Re-evaluation of fingerprints (drug design, chromatographic, foodomics, forensic): model building, method development and validation of predictive models

Final report for 2016-09-01 - 2021-08-31 time period

Project ID: K 119269

The essence of the project was the evaluation and comparison of the fingerprints and the different similarity metrics in various scientific fields. For this purpose, several new metrics, algorithms, and modeling workflows have developed. The multi criteria decision making and comparative algorithms have also developed/upgraded for the special tasks. As fingerprints can be applied in several research area, the project was entirely interdisciplinary, which helped us to provide a greater impact for the project. The typical application areas of the fingerprints in the research were the drug design, food chemistry (sensory attributes *etc*.), and analytical chemistry (spectroscopy, chromatography). We may decompose our work from the application point of view to the following cornerstones: i) similarity profiling (chemical-, interaction- and analytical fingerprints), ii) model comparisons and iii) multivariate studies of practical relevance. The below mentioned subtopics are in thematic-chronological order.

1. SIMILARITY PROFILING

a. Encoding fingerprints, molecular descriptions, and similarity of molecules

The Elsevier Publishing House invited us to write a book chapter, where we summarized chemical data formats, the various coding possibilities for fingerprints, molecular descriptions encoding the chemical structure of molecules and similarity matrices as well as their data fusion possibilities in detail.

1. Comprehensive Medicinal Chemistry III - Volume 3 In silico methods Eds: A. Davies, C. Edge, (2017) 329–378.

(In such handbooks there is no room for acknowledgement, no acknowledgement was given in other book chapters either.)

b. Metabolomics profiling

Recent increase in the use of qualitative metabolomic data, described by the presence (1) or absence (0) of particular metabolites, demonstrates great potential in the field of metabolomic profiling and fingerprint analysis. Nine qualitative metabolomic data sets covering a wide range of natural products and metabolomic profiles were applied to assess binary similarity measures for the fingerprinting of plant extracts and natural products. The measures were analyzed by the novel sum of ranking differences method (SRD), searching for the most promising candidates. Analysis of variance (ANOVA) revealed that concordantly and intermediately symmetric similarity coefficients are better candidates for metabolomic fingerprinting than the asymmetric and correlation based ones.

2. Metabolomics 14 (2018) 29 1-9 Q1 if(2018)=3.167

c. Similarity measures for interaction fingerprints

Similarity metrics that are viable alternatives to the commonly used Tanimoto coefficient were identified based on a comparison with an ideal reference metric (consensus). In a large-scale comparison, we have assessed the effect of similarity metrics and interaction fingerprint (IFP) configurations to several virtual screening scenarios with ten different protein targets and thousands of molecules. Particularly, the effect of considering general interaction definitions (such as Any Contact, Backbone Interaction and Sidechain Interaction), the effect of filtering methods and the different groups of similarity metrics were determined.

3. Journal of Cheminformatics 10 (2018) Article No: 48 D1 if(2018)=4.154

d. Fingerprint analysis of human eye-movement data

In Hungary, our research team is the first, which evaluates human eye-tracking data of food-oriented stimuli. As a first step of the project, a mini-review about the applications of eye-tracking methodology in food sciences was published in a bilingual scientific journal, Journal of Food Investigations (Élelmiszervizsgálati Közlemények). Since eye-movement patterns of participants are unique and can be easily influenced: the disturbing factors of eye-movements have been studied in detail. A publication about the influencing effect of mood was published in Food Quality and Preference. Fingerprint analysis of eye-movement data was presented in Conferentia Chemometrica 2017.

Journal of Food Investigations, LXIII. (2017) 1565-1576. [in Hungarian] and

4. Food Quality and Preference 61 (2017) 1-5. Q1 if(2017)=3.932

e. Multi perspective evaluation of phytonutrients

In this work, phytonutrient fingerprints of tomato landraces were evaluated, and comparison of the varieties was made based on seven phytonutrient values. Sum of ranking differences method is able to define the best tomato landrace based on their phytonutrient fingerprints.

5. Journal of Functional Foods, 33 (2017) 211-216. Q1 if(2017)=3.470

f. Similarity profiling of sensory data

In food sensory analysis, check-all-that-apply (CATA) questionnaires are widely used both in trained and consumer sensors tests. Despite its popularity, there has not been introduced any method to evaluate the discrimination ability of assessors completing such tests. Discrimination ability helps the researchers to define whether the assessors were able to perceive any differences between products. The presented methodology uses binary similarity measures to assess the discrimination ability of consumers, which enables the researchers to cluster the consumers based on their discrimination abilities. The introduced methodology fits into the international trends, as panel agreement in CATA tests has been introduced recently. Using both methods (agreement and discrimination ability), detailed information can be obtained regarding the performance of the participants.

