Vol. 10, No. 3, 2019
ISSN 2233-4203/ e-ISSN 2093-8950
ARTICLE
www.msletters.org | Mass Spectrometry Letters

Isomer Differentiation Using *in silico* MS^2 Spectra. A Case Study for the CFM-ID Mass Spectrum Predictor

Boris L. Milman[1]*, Ekaterina V. Ostrovidova[1,2], and Inna K. Zhurkovich[2]

[1]Institute of Experimental Medicine, ul. Akad. Pavlova 12, Saint Petersburg 197376, Russia
[2]Institute of Toxicology, ul. Bekhtereva 1, Saint Petersburg 192019, Russia

Received July 29, 2019; Revised September 14, 2019; Accepted September 15, 2019
First published on the web September 30, 2019; DOI: 10.5478/MSL.2019.10.3.93

Abstract : Algorithms and software for predicting tandem mass spectra have been developed in recent years. In this work, we explore how distinct *in silico* MS^2 spectra are predicted for isomers, i.e. compounds having the same formula and similar molecular structures, to differentiate between them. We used the CFM-ID 2.0/3.0 predictor with regard to (a) test compounds, whose experimental mass spectra had been randomly sampled from the MassBank of North America (MoNA) collection, and to (b) the most widespread isomers of test compounds searched in the PubChem database. In the first validation test, *in silico* mass spectra constitute a reference library, and library searches are performed for test experimental spectra of "unknowns". The searches led to the true positive rate (TPR) of (46-48 ± 10)%. In the second test, *in silico* and experimental spectra were interchanged and this resulted in a TPR of (58 ± 10)%. There were no significant differences between results obtained with different metrics of spectral similarity and predictor versions. In a comparison of test compounds vs. their isomers, a statistically significant correlation between mass spectral data and structural features was observed. The TPR values obtained should be regarded as reasonable results for predicting tandem mass spectra of related chemical structures.

Keywords : tandem mass spectrometry, isomers, prediction of mass spectra, mass spectral libraries, prediction performance, structural similarity

Introduction

Reference mass spectral libraries are essential data resources in non-target chemical analysis.[1-3] They have been used in both classical electron ionization mass spectrometry (MS) of volatile compounds and electrospray ionization tandem mass spectrometry combined with collisional activation (MS^2) which is advantageous for analysis of non-volatile compounds. However, MS^2 libraries have limited application because they do not cover many important compounds (e.g., metabolites and pollutants). Until quite recently, the compound sets whose MS^2 spectra had been entered into the libraries were measured in not more than the low tens of thousands.[1,2,4] This is much less than the number of known chemical compounds (tens of millions[5,6]) and most compounds are non-volatile, as can be concluded from their molecular masses (MM).

The generation of full MS^2 libraries is certainly a time-consuming and expensive enterprise. There is another way to generate these libraries, which is based on the prediction of the above spectra with the use of pertinent algorithms and corresponding software. A similarity between *in silico* mass spectra and experimental ones of the same compounds can be considered as the rate of the prediction performance. In the ideal but unlikely case, they are the same spectra.

Such computer predictors have been developed in recent years; some of them were already used in building of MS^2 libraries.[2,4] Estimating the above-mentioned similarity index/metric is involved in library searches resulting in the best answers for experimental spectra of unknown analytes. Thus, such searches demonstrate a mass spectrum prediction performance, and they should be a part of ongoing validation tests needed to estimate a progress level of mass spectrum predictors and to explore the possibilities of their wide use in MS^2 analysis.

In silico MS^2 spectra can find another use, namely in the case of the inability to record any experimental tandem mass spectrum or one of good quality that is sufficient for advanced searches in a reference MS^2 library of an

*Reprint requests to Boris L. Milman
E-mail: bmilman@mail.rcom.ru; bormilman@yandex.ru

