Title

How Well Can We Predict Mass Spectra from Structures? Benchmarking Competitive Fragmentation Modeling for Metabolite Identification on Untrained Tandem Mass Spectra

Permalink

https://escholarship.org/uc/item/5cx9079h

Journal

Journal of Chemical Information and Modeling, 62(17)

ISSN

1549-9596

Authors

Bremer, Parker Ladd
Vaniya, Arpana
Kind, Tobias
et al.

Publication Date

2022-09-12

DOI

10.1021/acs.jcim.2c00936

Peer reviewed

Published in final edited form as:
J Chem Inf Model. 2022 September 12; 62(17): 4049–4056. doi:10.1021/acs.jcim.2c00936.

How Well Can We Predict Mass Spectra from Structures? Benchmarking Competitive Fragmentation Modeling for Metabolite Identification on Untrained Tandem Mass Spectra

Parker Ladd Bremer,
Department of Chemistry, University of California Davis, Davis, California 95616, United States

Arpana Vaniya,
West Coast Metabolomics Center for Compound Identification, UC Davis Genome Center, University of California Davis, Davis, California 95616, United States

Tobias Kind,
West Coast Metabolomics Center for Compound Identification, UC Davis Genome Center, University of California Davis, Davis, California 95616, United States

Shunyang Wang,
Department of Chemistry, University of California Davis, Davis, California 95616, United States

Oliver Fiehn
West Coast Metabolomics Center for Compound Identification, UC Davis Genome Center, University of California Davis, Davis, California 95616, United States

Abstract

Competitive Fragmentation Modeling for Metabolite Identification (CFM-ID) is a machine learning tool to predict in silico tandem mass spectra (MS/MS) for known or suspected metabolites for which chemical reference standards are not available. As a machine learning tool, it relies on both an underlying statistical model and an explicit training set that encompasses experimental mass spectra for specific compounds. Such mass spectra depend on specific parameters such as collision energies, instrument types, and adducts which are accumulated in libraries. Yet, ultimately prediction tools that are meant to cover wide expanses of entities must be validated on cases that were not included in the initial training and testing sets. Hence, we here benchmarked the performance of CFM-ID 4.0 to correctly predict MS/MS spectra for

Corresponding Author: **Oliver Fiehn** – *West Coast Metabolomics Center for Compound Identification, UC Davis Genome Center, University of California Davis, Davis, California 95616, United States,* ofiehn@ucdavis.edu.

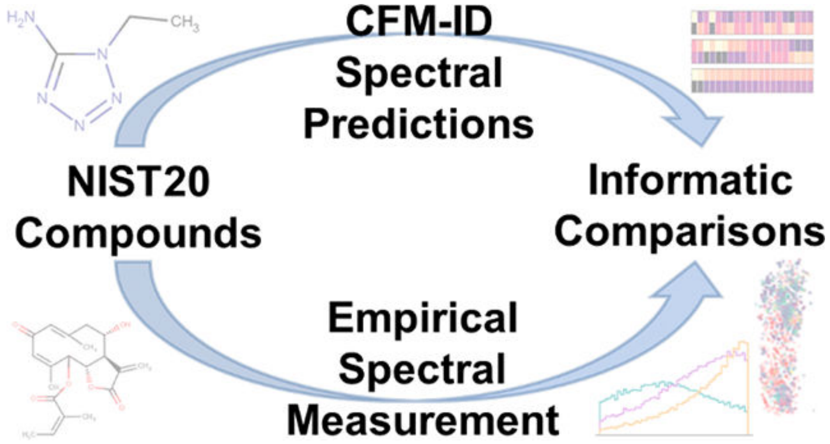
Author Contributions
P.L.B., A.V., and O.F. designed the analyses, T.K. assisted with training P.L.B. in cheminformatics, A.V. added data to MassBank.us, and S.W. helped with cheminformatics tools. P.L.B. performed analyses. P.L.B. and O.F. wrote the manuscript. All authors approved the final manuscript.

The authors declare no competing financial interest.

ASSOCIATED CONTENT
Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00936.
Figures that complement or expand on material in manuscript: complement to Figure 3b, where figure has been slightly modified because of different units on *x*-axis; complements to Figure 6, where t-SNE and PCA were used instead of UMAP; and confusion matrices and feature explanations that expand on our random forest analysis (PDF)

spectra that were not included in the CFM-ID training set and for different mass spectrometry conditions. We used 609,456 experimental tandem spectra from the NIST20 mass spectral library that were newly added to the previous NIST17 library version. We found that CFM-ID's highest energy prediction output would maximize the capacity for library generation. Matching the experimental collision energy with CFM-ID's prediction energy produced the best results, even for HCD-Orbitrap instruments. For benzenoids, better MS/MS predictions were achieved than for heterocyclic compounds. However, when exploring CFM-ID's performance on 8,305 compounds at 40 eV HCD-Orbitrap collision energy, >90% of the 20/80 split test compounds showed <700 MS/MS similarity score. Instead of a stand-alone tool, CFM-ID 4.0 might be useful to boost candidate structures in the greater context of identification workflows.