6. Foods, 10(5) (2021) Q1 if(2021)=4.350

g. Profiling of chromatographic columns

Two similar RP-HPLC columns are not equally suitable for the requisite separation, and there is no universal RP-HPLC column covering a variety of analytes. Therefore, column selection is a crucial segment of RP-HPLC method development, individual columns can also exhibit a unique character owing to specific polar, hydrogen bond, and electron pair donor−acceptor interactions. Our review is a comprehensive, authoritative, critical, and easily readable monograph of the most relevant publications regarding column selection and characterization in RP-HPLC covering the past four decades. The review not only enumerates the chromatographic column selection systems, and their historical evolution, but characterizes the columns from a point of view for multivariate data analysis.

7. Chemical Reviews, 119 (2019) 3674−3729 and 4818 D1 if(2019)= 52.758

h. Comparison of machine learning algorithms and variants of consensus modeling

Cytochrome P450 (CYP) enzymes play an important role in the metabolism of xenobiotics. Since they are connected to drug interactions, screening for potential inhibitors is of utmost importance in drug discovery settings. An extensive classification model for one of the major isoenzymes, namely the CYP P450 2C9 isoenzyme have carried out with the use of more than 45 000 different molecules and one, two- and three-dimensional molecular descriptors. Chemical, interaction, and pharmacophore fingerprints were also applied and compared in the modeling phase. A consensus modeling based on the applied two machine learning algorithms (boosted tree and multilayer feedforward backpropagation network) was successfully used to predict the inhibitory activity of the molecules. Chemical space was the largest available and reached its present limit for the 2C9 isoenzyme.

8. Journal of Computer-Aided Molecular Design, 34 (2020) 831-839 Q2 if(2020)= 3.686

i. Comparison of similarity measures used in drug design

In this study we carried out an extensive analytical study of the conditions leading to two comparative measures providing equivalent results over a given set of molecules. The differential consistency analysis (DCA) has been presented as a novel way to study the consistency between comparative similarity measures. As an example, DCA provided a clear and concise proof of the consistency between the Tanimoto and the Dice (Hodgkin-Richards) coefficients. With this tool we can reveal how the consistency can be established in an analytical way with minimal (or no) assumptions. DCA can be applied to each similarity coefficient.

9. Molecular Informatics, Vol. 40. 2060017 (2021) Q2 if(2020)= 3.353

j. Evaluation of spectral fingerprints

- Extracellular vesicles (EVs) are lipid bilayer–bounded particles that are actively synthesized and released by cells. The major components of EVs are lipids, proteins, and nucleic acids. Protein concentration in EVs was determined based on infrared spectroscopy. The spectral fingerprints of bovine serum albumin samples were used for multivariate calibration with PLS regression. The appropriate model was tested with red blood cell derived EVs (REVs). The predicted protein concentrations were similar to that obtained using the Bradford assay.

10. Analytical and Bioanalytical Chemistry 12 (2020) 4619-4628 Q1 if(2020)= 4.142

- The authentication and quality assurance of snack products have become important since these convenience foods are popular in the modern lifestyle. Here we have used non-destructive and environmental-friendly near infrared spectroscopy to classify the snacks based on their i) frying oil, ii) raw material, iii) place of origin and production technology. The spectral fingerprints were applied for the analysis with three well-known classification methods, such as PLS-DA, random forest, and multilayer feed-forward backpropagation network. The four models provide a solid base of NIR spectroscopy coupled with chemometrics to the quality control of snack products.

11. LWT Food Science and Technology, 133 (2020) 110130 D1 if(2020)= 4.952

2. MODEL COMPARISONS

a. Column selection and separation selectivity

A novel procedure has been elaborated to unravel patterns in chromatographic data. A parallel application of the sum of ranking differences and the generalized pair correlation methods has provided the same clustering of chromatographic columns and this pattern corresponds to the physicochemical properties of the highly similar columns studied. The new procedure is able to distinguish columns when other, classical statistical tests cannot. The dissimilarity measure (Euclidean distance) recommended in the literature is not able to find the dominant patterns. Similarly, Snyder's hydrophobic subtraction model leads to considerable information loss.