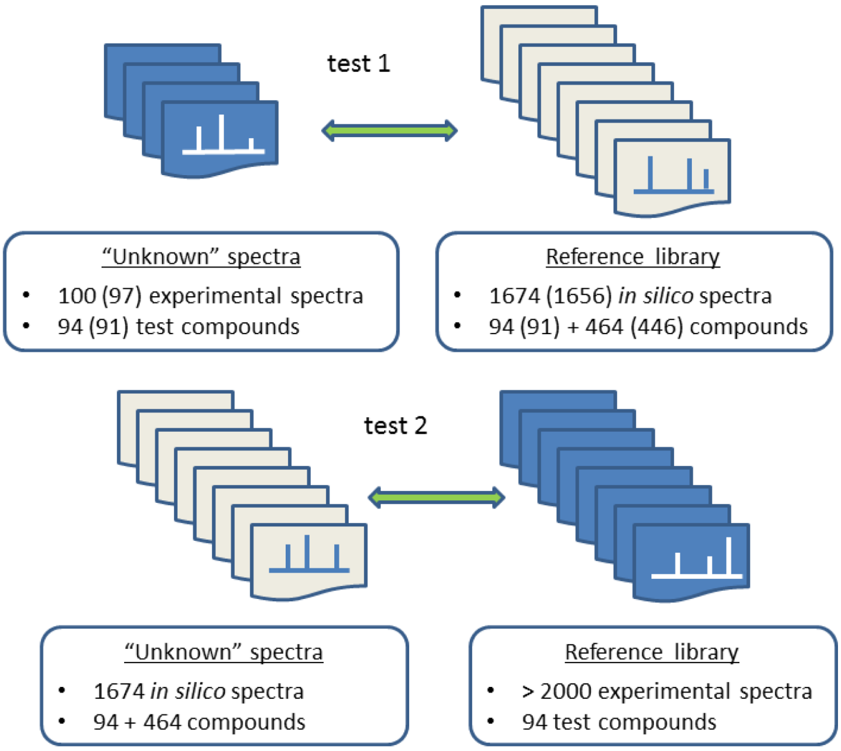

Figure 1. The schematic for comparison of MS^2 spectra. The CFM-ID 2.0 software was the starting predictor; the data for further calculations using the version 3.0 are in parentheses. Various spectra of the same compounds were basically differed in collision energy.

experimental origin. Here, it would be appropriate to predict tandem spectra for some candidate compounds derived from measuring exact ion masses and searching chemical databases and subsequently to perform MS^2 library searches with the *in silico* spectra as the features of unknown analytes.[7] This approach to identification also requires its own validation.

In this work, we would like to further study mass spectrum prediction performance by comparing *in silico* and experimental MS^2 spectra. It would be appropriate to see how unique *in silico* MS^2 spectra are generated for compounds having similar molecular formulas and structures. Among them, we focused on isomers, i.e. compounds having the same formula. They represent the most troubled sort of molecular similarity, which often results in very similar mass spectra and therefore unreliable differentiation of isomers.

To achieve the objective defined, pertinent collections of predicted and experimental MS^2 spectra should be formed. Since *in silico* spectra contain exact mass peaks by the nature of the prediction process, experimental spectra should be properly acquired by high resolution mass spectrometry (HRMS). We started by sampling a random subset of high resolution MS^2 spectra from the MassBank of North America (MoNA) collection of experimental spectra.[8] *In silico* spectra were generated for every compound whose spectra had been sampled (call them "test compounds") and for five of the most widespread (abundant, cited) isomers found for every test compound in the PubChem portal.[5] Both experimental and predicted spectra were entered into the special MS^2 libraries. The degree of similarity between *in silico* and experimental spectra was estimated in two different groups of library searches (Figure 1). First, the sample of experimental spectra and corresponding *in silico* spectra were considered as "unknown" and reference ones, respectively. This validation test simulated an identification procedure based on the use of a reference *in silico* mass spectral library. In the second examination, the two MS^2 subsets were reversed: all *in silico* spectra were considered as "unknown" data and all experimental ones of test compounds (not just the sample) were the reference spectra. The second test stimulated an identification option conceivable for HRMS in the absence of any or high-quality MS^2 data.

As usual, the two outcomes of library searches were eventually counted. The search and identification result observed when "unknown" spectrum and the most similar reference one (the 1st rank match) belonged to the same compound was defined as true positive (TP). The best match observed for different compounds implied a false positive result (FP). The true positive rate (TPR, sensitivity) was calculated in both tests by taking into account different spectral similarity indices (metrics). By definition, TPR was the percent of TP among all the searches. This quantity, TPR, was considered to be the performance measure/rate of mass spectral prediction.

We used CFM-ID 2.0/3.0 software as the predictor.[9-13] This is based on the probabilistic model ("Competitive Fragmentation Modeling") for MS^2 fragmentations and employs a machine learning approach for determining model parameters (e.g., probabilities of transitions between fragment ions from experimental mass spectra). There were five reasons to prefer this predictor. This program (1) predicts not only ion masses but also peak intensities, (2) it was successfully used in interlaboratory comparisons in the identification of low-molecular compounds,[10] (3) it is freely available online at the website,[11,13] (4) it calculates mass spectra rapidly. After all, (5) the problem of isomer identification/differentiation by this software was indirectly engaged in the work[9] that provides some basis for comparing the results that they achieved to our results.

