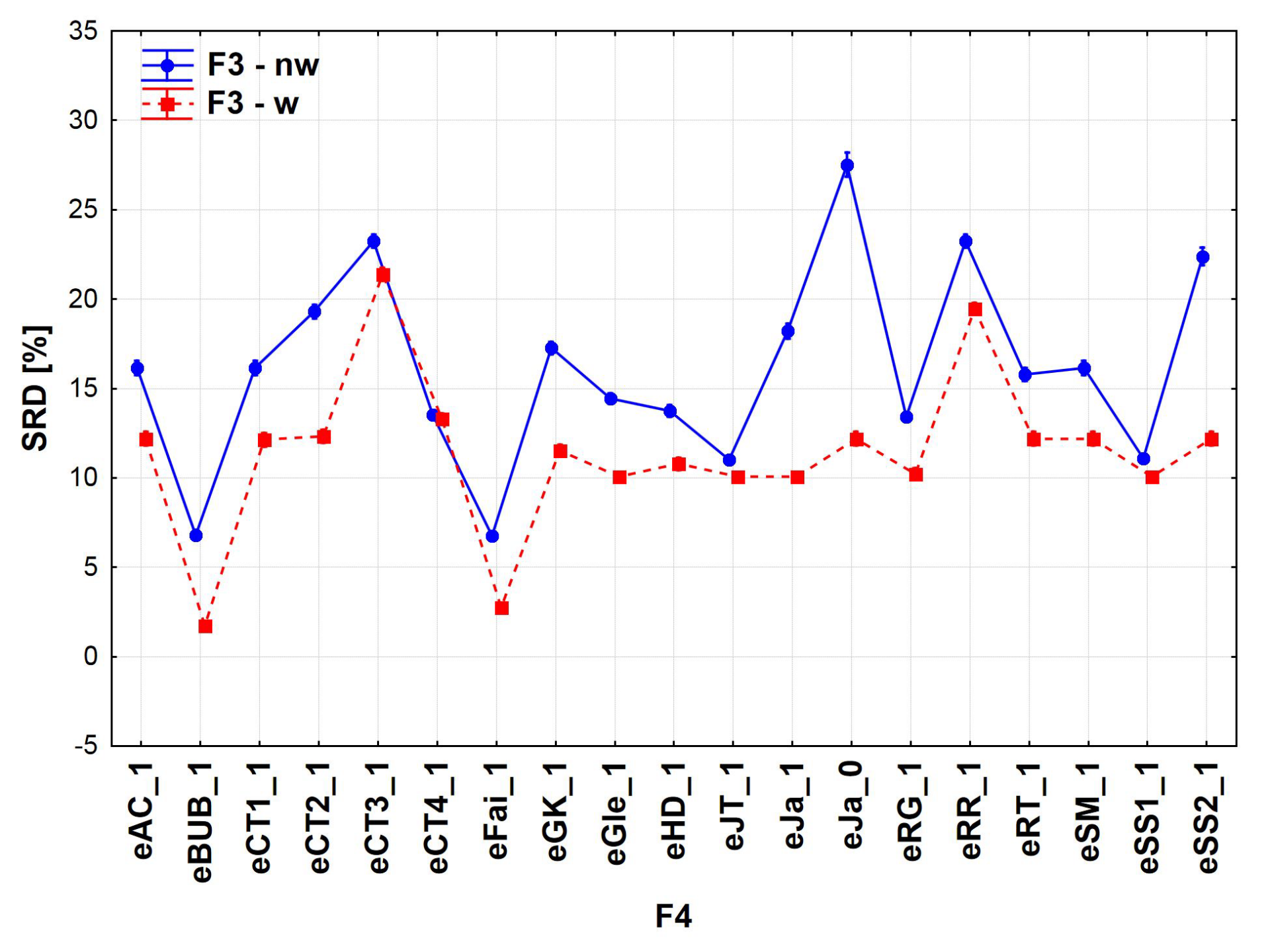
Ranking of extended indices

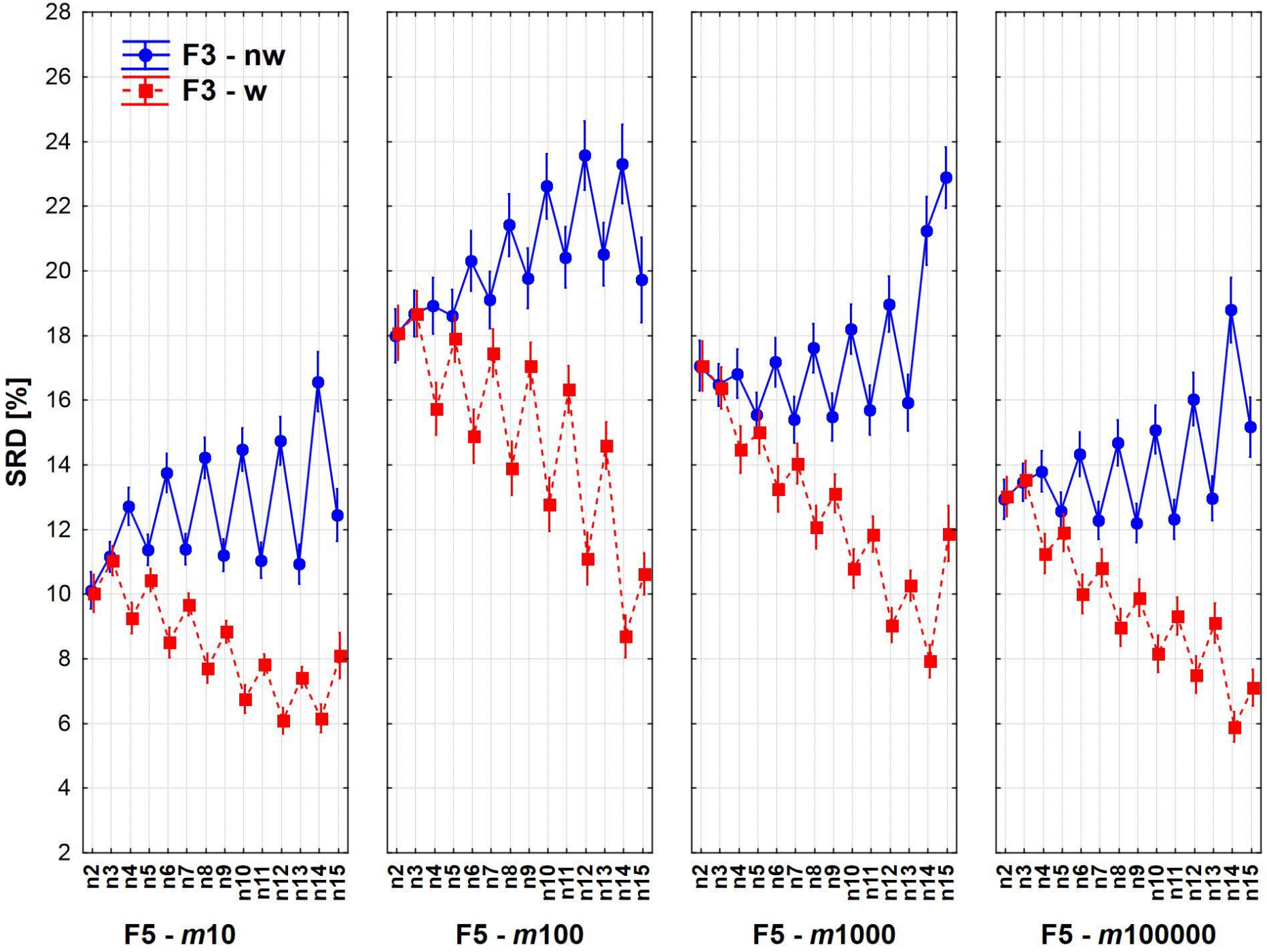



Median
25%-75%
Min-Max
SRD [%]
60
50
40
30
20
10
0
eAC_1
eBUB_1
eCT1_1
eCT2_1
eCT3_1
eCT4_1
eFai_1
eGK_1
eGle_1
eHD_1
eJT_1
eJa_1
eJa_0
eRG_1
eRR_1
eRT_1
eSM_1
eSS1_1
eSS2_1


15.5
15.0
14.5
14.0
13.5
13.0
12.5
12.0
SRD [%]
n2
n3
n4
n5
n6
n7
n8
n9
n10
n11
n12
n13
n14
n15
F2


F3 - nw
F3 - w
SRD [%]
22
20
18
16
14
12
10
8
6
4
n2
n3
n4
n5
n6
n7
n8
n9
n10
n11
n12
n13
n14
n15
F2


F3 - nw
F3 - w
SRD [%]
35
30
25
20
15
10
5
0
-5
eAC_1
eBUB_1
eCT1_1
eCT2_1
eCT3_1
eCT4_1
eFai_1
eGK_1
eGle_1
eHD_1
eJT_1
eJa_1
eJa_0
eRG_1
eRR_1
eRT_1
eSM_1
eSS1_1
eSS2_1
F4


F3 - nw
F3 - w
SRD [%]
28
26
24
22
20
18
16
14
12
10
8
6
4
2
n2 n3 n4 n5 n6 n7 n8 n9 n10 n11 n12 n13 n14 n15
F5 - m10
F5 - m100
F5 - m1000
F5 - m100000