12. Journal of Chromatography A, 1488 (2017) 45-56 Q1 if(2017)= 3.716

b. Improvement of sum of ranking differences procedure to fit the requirements of sensory data analysis.

More reliable results are gained in case of just-about-right (JAR) data if a consensus of many methods is determined. A specific approach is presented to compare and select JAR attributes of food products. Generalized Pair Correlation Method (GPCM) compares the impact of the JAR variables on overall liking pairwise and the probability weighted difference ordering was applied for ordering the attributes. A special data fusion is suggested based on the sum of ranking differences (SRD), primarily developed for method comparison. SRD was able to rank the JAR variables based on their differences from a benchmark defined by all the JAR evaluation methods in maximal performance. This also enables to group the product attributes. Moreover, it gives recommendations for how to optimize the products based on the results of several JAR methods and helps to gain a more reliable evaluation and selection of JAR attributes. The significant features can be identified easily when the SRD procedure is complemented by the frequencies of consumer evaluations. The same data matrix transposed is suitable to rank the evaluation methods using the average of all evaluation methods (consensus). From among the JAR evaluation techniques, GPCM proved to be closest to the average, *i.e.*, it can be used for substitution of the other techniques.

13.Chemometrics Intelligent Laboratory Systems, 161 (2017) 130-135 Q2 if(2017)=2.701

c. Comparison of regression models based on spectroscopy datasets

Fat and dry material contents (connected to moisture) are one of the most important parameters in the quality control of butter, margarine, and margarine spreads (dairy

spreads). More than a hundred margarine samples were used to model their fat and dry material content based on Fourier transform near infrared spectroscopy in transmission and reflectance modes for the quality control of margarine. We also carried out a systematic comparison of various modeling techniques such as partial least squares regression, principal component regression and support vector machines (SVM). The properly validated SVM models proved to be the best for all four datasets.

14. Analytical Methods, 10 (2018) 3089-3099 Q1 if(2017)=2.000

d. Modeling of eye movements

Eye tracking data were mapped to a low-dimensional space, in which they can be easily clustered. Different food choice tasks were presented, the participants had to choose 1 product of the presented four and later from eight alternatives. A new measure was introduced based on all three consecutive points from the fixations, and the areas of the triangles formed by these three points were computed. The new eye-movement index captures the temporal variation and considers the orientation of the fixation points.

15. Journal of Chemometrics 32 (2018) e3003 Q2 if(2018)=1.847

e. Toxicity prediction

The models were generated with multiple linear regression (MLR), principal component regression (PCR), partial least squares regression, artificial neural networks, and support vector machines (SVM). The comparisons were made by the sum of ranking differences and factorial analysis of variance. The largest bias and variance could be assigned to the MLR method. The generated models were also compared based on their basic performance parameters (R^2 and Q^2). MLR produced the largest gap between these parameters, while PCR gave the smallest. Although PCR is the best validated and balanced technique, SVM always outperformed the other methods.

16. SAR & QSAR in Environmental Research 29 (2018) 661-674 Q2 if(2018)= 2.287

f. Disjoint class modelling *vs*. Classification

The two nonparametric methods: sum of ranking differences (SRD) and generalized pairwise correlation method (GPCM) have been used to rank and group classifiers obtained from six case studies. While SRD and GPCM are sensitive to the reference selection (supervisor), this effect could be eliminated with comparisons with one classifier at a time (SRD-COVAT) and the resulting heatmaps support and validate the grouping pattern found by using the above two techniques. Considering highly different and deviating data sets, soft independent modeling of class analogies has proven to be of weak performance (worst among the studied methods in numerous cases), despite its advantages and unique theoretical background.

17. RSC Advances 8 (2018) 10-2 Q1 if(2018)=3.049

g. Comparisons of distance measures

A combination of methodologies can determine a proper ranking of items. Three well-established metrics: Kendall tau, Spearman footrule, and Cayley distance and a novel metric created by the combination of Cayley and Spearman footrule metrics were compared.

Chemometric data of phytonutrients of tomato varieties and sensometric data of orange juices were used to test the performance of the ranking distance metrics.