In addition to the main purpose of the study, which was to determine a level of prediction trueness for mass spectra of related compounds, there were several other goals of the research. They were to find out the effects of various metrics of spectral similarity on TPR and to correlate this quantity to a general structural similarity of isomer compounds corresponding to true and false results of library searches.

Methods

Samples of mass spectra and test compounds

The formation of the mass spectral subset sampled from the MoNA collection was detailed in the article.[14] Briefly, the file of various 44157 MS^2 spectra was imported from MoNA[8] and downloaded into the NIST MS Search 2.3 software.[15] According to our estimation, the file contained approximately 1.1×10^4 positive ion high-resolution tandem mass spectra of a satisfactory quality. The "unknown" spectra for the first test and reference spectra of the second test were originated from that initial file. Two independent raw subsets of 1% of initial spectra were *randomly* sampled to form two selections of spectra of "unknown" compounds. Both samples resulted in statistically indistinguishable outcomes of library searches in the previous research;[14] here only one of them was used. The sample was largely refined before the use. One/two-peak spectra, very noisy ones, low-resolution and negative ion data, unique spectra of any compound, and some other items unacceptable for identification for a variety of reasons were removed from the sample. The rest of the 109 items were used for the research.[14]

From these 109 spectra, 100 of them were considered as "unknown" spectra in the first test using the version 2.0. The rest were MS^2 spectra not recorded for $[M+H]^+$ precursors (as demanded by the CFM-ID[11] program) or spectra belonging to compounds not found in PubChem.[5] Those 100 spectra were originally acquired for 94 unique compounds (test compounds, Table 1), with six of them having two different spectra. For every from 94 items, five of the most widespread/abundant isomers were found in Pubchem. Optical and geometrical isomers of test compounds known to result in the same or very similar mass spectra were not taken into account. Isomers were searched in PubChem by corresponding molecular formulas. Compounds that were found were ranked by the relevance feature and correlated with a data count ("the degree of annotation"[16]). The last feature could reasonably be expected to depend on a relative spread/abundance/ citation of chemical compounds.[3] The top five isomers of every test compound were eventually selected. The final subset of test compounds and their isomers contained 558

Table 1. Test compounds and mass spectrometers used for recording MS^2 spectra.[8]

#	Name	Test 1*	Test 2*
1	1-[4-(2-Chlorophenyl)piperazin-1-yl] ethanone	1	1
2	17α-Ethynylestradiol	1	1
3	1-Chlorobenzotriazole	2	1,2
4	1-Naphthonitrile	3	3
5	1-Octadecylamine	3 **	3
6	2,4-Dimethylphenylformamid	2	1,2
7	2-Amino-3-[2,4-dichloro-6-hydroxy-3-[2-[imidazole-1-carbonyl(propyl)amino] ethoxy]phenyl]sulfanylpropanoic acid	1	1
8	2-Bromoaniline	3	3
9	2-Phenylethylamine	3	3
10	2-Toluenesulfonamide	1	1
11	3,4-Methylenedioxy-N-methylamphetamine	1	1, 2
12	3-[(2-Chloro-1,3-thiazol-5-yl)methyl]-5-methyl-1,3,5-oxadiazinan-4-imine	1 **	1
13	3-Bromo-N,N-dimethylaniline	3	3
14	3-Methylimidazo[4,5-f]quinolin-2-amine	3	3
15	4,5-Dichloro-2-n-octyl-3(2H)-isothiazolone	3	2,3
16	4,7-Phenanthroline	3	3
17	4-[Formyl-(6-methoxyquinolin-8-yl)amino]pentanoic acid	1	1
18	4-Aminoantipyrine	3	2,3
19	4-Amino-N,N-dimethylbenzenesulfonamide	3	1,3
20	4-Methoxycinnamic acid	1	1
21	5-Methoxyflavanone	2	2
22	Adenine	3	1,3
23	α-Codeimethine	2	2
24	Amidosulfuron	1	1

Table 1. Contined.