Graphical Abstract



INTRODUCTION

The expanse of metabolites observed in humans, plants, and other forms of life is enormous. The Human Metabolome Database (HMDB) alone currently contains well over 100,000 documented metabolites, and the total plant metabolome is believed to span over 1 million compounds.[1,2] In liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS)-based metabolomics, a compound in a sample is commonly annotated by comparing their experimental mass spectra to reference mass spectra that are contained in a mass spectral libraries.[3] Classically, these libraries are developed by acquiring mass spectra from authentic analytical standards. In practice, however, reference mass spectra are available for only a small fraction of the metabolome.[4,5] The coverage of compounds in PubChem that have associated mass spectra is estimated to be less than 1%.[4] Therefore, millions of compounds do not have associated experimental mass spectra, and moreover, most of them are not commercially available. Hence, mass spectra for these compounds must be predicted by in silico tools to facilitate compound identification in untargeted metabolomics.[6] Predicted reference MS/MS spectra are in untargeted metabolomics because it is estimated that more than 80% of unknown MS/MS spectra remain unidentified.

Numerous computational tools have been developed for compound identification or structure elucidation.[7] The three basic approaches are as follows: (1) rule-based fragmentation tools,[8] for which fragmentation trends are identified by either classic organic chemistry based rules such as hydrogen-rearrangement rules[9] or literature based reaction rules,[10] (2) quantum chemistry tools,[11,12] in which first principle theory is applied to simulate fragmentation of a compound of interest [Quantum chemistry tools such as quantum-chemical electron ionization mass spectra (QCEIMS) are generally applied to electron ionization spectra, but there have been recent works to predict ESI-MS spectra.[13]], and (3) machine learning tools,[14,15] for which statistical models are parametrized to generate spectra based on compound and spectrum relationships. These tools produce millions of in silico reference mass spectra relatively quickly and easily in hopes to alleviate the pressing demand for reference MS/MS spectra. The success of enhancing experimental libraries within silico libraries has been demonstrated; however, it is also clear that as stand-alone tools, they are not sufficient.[16] Other machine learning tools attempt to predict chemical structures or chemical fingerprints from spectra. Examples are CSI:FingerID, the structure classifier Canopus, or ChemDistiller.[17–19]

CFM-ID 4.0, the tool tested in this publication, is a machine learning software based on a stochastic homogeneous Markov process, with additional hard-coded fragmentation rules for certain classes of compounds such as complex lipids.[8] Therefore, it is important to highlight that in this paper we examine the underlying statistical model in conjunction with its default training set. However, CFM-ID comes with the capacity to reparametrize according to whatever example set the user might provide. CFM-ID was trained on a set of 12,165 Q-TOF fragmentation spectra for the $[M + H]^+$ adducts and 6,120 MS/MS spectra for the $[M – H]^–$ adducts, covering collision energies of 10, 20, and 40 eV.[4] Accordingly, CFM-ID predicts spectra for these collision energies for any given input compound.

The chemical space of the metabolome is more expansive than any training set. The higher accessibility of high accuracy mass spectrometers today enables the use of training sets that are representative of both orbital ion trap and Q-TOF mass spectrometers equally. We therefore tested CFM-ID's prediction capabilities for compounds, fragmentation methods, and collision energies that it has not yet encountered. To accomplish this, we predicted spectra for the highly curated and reliable NIST20 MS/MS library, which contains compounds that are not included in CFM-ID's training set that were measured on both Q-TOF and orbital ion trap instruments.