18. Journal of Chemometrics 32 (2018) e3011 Q2 if(2018)=1.847

h. Extraction of soil and sediment samples

Three extraction techniques (conventional, microwave and ultrasound) of four sequential element extraction steps from soil and sediment samples were ordered by SRD technique. The concentrations of elements were determined through ICP OES, and the elements were ranked by extraction efficiency. Ordering the elements is useful for three purposes to: i) find possible associations among the elements; ii) find possible elements that have outlying concentrations; iii) detect differences in geochemical origin or behavior of elements. Cross-validation of the SRD values with cluster and principal component analysis revealed the same groups of extraction steps and techniques.

19. Chemosphere 198 (2018) 103-110 D1 if(2018)=5.108

i. Comparison of data fusion methods as consensus scores for ensemble docking

Ensemble docking is a widely applied concept in structure-based virtual screening–to account for protein flexibility at least partly–usually granting a significant performance gain at a modest cost of speed. Nonetheless, there are several fusion rules that can be applied in this procedure. We carried out a detailed statistical comparison of seven fusion rules for ensemble docking, on five case studies of current drug targets, based on four performance metrics. The preferable usage of geometric or harmonic means was proven in ensemble docking.

20. Molecules 24 (2019) Article No: 2690 Q1 if(2019)=3.267

j. Multi-level comparison of machine learning classifiers and their performance metrics

- Several performance parameters give conflicting results frequently in classification models for toxicity. We performed a multi-level comparison with the use of different performance metrics and machine learning classification methods based on molecular descriptors (fingerprints). The effect of dataset composition (balanced *vs*. imbalanced) was decomposed and 2-class *vs*. multiclass classification scenarios was also investigated. We proved that most of the performance metrics are sensitive to dataset composition, especially in 2-class classification problems in QSAR/QSPR analysis.

21. Molecules 24 (2019) Article No: 2811 Q1 if(2019)=3.267

- We summarized the current trends in the QSAR model development for the ADME (absorption, distribution, metabolism, and excretion) and toxicity endpoints in a comprehensive review. The study focused on the machine learning-driven classification models from the last six years with large datasets (above 1000 molecules), to provide a comparative analysis of the applied algorithms, validation protocols, descriptor sets (fingerprints and classical molecular descriptors), software, endpoint-specific performances, and dataset sizes. The thorough meta-analysis showed that the "forest-based" machine learning algorithms still dominates the field of drug safety prediction.

22. Molecular Diversity, Vol. 25, pp. 1409–1424 (2021) Q2 if(2020)=*2.943*

k. Comparison of intercorrelation limits in molecular descriptor preselection

The generation and selection of molecular descriptors (fingerprints) is an essential part of QSAR/QSPR (quantitative structure-activity/property relationship) modeling. Our goal was to propose guidelines for selecting the intercorrelation limit(s) for descriptor selection. QSAR models were generated with a wide range of intercorrelation limits, based on four case studies from the literature, with diverse endpoints. SRD, as a multicriteria decision making tool, can distinguish suboptimal solutions well.

23. Molecular Informatics 38 (2019) Article No: 1800154 Q2 if(2019)=2.741

l. Comparison of validation variants

Three case studies have been selected carefully to reveal similarities and differences in validation variants. In special circumstances, any of the influential factors for validation variants can exert significant influence on evaluation by sums of (absolute) ranking differences (SRDs): stratified or repeated resampling and data set splits (5-7-10). The optimal validation variant should be determined individually. A random resampling with sevenfold cross-validations seems to be a good compromise to diminish the bias and variance. For unknown data structures, a randomization of rows is suggested before SRD analysis. On the other hand, the differences in classifiers, validation schemes, and models proved to be always significant, and even subtle differences can be detected reliably by SRD and analysis of variance (ANOVA) for SRD scores.

24. Journal of Chemometrics 33 (2019) 1-14 Article No: e3104 Q2-Q3 if(2019)= 1.633

m. Modeling of overall liking

Buckwheat-pasta enriched with silkworm powder was produced and numerous technological parameters (*e.g.*, color change, cooking quality, *etc*.) were defined and altered. Generalized pair correlation method (GPCM) was used to assess the sensory attributes influencing the overall liking the most. Application of GPCM in food sensory sciences was introduced by our research group and this project dealt with a practical application. The highest overall liking scores resulted 10 % insect powder content. Insect powder is a suitable material for enriching less liked, but basically healthy products.

25. LWT - Food Science and Technology 116 (2019) Article No: 108542 D1 if(2019)=4.006

n. Multi-objective optimizations

- Sum of Ranking Differences (SRD) is an innovative statistical method that ranks competing solutions based on a reference point. The latter might arise naturally or can be aggregated from the data. We provide two case studies to feature both possibilities. Apportionment and districting are two critical issues that emerge in relation to democratic elections. Theoreticians invented clever heuristics to measure malapportionment and the compactness of the shape of the constituencies, yet there is no unique best method in either case.