#	Name	Test 1	*Test 2 *
22	Adenine	3	1,3
23	α-Codeimethine	2	2
24	Amidosulfuron	1	1
25	Amisulpride N-Oxide	1	1,2
26	Ampicillin	2	1
27	Apomorphine	2	3
28	Aspirin	1	1
29	Asulam	3	3
30	Atenolol	3	2,3
31	Atenolol-desisopropyl	1	1,2,3
32	Azobenzene	3	3
33	Benalaxyl	1	1,3
34	Benzidine	3	3
35	Bicalutamide	3	2,3
36	Caffeine	3	1,2,3
37	Carbamazepine-10,11-epoxide	3	2,3
38	Cetirizine N-Oxide	2	1,2
39	Chlorcyclizine	1	1
40	Chlorfenvinphos	3	1,3
41	Chromomycin A3	2	2
42	Cimetidine	1	1
43	Clopidogrel carboxylic acid	3	3
44	Dextromethorphan	3	2,3
45	Diatrizoate	3	3
46	Dihydroergotamine	2	2
47	Dioxoaminopyrine	3	3
48	Diphenyl phthalate	1	1
49	Doxylamine	1	1,2
50	Erythromycin	3	1,2,3
51	Estrone	3	1,3
52	Ethopabate	2	2
53	Ethoprop	1	1
54	Fenoterol	2	2
55	Fenthion	1	1,2
56	Fenthion-sulfoxide	1	1,2
57	Florfenicol	1	1
58	Forchlorfenuron	1	1
59	Gabapentin	2	2,3
60	Genistein	4	4
61	Ibuprofen	3	3
62	Iomeprol	3	3
63	Isoproturon	3 **	2,3
64	Kresoxim-methyl acid	1	1
65	L-Arginine	1	1,3

Table 1. Contined.

#	Name	Test 1	*Test 2 *
66	Levamisole	3	2,3
67	L-Threonine	2	1
68	Mesotrione	3	2,3
69	Metamitron-desamino	2 **	2,3
70	Metformin	3	1,2,3
71	Methsuximide	1	1
72	Methylprednisolone	3	2,3
73	Mirtazapine	1	1,2
74	N'-(2,4-Dimethylphenyl)-N-methylformamidine	3	2,3
75	N,N-Dimethyldecylamine N-Oxide	3 **	3
76	N2-Isobutyryl-2'-deoxyguanosine	2	2
77	N4-Acetylsulfamethoxazole	3	2,3
78	N-Desmethylvenlafaxine	3	3
79	Noscapine	2	1,2,3
80	Penconazole	3	1,3
81	Perindopril	1	1,2
82	Phenazine	3	1,3
83	Phenylbutazone	1	1,2
84	Prednisolone	2	2,3
85	Pymetrozine	3	3
86	Rimsulfuron	3	3
87	Sulfadiazine	2 **	2
88	Sulfadimethoxine	3	1,2,3
89	Tenoxicam	2	2
90	Terbinafine	1	1
91	Ticlopidine	1	1
92	*trans*-Zeatin	2	2,3
93	Tri(butoxyethyl) phosphate	1	1,3
94	Venlafaxine	2	1,2,3

* 1: Orbitrap (Q-FT), 2: Q-ToF, 3: Orbitrap (IT-FT), 4: different instrument.

** Two experimental spectra per a compound

unique compounds (there were some coincidences among 94 × 6 = 564 potentially different chemical species). Their 1674 MS^2 spectra were calculated by CFM-ID 2.0[11] with three different spectra for each compound predicted at three collisional energies.

All the predictions were repeated employing the version 3.0, and three of them were failed. As a result, 97 *in silico* spectra of test compounds and corresponding isomer mass spectra (Figure 1) were taken into account.

Library search

In the first test (Figure 1), each one from the “unknown” spectra (100 or 97) was compared to the predicted ones by performing searches in the library of 1674 (1656) *in silico* spectra. In the hit list of the library search, the formal 1st rank of the reference spectrum of the same compound as "unknown" meant the TP result of the search and, consequently, conventional true identification. Following all of the searches, TPR values were calculated. This rate was considered not only as the performance of library searches but also that of mass spectrum prediction. The TPR quantity might depend on an index/metric of spectral similarity.

The second test was the reverse comparison where *in silico* spectra were considered as ones of "unknown" compounds (Figure 1). All experimental spectra of the test compounds (> 2000 items initially contained in the MoNA collection including the spectral sample) constituted the reference library. The TPR calculation was more complicated than for the first test. For every test compound (under the research arrangement, six compounds were taken into account twice, see above) and each one from its five isomers, the spectrum was chosen from three *in silico* ones which led to a maximum value of the similarity metric/index with one or another reference spectrum of corresponding test compound. The search result was TP when the similarity metric was the highest for one or another *in silico* spectrum of the test compound. Conversely, the highest value of one or another isomer *in silico* spectrum resulted in FP. Then, the TPR was calculated for all 100 search groups and for every metric. The second test was carried out for only the version 2.0 and two of four similarity metrics (see below).