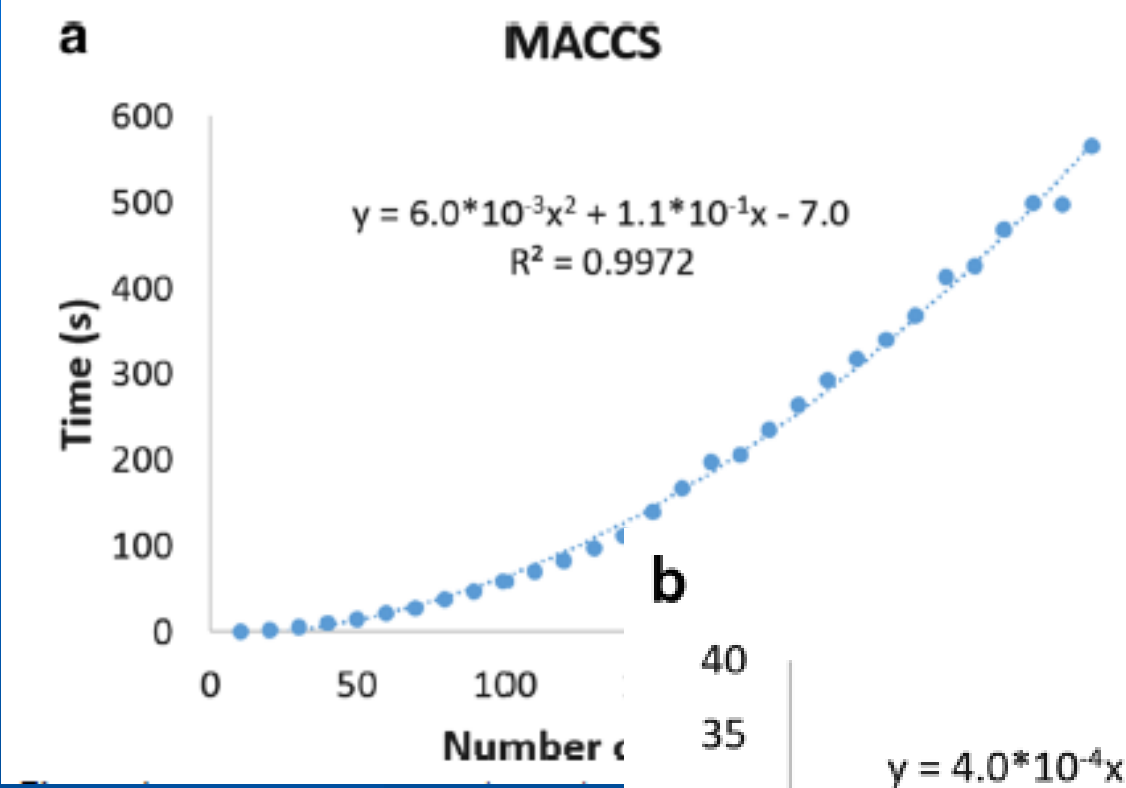
Summary

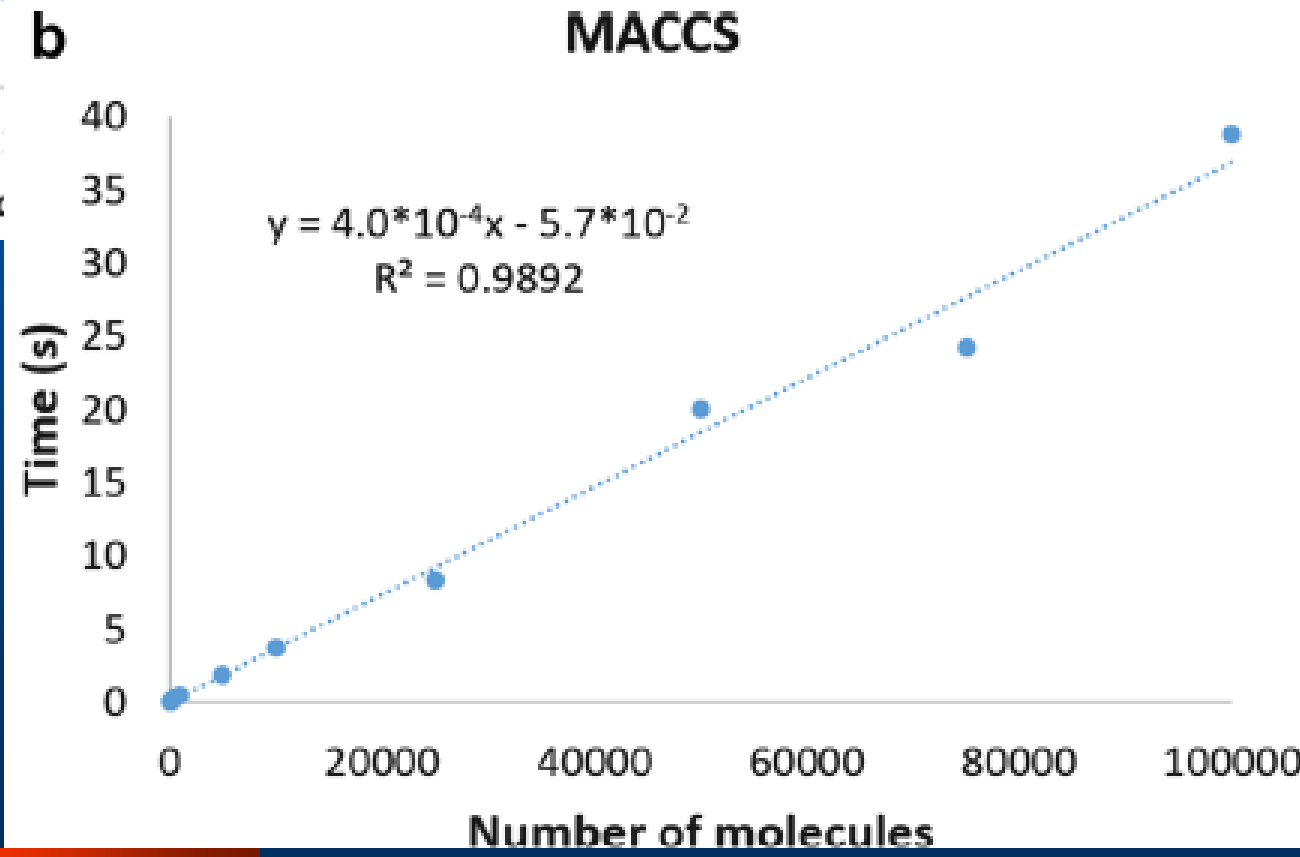
- By analogy (?) with the previous similarity indices, we have created 19 new, symmetric indices allowing multiple comparisons.
- Their properties were extracted:
- Using ANOVA, we decomposed the effect of the factors: *n*, *m*, *w*, *S*, *n***w*, *w***S*, ...
- Using SRD, we found optimal factors, consistent solutions.