26. Plos ONE 15 (2020) e0229209 D1 if(2020)=3.240

- We used atomic force microscopy to measure the surface roughness of polyethylene terephthalate (PET) fibers. Samples were measured multiple times at different locations, in four scan sizes. The surface roughness was expressed in terms of nine roughness

parameters. Simple statistics was not able to detect significant differences in roughness before and after plasma treatment. A factorial ANOVA of sum of ranking differences scores established that (i) the plasma treatment had roughened the PET fiber surface; (ii) the roughness increases with the scanned area in the measured range; and (iii) what the best roughness parameters are in discriminating between surfaces before and after treatment.

27. ACS Omega 5 (2020) 3670-3677 Q1 if(2020)=3.512

- The most widespread advanced technology in drying industry is freeze-drying. Freeze-drying is a fast, efficient method, which preserves the nutritional and sensory quality of fruits and vegetables. Conventional drying methods cannot be replaced by freeze-drying in most of the cases. We optimized conductive drying methods to achieve a sample quality like the freeze-dried samples. A detailed technological evaluation provided the fingerprints of the different methods, which included 16 dimensions of the products. Special conditions of drying Polana raspberry are adequate alternative to freeze-drying.

28. Journal of Chemometrics 34 (2020) e3224 Q2 if(2020)=2.467

o. Development and comparison of novel pasta products

Development of a novel pasta product containing silkworm powder was completed. The new food products were subjected to numerous technological measurements as well as a sophisticated consumer sensory study in order to define the best composition having the highest nutritional qualities and the highest consumer acceptance scores. These measurements enabled us to create not only the technological but the sensory fingerprints of the products. Silkworm enriched pasta samples have good nutritional parameters, and the pasta with 10 % insect content reached the highest overall liking score.

29. LWT Food Science and Technology 116 (2019) 108542 D1 if(2019)=4.006

p. Prediction of consumers' preference

- Eye-tracking measurements give information about the subconscious mind of the participants. This research uncovered factors influence the eye-movements. The latter are often treated as fingerprints of the human interest. Product orientation significantly influenced the time to first fixation and first fixation duration parameters. Stimulus size significantly increased fixation and dwell count, while background showed no significant effects. Significant relationships were found between the number of presented images and eye-movements and decision times.

30. Food Quality and Preference 83 (2020) 103915 Q1 if(2020)=5.565

- Mapping the subconscious mind of participants regarding meat alternatives was done during this project. We aimed to map the mind of US residents from two major regions, New York and California. Altogether 400 participants were involved in a set of studies. The used approach, ConJoint analysis enables us to define and describe the way people think of a given topic. Researcher knows that data from the "many," from total panels, might give overgeneralized, and seeming irrelevant, "bland" results, which do not teach or advance science, and simply "do not work in practice." The introduced mind-sets give a new tool to address the people with newly developed meat-alternative based food products.

31. Sustainability 2020, 12(13), 5352 Q1 if(2020)=3.251

- Sensory evaluation of wines is a widely used method to assess the sensory fingerprint of different wines. It is used by all winemakers and gives the most reliable information about the quality of the wines. Many producers use the sensory fingerprints of their products as a label or brand to prove the superiority of the wines. On one hand, the most widely used method for assessing wine quality (called OIV method) is an overgeneralized one, which cannot differentiate the samples. On the other hand, the ISO created a standardized methodology, called quantitative descriptive method (QDA), which can be adapted to any kind of food or non-food products/services. We compared the two methods by conducting sensory tests of different Hungarian red wines. Although OVI can be performed faster and requires less training, the QDA approach gives much detailed information about the samples. QDA explains wine faults better and its results give more feedback to producers than OIV.

32. Journal of Chemometrics 34 (2020) e3219 Q2 if(2020)=2.467

- Polyphenols are one of the most important compound classes in wines, originating mostly from the grape, and to some extent from the barrels used for aging. We have measured the concentration of two major components, namely trans-resveratrol, and anthocyanin during the 24-month aging process of three different wine varieties. A Factorial ANOVA of aging time and the wooden barrel showed that the chemical "fingerprints" of the wines are changing during the aging period and the wine variety is also a significant factor based on the anthocyanin and *trans*-resveratrol content of the wines.