In library searches using the above-mentioned MS Search program, we chose the relevant search option of Identity, MS/MS. The previous[14] *m/z* tolerance range of ± 0.01 Da was conventionally set for matching both precursor and product ions. This range simulated the search for HR mass spectra. The option of ignoring peaks in the 1.6 Da range around precursor *m/z* was set to exclude some false matches as was preliminarily proved. All of the spectral similarity indices/metrics were automatically calculated by the MS Search program. The metrics were as follows: (a) the dot product of the "unknown" and library spectrum (DotProd in the MS Search 2.3 program), (b) the reverse dot product (Rev-Dot, the close index which ignores non-matching peaks in the unknown spectrum), (c) the modified dot product[17] common for NIST MS Search software (denoted as Score), and (d) the number of "unknown" spectrum peaks matched to those in the reference spectrum (NumMP) within the mass tolerance. The first three indices have a range from 0 to 999 with higher values indicating closer spectral similarity. The NumMP metric is commonly a low number, the average is six peaks. This might be expressed as the percentage in relation to its maximum value (i.e., the number of peaks in the "unknown" spectrum matched to itself). In some cases, the same NumMP values were found for reference spectra of several compounds. In this case, matching metric values and corresponding candidates for identification were conventionally ranked by the DotProd index.

The statistical errors[18] of result rates (assuming that the sample size of mass spectra under consideration, is much smaller than the initial MoNA size) as well as the significance of their differences[19] were estimated at 95% probability. For these rather small samples, the error level was relatively high (up to 22% rel.). On the other hand, with small samples, it was possible to manually review many spectral matches that provided generally reliable conclusions.

Structural similarity

The structural similarity was estimated by Tanimoto indices which were calculated at the site.[20] The Tanimoto coefficient is defined as $x/(x + y + z)$, which is the proportion of the substructures shared among two chemical compounds divided by their sum. The quantity x is the number of subunits common in both compounds, while y and z are the number of substructures that are unique in one or the other compound, respectively. There are two different sorts of substructures/subunits (structural descriptors). They are (a) atom pairs (AP), defining the AP Tanimoto index, and (b) the maximum common substructures (MCS), providing the estimation of the MCS Tanimoto index. Both Tanimoto indices have a range from 0 to 1 with lower values indicating lesser similarity than higher ones. The statistical significance of the correlation of structural similarity and spectral ones was estimated at the site.[21]

Results and Discussion

Rates of library searches

Results of library searches for different similarity metrics (see above) are given in Table 2. Commenting on this Table, the amazing thing that should be noted for Test 1 is that all metrics resulted in the same or very similar TPR. It is for this reason that we limited the estimates to only two (common) similarity indices and the only predictor version in Test 2. Here, more searches were performed and no differences were also found between DotProd and NumMP. Thus, choosing a certain metric and software version was of no particular importance at this statistical uncertainty level. Because the NumMP index counting only ion masses provides similar search outcomes, correct prediction of mass predominates over that of peak intensity. For the same compound, different metrics take into account the same mass values or many of them (Rev-Dot) and thus resulted in similar (a) orders of metric values of isomers and (b) eventual TPRs. Another conclusion is that the later CFM-ID version 3.0 does not ensure significant progress in mass spectra prediction of

Table 2. True positive rates.

Metric	TPR (%)		
	Test 1, 2.0	Test 1, 3.0	Test 2, 2.0
Dot product (DotProd)	46 ± 10	48 ± 10	58 ± 10
Reverse dot product (Rev-Dot)	46 ± 10	47 ± 10	not estimated
NIST-modified dot product (Score)	46 ± 10	46 ± 10	not estimated
Number of matching peaks (NumMP) *	46 ± 10	48 ± 10	58 ± 10

* The conditional estimate (see above) taking into account that the metric value might be the same for the spectra of several compounds

compounds under the study. Such prediction improvement was manifested mainly for lipids[12] that were not available in the random compound sample (see above). It should be additionally noted that there were no substantial differences in TPRs obtained for different tandem instruments (Table 1).

With this error level, differences between both tests were insignificant. Taking into account that approximately one in two *in silico* mass spectra of test compound were predicted better than corresponding isomer spectra, the TPR values obtained, 46-48% and 58%, should be regarded as the satisfactory level of predicting tandem mass spectra of similar chemical structures.