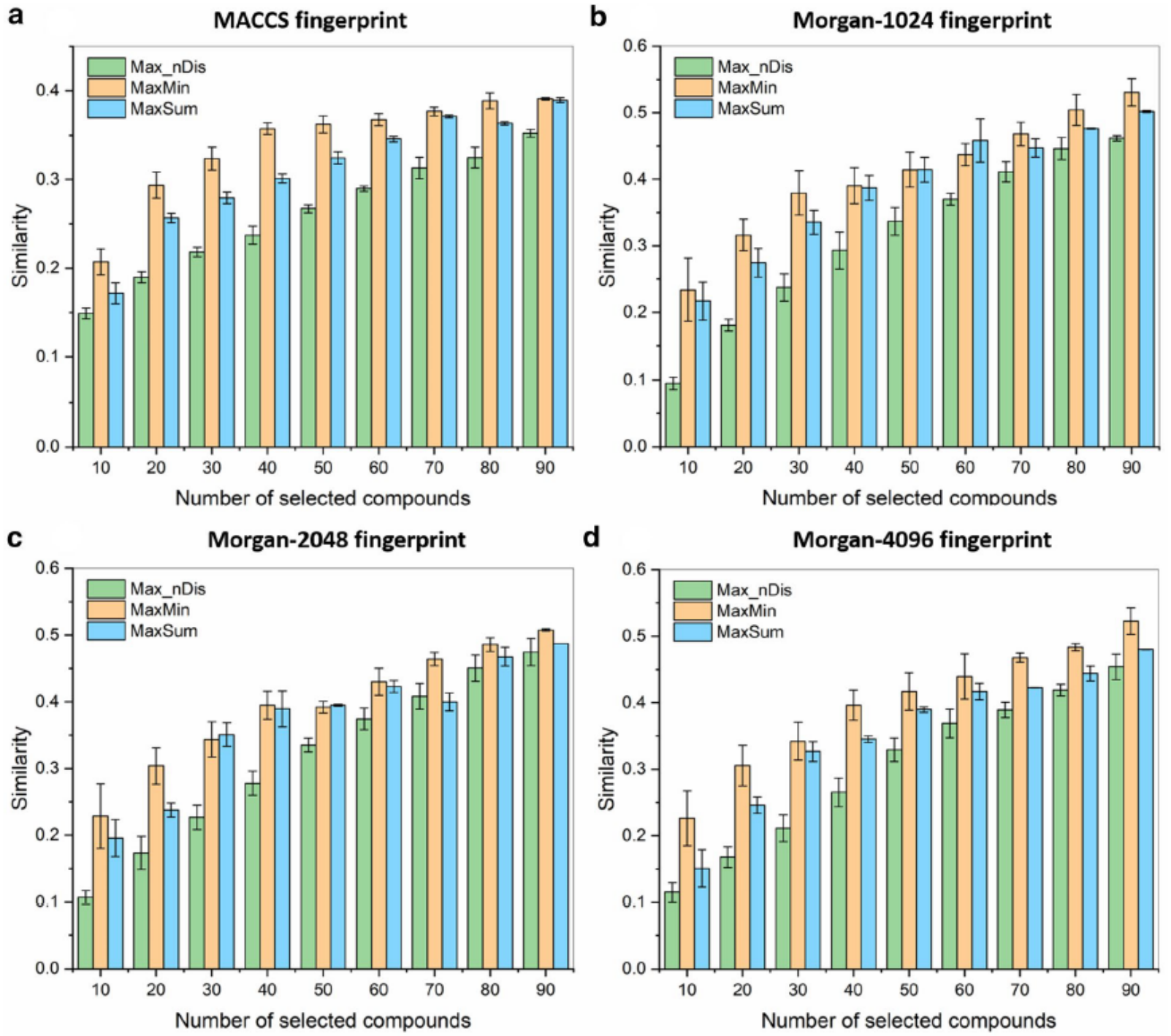
Conclusions

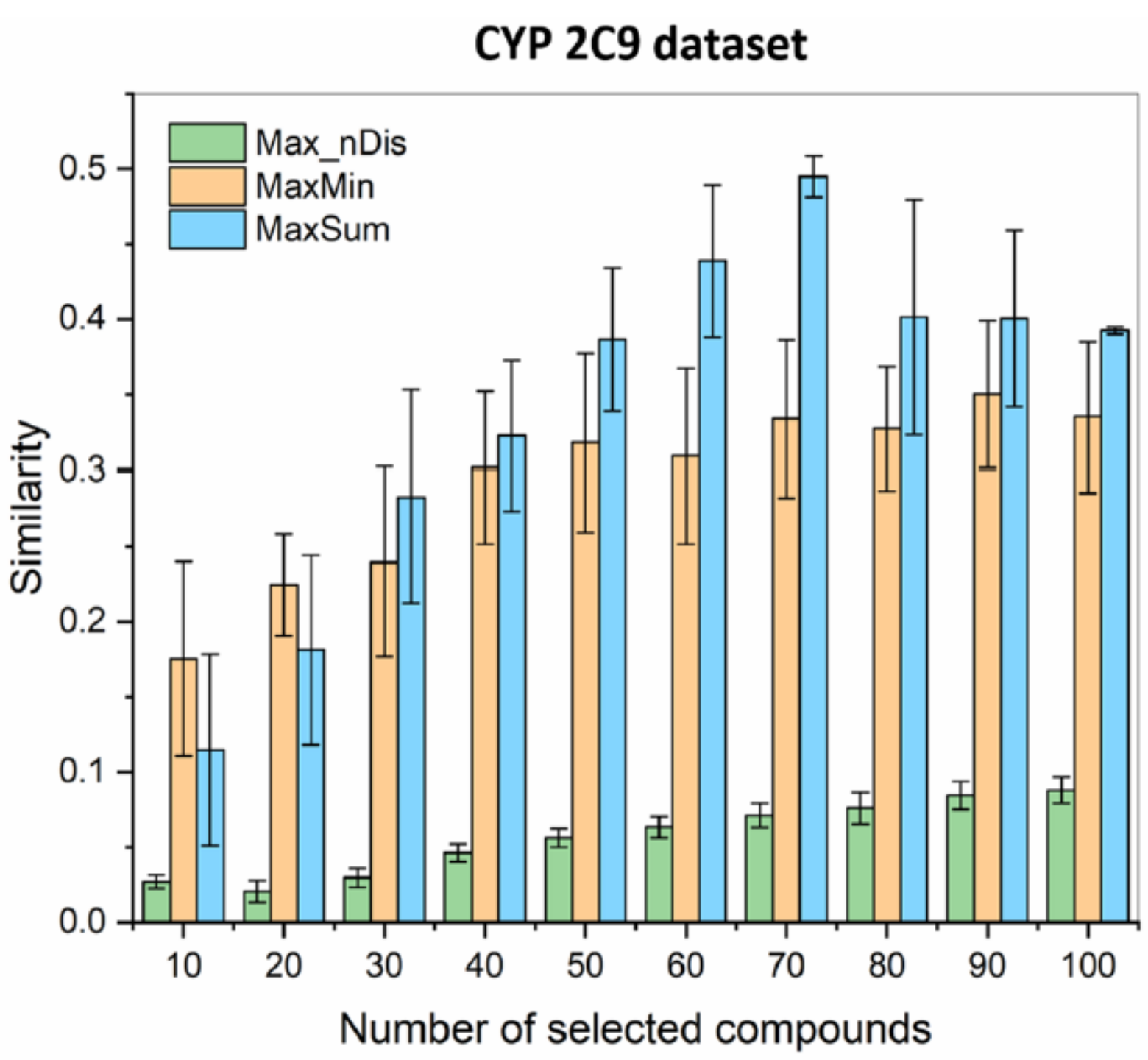
- The optimal comparison number? (*n*=13)
- Do you need weighting? Yes (optimum: *n*=14)
- Extensions of the Baroni-Urbani-Buser (eBUB) and Faith (eFai) coefficients (counting similarities of 1) are most recommended.
- The fingerprint length dependence of some indices (eCTi and eGK) is significant.
- Optimal fingerprint length? (10 or 100000, but an optimal index can be searched for a realistic 1000)

ion time


a
MACCS
y = 6.0*10-3x2 + 1.1*10-1x - 7.0
R² = 0.9972
Time (s)
600
500
400
300
200
100
0
0
50
100
Number


b
MACCS
y = 4.0*10-4x - 5.7*10-2
R² = 0.9892
Time (s)
40
35
30
25
20
15
10
5
0
0
20000
40000
60000
80000
100000
Number of molecules


a
MACCS fingerprint
b
Morgan-1024 fingerprint
c
Morgan-2048 fingerprint
d
Morgan-4096 fingerprint
Max_nDis
MaxMin
MaxSum
Similarity
Number of selected compounds
0.0
0.1
0.2
0.3
0.4
0.5
0.6
10
20
30
40
50
60
70
80
90


CYP 2C9 dataset
Max_nDis
MaxMin
MaxSum
Similarity
0.0
0.1
0.2
0.3
0.4
0.5
10
20
30
40
50
60
70
80
90
100
Number of selected compounds

Advantages of *n*-ary

- **Less computing power**, faster algorithm, larger datasets,
- **Diversity detection** significantly better,
- Possibility of **clustering** (HCA),
- Definition of **internal and external consistency indices** & their optimization is possible.