33. Acta Alimentaria 48 (2019) 349-357 Q3 if(2020)=0.650

3. MULTIVARIATE STUDIES OF PRACTICAL RELEVANCE

a. Analysis of subconscious decision factors of food consumers

Insect-based food products were chosen as the subject of our study, which provided good quality data and the results proved to be good enough to create a basis of our next studies. Subconscious decision factors may influence the consumers' decisions differently. Therefore, segmentation (or clustering) of consumer groups is essential. However, the key to understand the created clusters is to define their way of thinking, the fingerprint of their subconscious behavior. In order to test our methods on these kinds of data, a highly divisive food product was used to map the opinions of consumers.

34. Food Quality and Preference 59 (2017) 81-86. D1 if(2017)=3.652

b. Ranking and grouping of performance parameters

We compared various performance parameters to be able to characterize the different fingerprint models. Again, the Springer Publishing House invited us to write a chapter.
We have revisited the vivid discussion in the QSAR-related literature concerning the use of external *versus* cross-validation and have presented a thorough statistical comparison of model performance parameters with sum of (absolute) ranking differences, SRD and analysis of variance (ANOVA). It was shown unambiguously for both case studies that the performance merits are significantly different, independently from data preprocessing. While external merits are generally less consistent (farther from the reference) than training and cross-validation based merits; a clear ordering and a grouping pattern of them could be acquired. The results presented here corroborate our earlier findings that external validation

is not necessarily a wise choice and is frequently comparable to a random evaluation of the models.

35. Chapter 3. pp. 89-104 in Advances in QSAR Modeling, Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences (editor: K. Roy)

c. Mushroom composition

The elemental composition of the Agaricus spp. according to the area of origin has been determined:

36. Journal of Food Composition and Analysis 72 (2018) 15-21 Q1 if(2018)=2.994

d. Comparison of insect species for human consumption

This paper revealed which insect species should be proposed for human consumption. The latest developments of sum of ranking differences allowed evaluation of literature data from the past decades. There are no "best" or "healthiest" insect species because the assessment of "best" heavily depends on the nutrients in question. Adult mealworms should be chosen for best protein composition, while waxworm larvae (*Galleria mellonella*) have the highest mineral content among the evaluated items.

37. Innovative Food Science and Emerging Technologies 52 (2019) 358-367 D1 if(2019)=4.477

e. Prediction of consumers' preference

The paper discusses an evergreen question: the problems of too many samples in food sensory evaluations. Altogether 15 samples were created and tested by 42 participants. Attribute ranks provided by the generalized pair correlation method were used to define the attributes, which influencing the acceptance the most. Poppy seed was preferred over poppy seed flour, whole poppy seeds were preferred over conched and ground poppy seed and lower amounts (2 % and 4 %) were preferred over higher amounts (6 %, 8 % and 10 %).

38. LWT - Food Science and Technology 103 (2019) 162–168. D1 if(2019)=4.006

f. Accumulation of metals in foods

We studied the metal uptake behavior of *Russula cyanoxantha* mushroom relying on the soil properties. We sampled mushroom and soil from six forests according to an urbanization gradient, and two city parks in Cluj-Napoca (Romania). The elements were quantified using inductively coupled plasma optical emission spectroscopy (ICP-OES). The concentrations of some elements differed significantly. We observed a definite negative trend in the mineral accumulation potential of this fungus along the urbanization gradient. The fungus turned from a cadmium-accumulator to a cadmium-excluder. This highlights a positive environmental influence of the urbanization over the toxic metal uptake of *R. cyanoxantha*.

39. Chemosphere, 238, Article No: 124566 (2020) Q1 if(2020)= 7.086

g. Prediction of corneal permeability

In this work two extensive QSAR models have been developed with outstanding prediction performance and 189 diverse compounds for their corneal permeability and corneal membrane retention. Extended connectivity fingerprints (ECFP) with classical 1D, 2D and

3D molecular descriptors were applied for PLS regression models. There is no significant correlation between the corneal permeability and the CaCo-2 permeability, jejunal permeability or the blood-brain partition coefficient. The goodness of the models (R2) was above 0.90 for the training and above 0.85 for the validation in both cases, thus they are capable for the determination of the corneal permeability of the potential drug candidates in a very short time with appropriate precision.