With regard to isomer differentiation, our results should be matched to previous ones achieved in the work of the authors of this predictor.[9] There were comparisons of predicted MS^2 spectra of approximately 1000 metabolites to experimental spectra of these metabolites and compounds similar to them in MM. In that research, the TPR was only 10-12% with 88-90% of the candidate compounds belonging to different groups of isomers. Thus, it was also essentially a test for differentiation of isomers with the same predictor software and worse outcomes (10-12% vs. 46-48% or 58%). There may be several reasons to explain this difference. Possibly, the discrepancy between two results is because an earlier PubChem release was used in the work[9] as the source of structures. Furthermore, different metrics of spectral similarity were exploited in our research and in works.[9,12] Another possible reason was that optical and geometrical isomers having very similar mass spectra were not excluded from the list of candidate compounds as opposed to our research (see above). It also can be concluded that in our case only the most widespread candidate compounds were taken into account. Their structures (and consequently mass spectra, see below) might be more dissimilar and discriminated with the higher rate.

Mass spectral vs. structural similarity

It is reasonable to suppose that the greater the structural difference of a test compound and its isomer, the better the chance that the *in silico* mass spectrum of the first compound rather than that of the isomer will be more similar to the test experimental spectrum. In our data, a close similarity between experimental and *in silico* spectra of the same compounds, if achieved, is commonly semi-quantitative. That is, *m/z* values of principal ion fragments, but not necessarily the order of relative intensities of corresponding peaks, are the same (Figure 2, a, b). The similarity between the *in silico* spectra of candidate isomer compounds and the corresponding experimental spectra of the test compound is often much smaller (Figure 2, c, d). In this context, some *principle of minimalism* should be advanced in solving identification problems of several isomers. This is that not very close similarity of *in silico* and experimental mass spectra of the compound under identification is of the first value, a minor resemblance of *in silico* mass spectra of other candidate compounds in relation to the compound to be identified may be of increasing importance.

The relationship between the similarity of the mass spectra and that of structures was explored in detail by estimating the correlation between the DotProd metric derived from Test 2 and the corresponding Tanimoto indices of structural similarity.[20] The difference between two DotProd values obtained in the library searches was calculated for every test compound. The first one was the best (i.e., the maximum value from three ones) similarity index between *in silico* and reference/experimental mass spectra of the test compound. This quantity can be denoted as DotProd(test) (e.g., 774, Figure 2 for spectra a and c). The second index was the maximum value from many tens of pair comparisons between all *in silico* spectra of isomers of that test compound and all its experimental spectra. This is the DotProd(candidate) value (e.g., 273 in Figure 2, b, d). The difference of DotProd(test_candidate) = (DotProd(test) - (DotProd(candidate)) is positive (501, Figure 2) when the mass spectra of test compounds are predicted better than their isomers and negative if isomer candidates provide the better/false prediction and correspondingly higher DotProd values. The DotProd(test_candidate) difference as the prediction performance was further correlated with the structural similarity of corresponding pairs of test compounds and their "the best" isomers (Figure 3).

Figure 3 demonstrates that the DotProd(test_candidate) variable decreases with both coefficient of structural