Acknowledgement

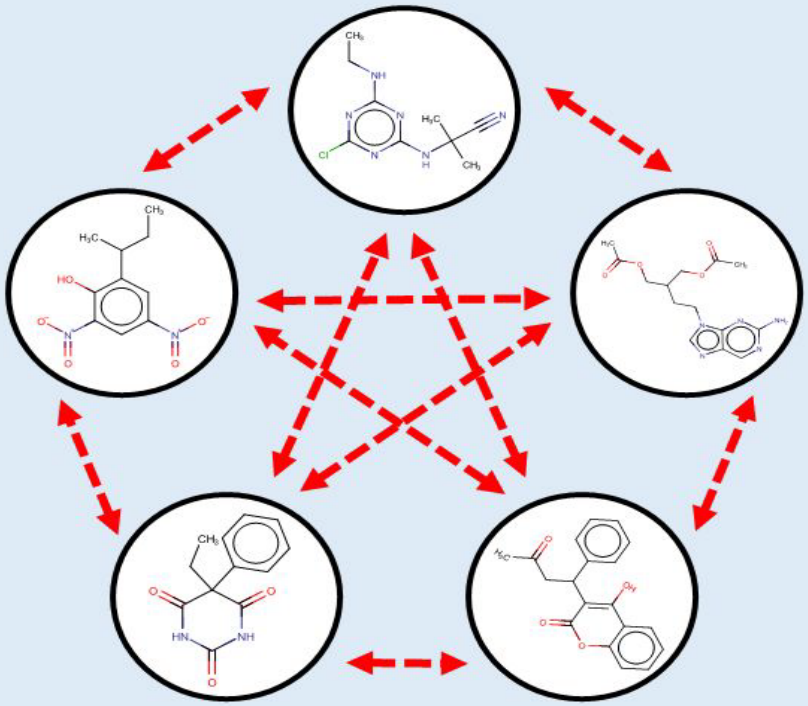
- National Research Development and Innovation Fund (Hungary): OTKA K134260
- University of Florida startup grant (RAMQ)
- Covid pandemic

Releted papers

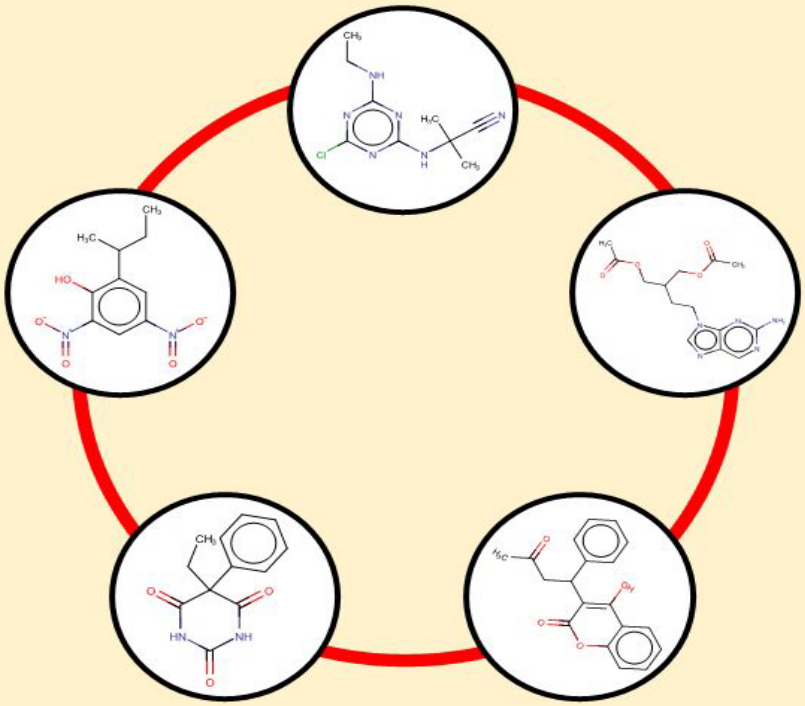
- Ramón Alain Miranda-Quintana*, Dávid Bajusz, Anita Rácz, Károly Héberger*, Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: Theory and characteristics *Journal of Cheminformatics,* 13, Article No: 32 (2021) if(2020)=5.514 (D1)
- Ramón Alain Miranda-Quintana*, Dávid Bajusz, Anita Rácz, Károly Héberger*, Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 2: speed, consistency, diversity selection *Journal of Cheminformatics* 13, Article No: 33 (2021) if(2020)=5.514 (D1)
- Ramon Miranda-Quintana*, Dávid Bajusz, Anita Rácz, Károly Héberger*, Differential Consistency Analysis: Which similarity measures can be applied in drug discovery? *Molecular Informatics,* 40, Article No.: 2060017 (2021) if(2020)=3.353 (Q2)
- Dávid Bajusz, Ramón Alain Miranda-Quintana*, Anita Rácz, Károly Héberger*, Extended many-item similarity indices for sets of nucleotide and protein sequences *Computational and Structural Biotechnology Journal* 19, 3628-3639 (2021) if(2020)=7.271 (Q1)