40. Journal of Pharmaceutical and Biomedical Analysis, 203 114218 (2021)

Q1 if(2021)= 3.935

h. Multicriteria decision making for evergreen problems in food science

Sum of ranking differences (SRD) method has gained a well-deserved popularity in the field of chemometrics. Numerous applications by multiple research groups all over the world have been published in the last decade and the number of publications show an increasing tendency. Despite its high popularity in chemometrics, there are only a limited publications from food sciences. Due to the nature of the method, SRD can provide solutions to a range of problems arising in food science projects, therefore our research team aimed to introduce some of these applications for food scientists. The paper introduces twelve case studies covering various fields in food science such as analytics in food chemistry, food chemistry, food technology, food microbiology, food quality control, food sensory analysis. SRD was used to analyze already published data of the papers and was able to provide additional information presented by the original authors, therefore SRD proved its significant potentials in food sciences, too.

41. Food Chemistry, 344, 128617 (2021) D1 if(2020)=7.514

4. SOFTWARE

a. FingerPrint Kit - Python-based cheminformatics package for fingerprint-related tasks. https://github.com/davidbajusz/fpkit/, DOI: 10.5281/zenodo.1217969

The FPKit Python package was updated for molecular fingerprint formats through the popular cheminformatics package cinfony (http://cinfony.github.io/).

b. SRDonline R version 3.5.1 (Feather Spray) Version No: v0.1
Availability: https://attilagere.shinyapps.io/srdonline/
Authors: Attila Gere, Klára Kollár-Hunek, Károly Héberger,

5. SCIENTOMETRICS

In summary, 41 publications were written in the five years (4+1) of the project, which clearly shows the effectiveness of the project. The sum of the impact factors is 185.627 for the total period. This means that the average impact factor per paper was 4.759 (see Figure 1.) for the five years.