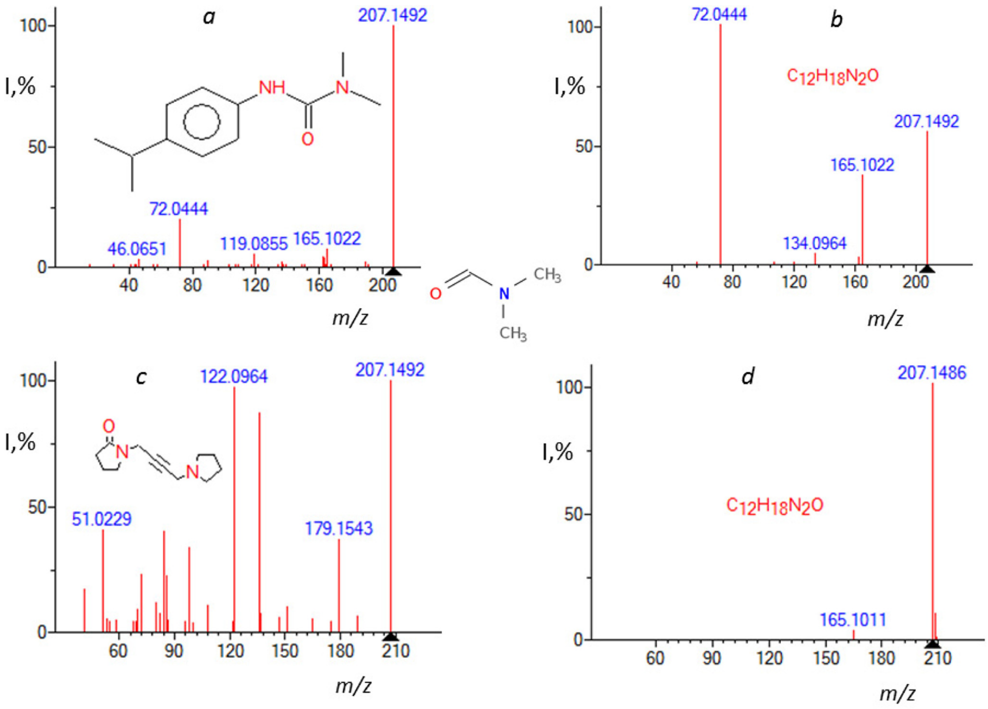

Figure 2. Test 2, *in silico* mass spectra of isoproturon (a) and its isomer ("the best" one), oxotremorine (c), and reference experimental spectra of isoproturon (b and d, respectively) most similar to these *in silico* spectra by DotProd indices. The last are 774 and 273, respectively, which are the maximum values for this test compound and all of its isomers. The pair structural similarity (the common substructure is in the centre; AP Tanimoto: 0.0194, MCS Tanimoto: 0.2000) is one of the lowest among our data; the DotProd difference (501) is one of the highest values.

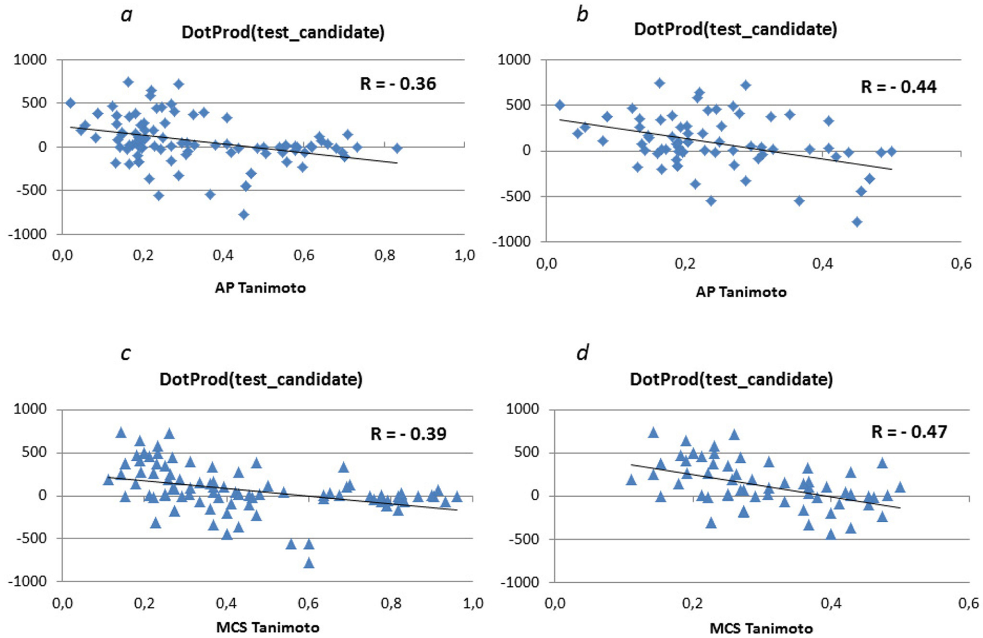

Figure 3. Test 2, the dependences of the DotProd(test_candidate) quantity as the relative prediction performance of tandem mass spectra on corresponding Tanimoto coefficients as the pair similarity of structures of test compounds and their "the best" isomers. Dependences *a* and *c* include all of the data, *b* and *d* exclude points with strong structural similarity (Tanimoto values > 0.5). All of the R reverse correlation coefficients are significant for the probability 95%.

similarity of corresponding pairs of test compounds and "the best" isomers. The R correlation coefficient (reverse correlation) is not strong but statistically significant. It is remarkable that the exclusion of data for relatively strong structural similarity (Figure 3, b, d) improves the correlation. Clearly, this is the expected result because many FPs are hardly avoidable when mass spectra for structurally very similar compounds are predicted. With AP Tanimoto and MCS Tanimoto being > 0.5, (i.e., in the case of more similar structures) TPR was only 43% and 39%, respectively. If more dissimilar structures are compared (AP Tanimoto and MCS Tanimoto are ≤ 0.5), 65% and 70% of true identifications were recorded. It should be noted that a stronger correlation of similarity indices of spectra and structures is hardly expected. The reason is that the difference in mass spectra (i.e., the dissimilarity in parent ion fragmentation) depends not only on the general structural difference but also on special fragmentation features determined by particular functional groups on their protonation. A noteworthy detail is that we did not find any single substructure ensuring that many mass spectra of test compounds are predicted worse than their isomers.