COMPUTE TIME

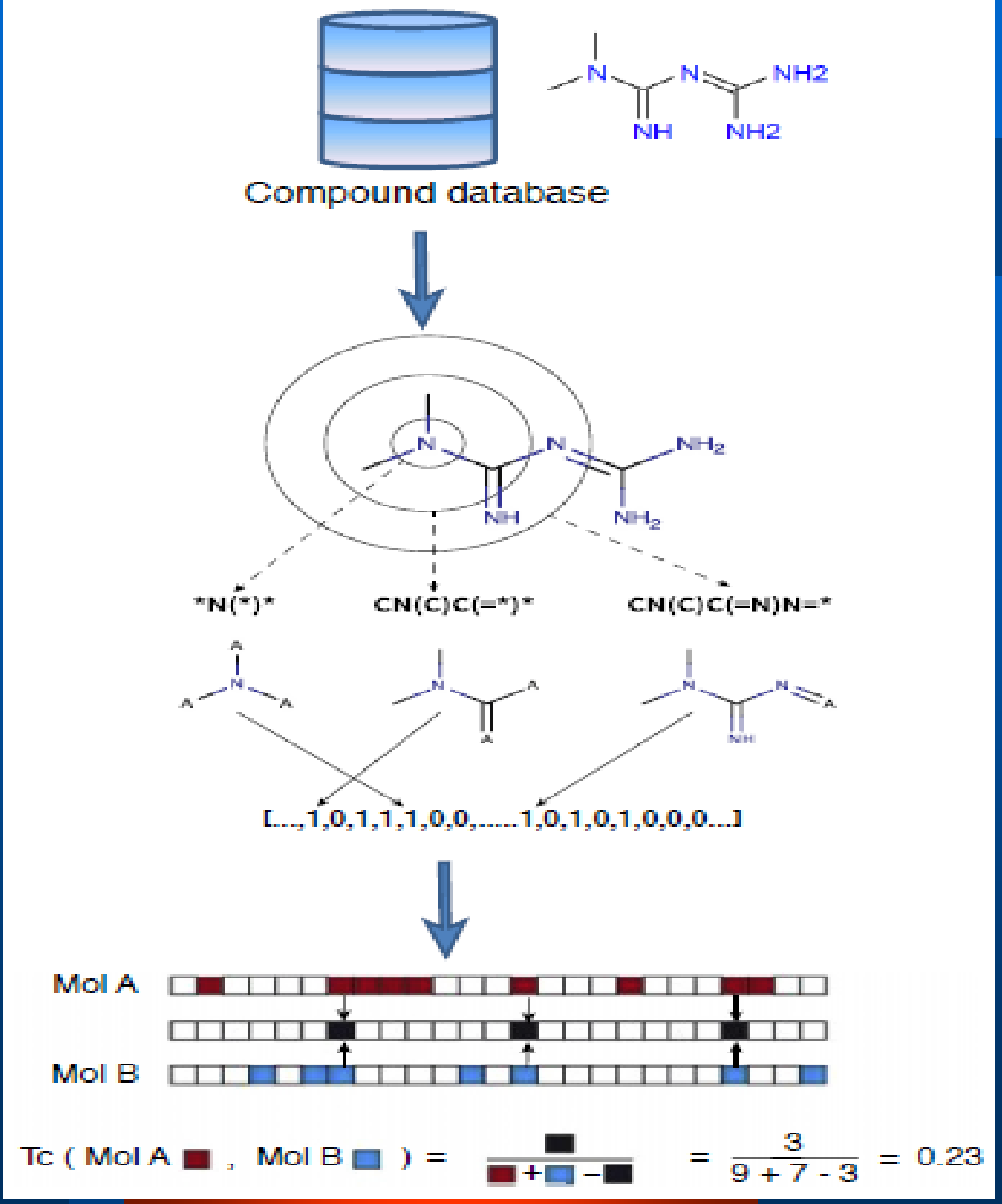
Binary similarity



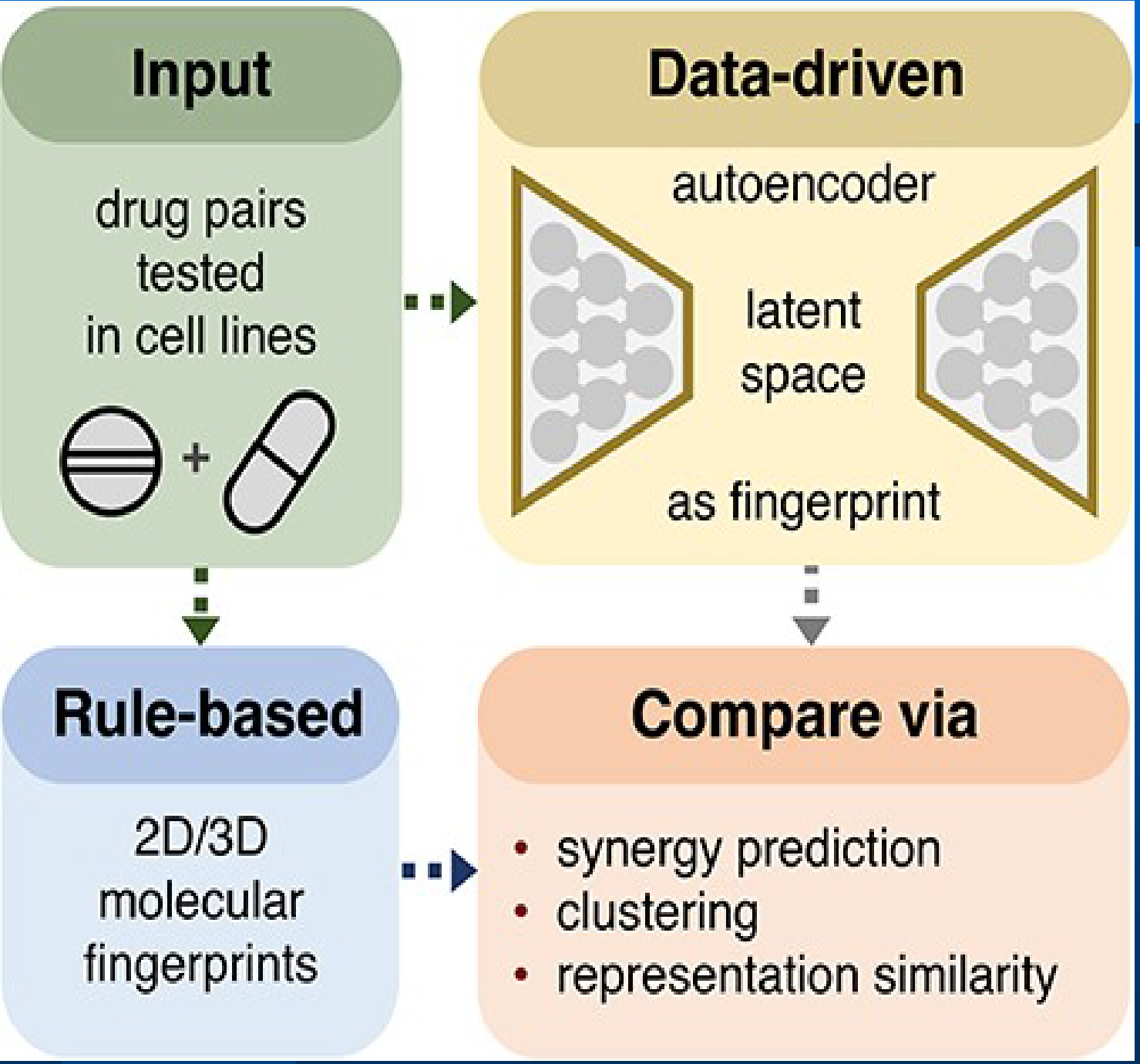
Extended (*n*-ary) similarity



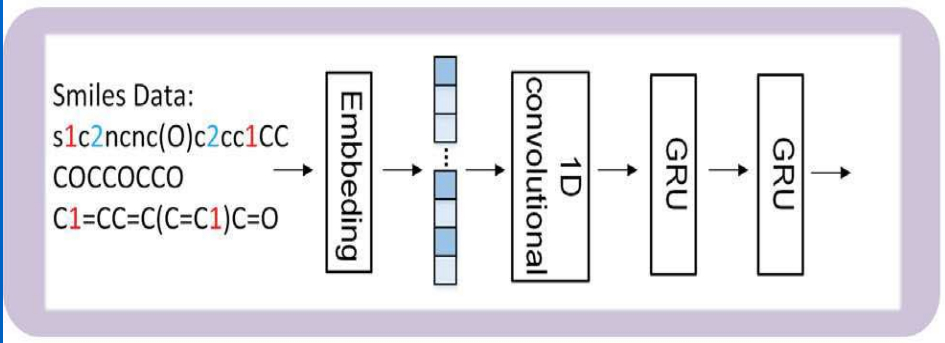
DIVERSITY

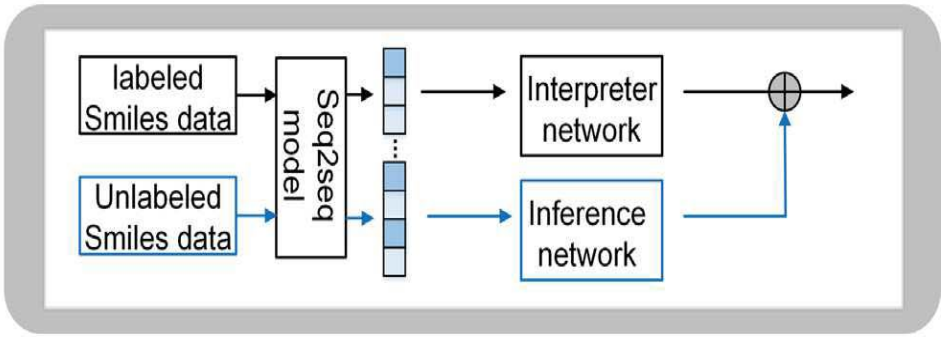


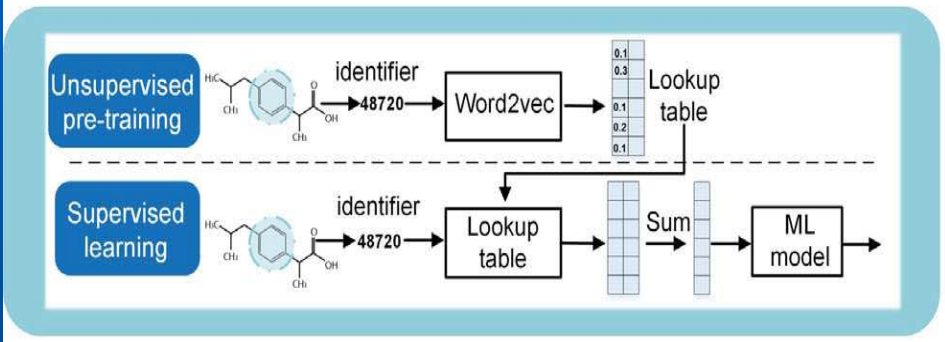
Yu-Chen Lo, Stefano E. Rensi, Wen Torng and Russ B. Altman,
Machine learning in chemoinformatics and drug discovery