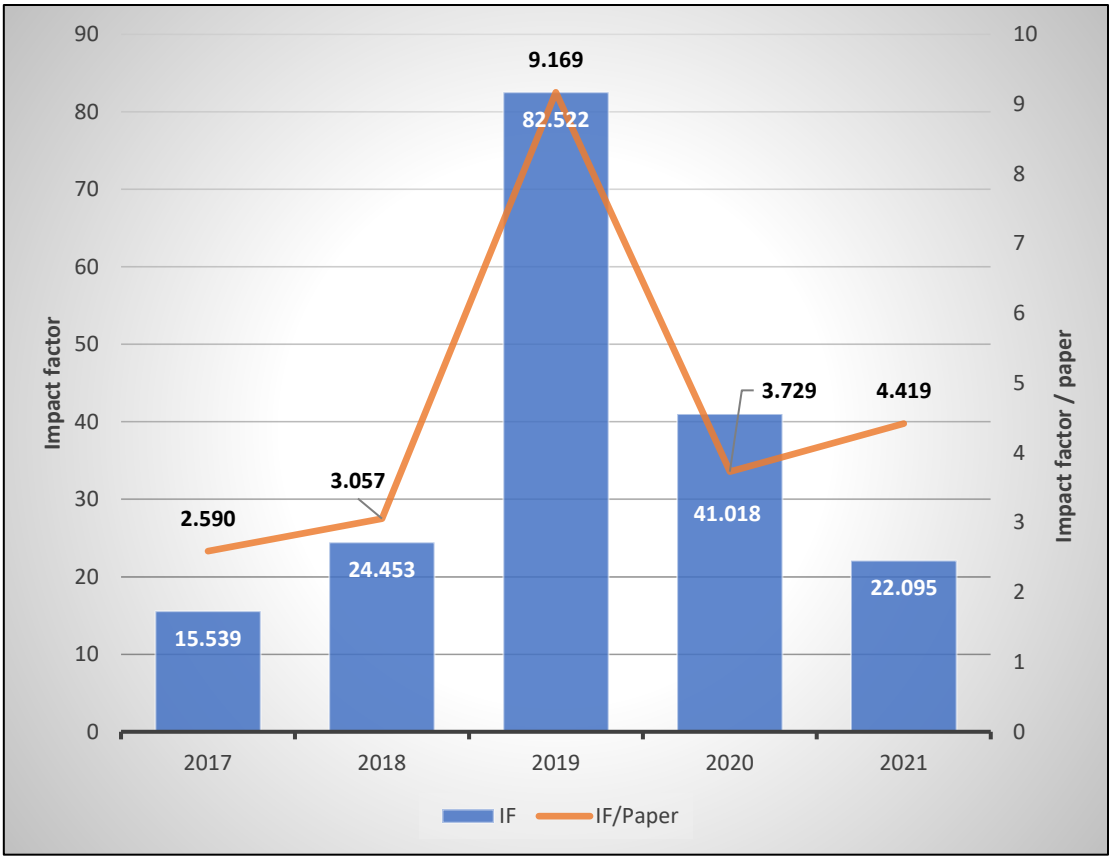
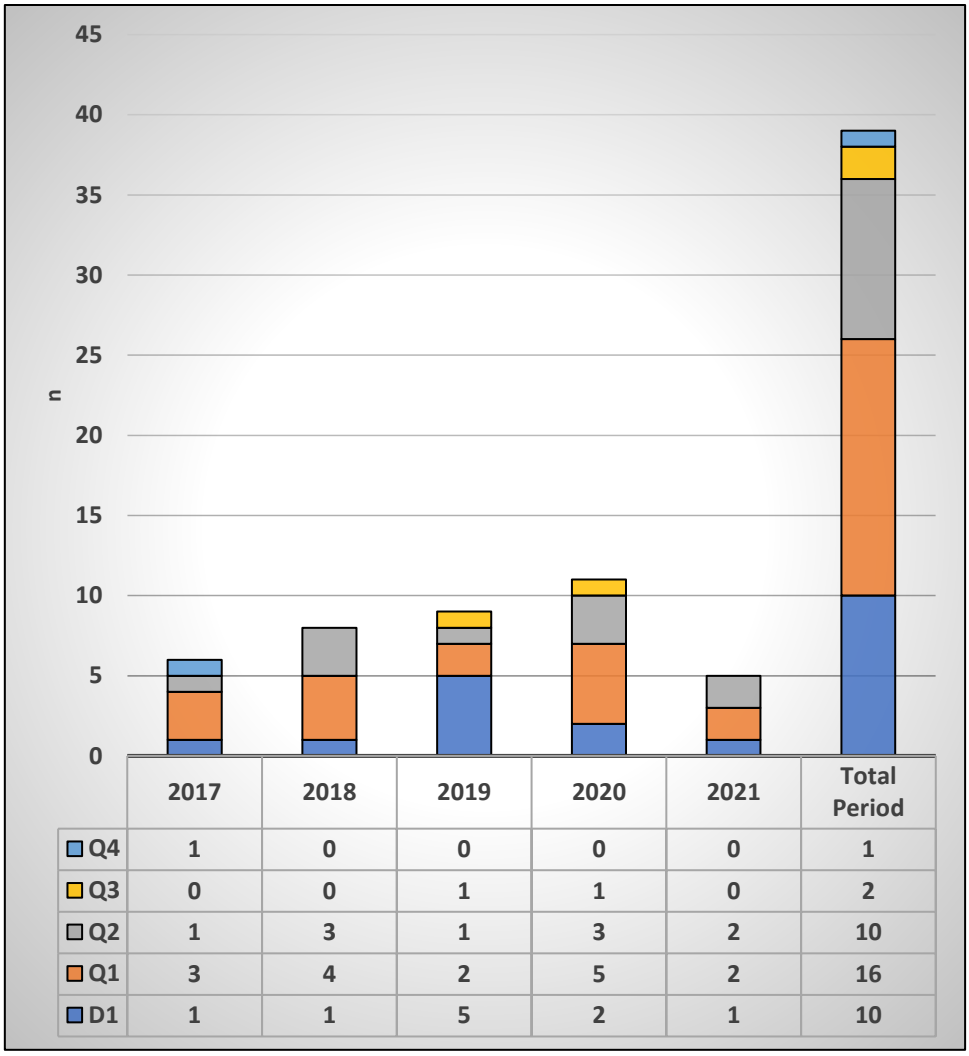


Figure 1. The distribution of the impact factors and impact factor / paper for each year.

From another point of view, we can evaluate the number of papers based on the different SJR quartiles journals. In Figure 2. we present, that most of the papers in each year were published in D1-Q2 journals. It can support that our project was successful from the quality side as well. As it can be expected, the number of publications has increased from the starting year to the 4th year.



	2017	2018	2019	2020	2021	Total Period
Q4	1	0	0	0	0	1
Q3	0	0	1	1	0	2
Q2	1	3	1	3	2	10
Q1	3	4	2	5	2	16
D1	1	1	5	2	1	10

Figure 2. The number of publications in the different quartiles for each year and the total period.