Conclusion

We advanced validation tests for the prediction process of tandem mass spectra. In these, the prediction results were compared with compounds of the same molecular formula and a relatively similar structure which eventually were chosen by the feature of relatively high occurrence/ abundance in the literature or internet databases.

Two kinds of non-target mass spectrometry analysis were stimulated in the tests. (1) The reference mass spectral library consists of *in silico* mass spectra. Library searches are performed for experimental spectra of unknown compounds. (2) With a given molecular formula as a reliable result from the use of HRMS, a search in chemical databases lead to several candidate molecules of isomers which could be further ranked by their abundance. Mass spectra of highly ranked candidates can be predicted and compared to the reference library of mass spectra of experimental origin. The top position (1st rank) of reference mass spectra in the hit list provisionally determines a true positive result.

We used the CFM-ID 2.0/3.0 predictor with regard to (a) test compounds whose experimental mass spectra had been randomly sampled from the MoNA collection, and to (b) the most widespread isomers of test compounds searched in the PubChem database. Our comparisons led to TPR of (46-48 ± 10) % and (58 ± 10) % in those tests, respectively. With the uncertainty ranges, these quantitates were independent of the kind of spectral similarity indices/ metrics and the predictor version. Taking into account that approximately a half of actual candidates for identification were correctly predicted, the TPR values obtained should be regarded as reasonable results for predicting tandem mass spectra of isomeric chemical structures. Further development of approaches to efficient mass spectrum prediction is much required. Now, the predictor under study and some other computer programs[2,22,23] can be used for forming lists of candidate molecules (identification hypotheses[3]) in non-target analyses. With increasing dissimilarity of related structures, the probability of erroneous prediction of foreign molecules will be reduced.

Acknowledgments

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Milman, B. L.; Zhurkovich, I. K. *Trends Anal. Chem.* **2016**, 80, 636.
2. Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; Fiehn, O. *Mass. Spec. Rev.* **2018**, 37, 513.
3. Milman, B. L. *Chemical identification and its quality assurance,* Springer: Heidelberg, **2011**.
4. Wolfender, J. L.; Nuzillard, J. M.; Van Der Hooft, J. J.; Renault, J. H.; Bertrand, S. *Anal. Chem.* **2018**, *91*, 704.
5. PubChem; See https://pubchem.ncbi.nlm.nih.gov.
6. Chemical Abstracts Service; See https://www.cas.org.
7. Milman, B. L.; Zhurkovich, I. K. *Trends Anal. Chem.* **2017**, 97, 179.
8. MassBank of North America; See http://mona.fiehnlab.ucdavis.edu/downloads.
9. Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2015**, 11, 98.
10. Allen, F.; Greiner, R.; Wishart, D. *Curr. Metabolomics* **2017**, 5, 35.
11. CFM-ID. 2.0; See http://cfmid.wishartlab.com/predict.
12. Djoumbou-Feunang, Y.; Pon, A.; Karu, N.; Zheng, J.; Li, C.; Arndt, D., Gautam, M.; Allen, F.; Wishart, D. S. *Metabolites* **2019**, 9, 72.
13. CFM-ID. 3.0; See http://cfmid3.wishartlab.com/predict.
14. Milman, B. L.; Zhurkovich, I. K. *MSL* **2018**, 9, 73.
15. The National Institute of Standards and Technology. NIST Mass Spectral Search Program, version 2.3; See https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:nist17.
16. Bolton, E. (National Institutes of Health, USA). Personal communication, 2018.
17. Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994**, 5, 859.
18. Sample Size Calculator; See https://www.surveysystem.com/SSCALC.HTM.

19. Statistical Significance Calculator; See https://www.easy-calculation.com/statistics/statistical-significance.php.
20. ChemMine Tools; See http://chemminetools.ucr.edu.
21. P value from Pearson (R) Calculator; See https://www.socscistatistics.com/pvalues/pearsondistribution.aspx.
22. Andersen, J. L.; Fagerberg, R.; Flamm, C.; Kianian, R.; Merkle, D.; Stadler, P. F. *MATCH Commun. Math. Comput. Chem.* **2018**, 80, 705.
23. Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. *Nat. methods* **2019**, 16, 299.