Drug Discovery Today **23**, 1538-1546 (2018)

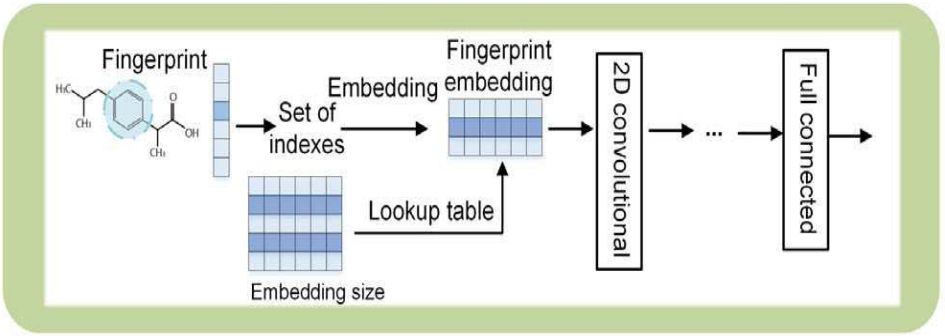


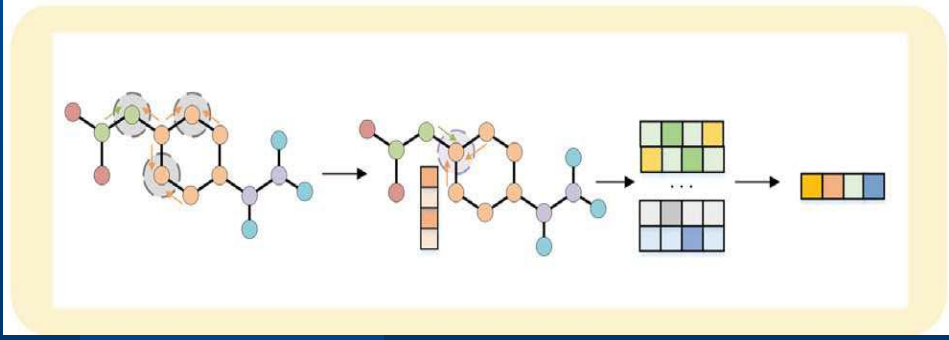
B. Zagidullin, Z. Wang, Y. Guan, E. Pitkänen and J. Tang: Comparative analysis of molecular fingerprints in prediction of drug combination effects *Briefings in Bioinformatics, 00(00)*, 2021, 1–15 https://doi.org/10.1093/bib/bbab291

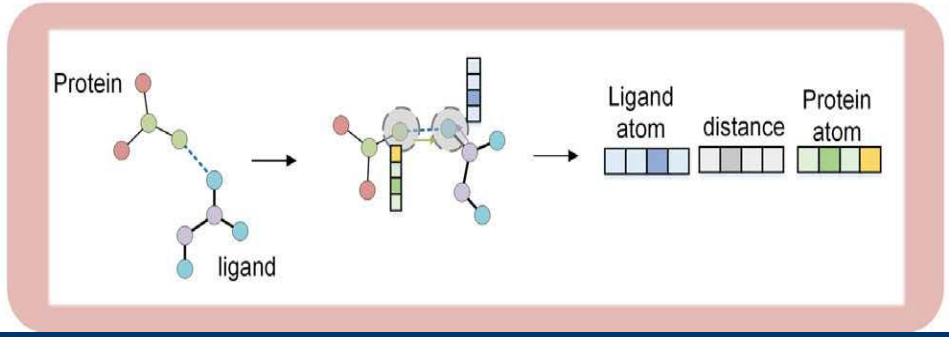






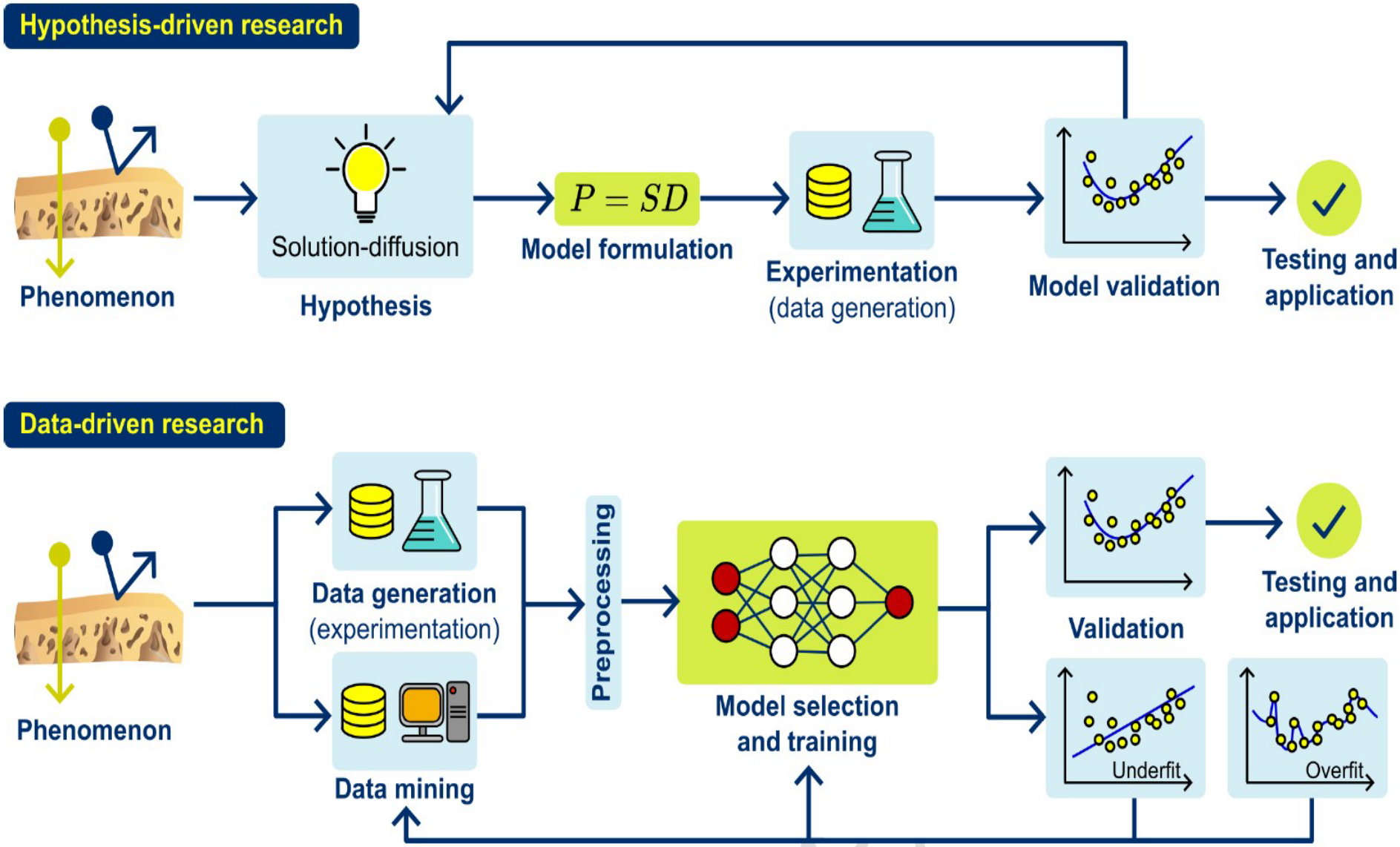






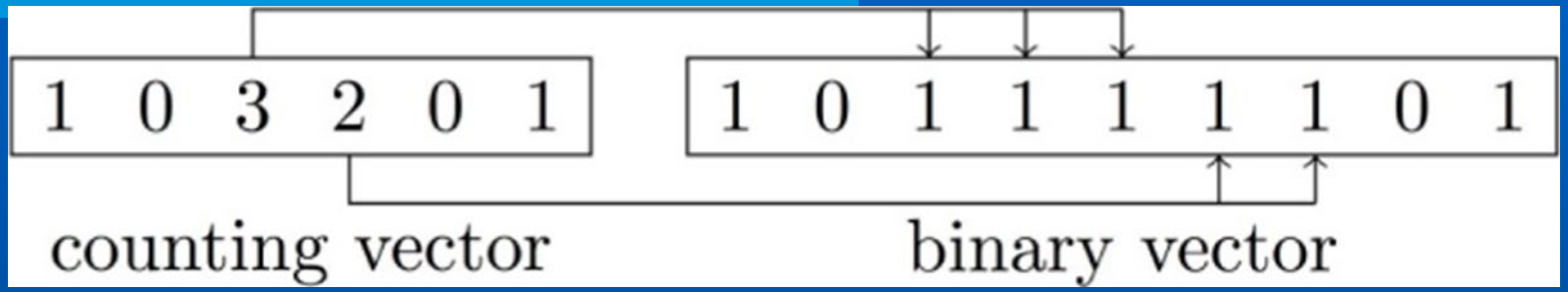
Youjun Xu, Chenjing Cai, Shiwei Wang, Luhua Lai, Jianfeng Pei: Efficient molecular encoders for virtual screening. *Drug Discov Today Technol*. 2019 Dec; 32-33:19-27. doi: 10.1016/j.ddtec.2020.08.004 Epub 2020 Oct 4.

Machine learning *vs.* classic

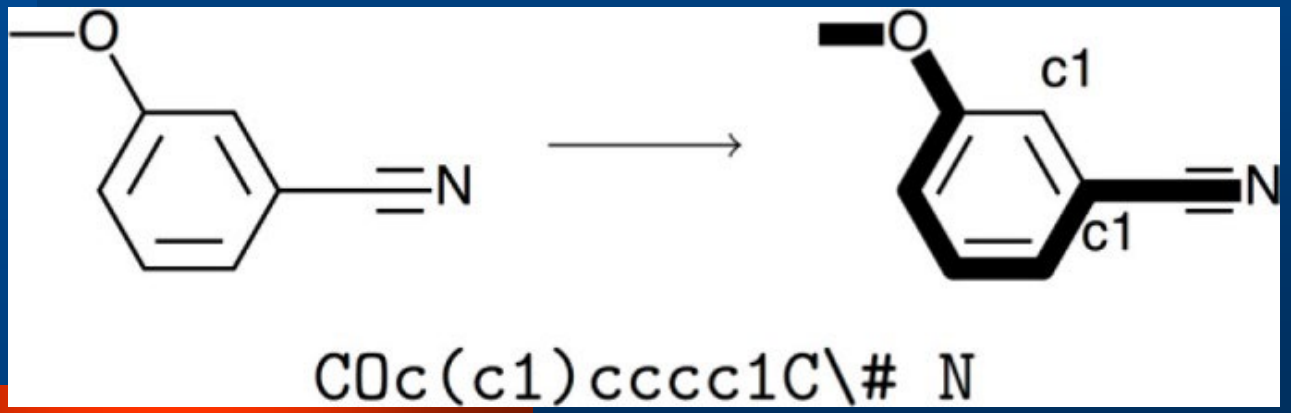


Molecular representations

- A ***binary fingerprint*** describes the molecule as a set of features:



- A ***counting vector*** allows for a more detailed description of the molecule as a multi-set of features
- Linear representation: SMILES:



Main steps of drug design