METHODS

The workflow for our methods is shown in Figure 1. We used the highly curated NIST20 library from the U.S. National Institute of Standards and Technology (NIST) as input of spectra and molecules into the benchmarking test.[20] Compounds found in NIST17[21] or the CFM-ID training set were removed from the NIST20 library set. The remaining chemical structures were used to predict MS/MS spectra using the CFM-ID 4.04 and the Mass Spectrum Rule-Based Fragmenter (MSRB) 1.1.3 software programs that were provided in Docker image format from the David Wishart laboratory (University of Alberta, Canada).[22] The software performance was evaluated by matching predictions against experimental NIST20 library MS/MS spectra using the unweighted dot product with a mass tolerance of 10 ppm and excluding all ions within 2 Da of precursors. All spectra were normalized to relative abundance before calculating mass spectral similarities. Compound structures were classified according to the Wishart laboratory ClassyFire tool using the batch version implemented at http://cfb.fiehnlab.ucdavis.edu.[23] To test our similarity-prediction model, the Vaniya/Fiehn Natural Product Library set of Q-Exactive HF orbital ion trap accurate mass MS/MS spectra (VFNPL) was freely downloaded from the Massbank of North America (https://massbank.us). For all chemical structure data sets, CACTVS molecular fingerprints were obtained using the PubChem web tool.[24] All analyses were conducted using custom python scripts.

RESULTS AND DISCUSSION

Selecting Experimental MS/MS Spectra.

The NIST20 MS/MS library is composed of 27,613 compounds that generated 1,026,712 MS/MS mass spectra. This library is commercially available to the public and is released in three-year intervals after extensive curation. Only spectra for the most often observed $[M + H]^+$ and $[M – H]^−$ adducts were used to yield a consistent and large benchmarking data set. Compared to the 2017 release (NIST17 library), there was a significant increase with 15,961 compounds and 609,456 spectra newly added. A few NIST20 molecules were already used in CFM-ID training libraries and consequently removed, leaving 15,328 and 15,494 compounds in the $[M + H]^+$ and $[M – H]^−$ benchmarking set, respectively. While CFM-ID was solely trained on Q-TOF mass spectra, we included Q-TOF as well as orbital ion trap spectra. Orbital ion trap spectra included both higher energy collisional dissociation (HCD) and collision induced dissociation (CID) fragmentations.

Creating the CFM-ID Library.

For these filtered NIST20 compounds, a CFM-ID 4.0.4 spectral library was created that was patched with CFM-ID predictions for molecules for which a rule-based upgrade model was available, MSRB 1.1.3. The MSRB-Fragmenter patch is an add-on tool that predicts spectra based on rules. The CFM-ID web tool shows users rule-based predictions when available, instead of machine-learning based predictions. Therefore, to replicate user experience, we utilized the MSRB predictions when possible.[4] In total, the MSRB-Fragmenter yielded 834 spectra for 278 compounds for $[M + H]^+$ adducts and 822 spectra for 274 compounds for $[M – H]^−$ adducts.

Overall CFM-ID Performance.

We aimed at benchmarking the performance of CFM-ID on spectra that were not included in either training, testing, or validating CFM-ID software.[25] CFM-ID version 4.0 was created in early 2020. For that reason, we utilized the NIST20 MS/MS library that was released in June 2020 and removed all compounds that were present in NIST17 or the CFM-ID 4.0 training set. For each remaining compound, we generated CFM-ID predictions for three collision-induced dissociation energies, 10, 20, and 40 eV. After removing CFM-ID training compounds, NIST17 compounds, and uncommon adducts, 248,207 spectra remained. For each spectrum, we obtained the dot product similarity score with all three energy predictions for CFM-ID. We did not include any peak within 2 Da of the precursor ion because the precursor ion signifies the intact molecule and must be considered as orthogonal to MS/MS fragment spectra and because the intensity of precursor ions varies considerably between instrument types and collision energies. For each experimental spectrum, we saved only the score with the greatest similarity among its three comparisons.

We hypothesized that that the quality of CFM-ID predictions of these spectra might depend on (a) instrument type and type of collision induced-fragmentation, (b) adduct type (a complexity which we limited by constraining to only protonated and deprotonated molecules), and (c) collision energy and, finally, the actual compound structure (defined by InChI Codes which were hashed as InChIKeys). We first partitioned 248,207 NIST20 mass spectra into six groups defined by instrument type and adduct type as given in Table 1.

When subjecting these molecules to in silico fragmentation by CFM-ID 4.04 and benchmarking these spectra against the NIST20 experimental mass spectra, we were surprised to see a clear dichotomy of matches in a histogram plot (Figure 2), with very disparate frequencies of a number of compounds that excellently matched to experimental mass spectra (at dot-score similarity > 950) and many more compounds that did not show satisfying MS/MS similarities (<50 dot-score similarity). Between these two boundaries we found a nearly flat distribution of a few other compounds. For Q-TOF spectra, the low total number of compounds may have hampered finding any good MS/MS matches at all.

Impact of Collision Energy on CFM-ID Performance.

Next, we analyzed the impact of collision energies. We first focused on the 157,407 protonated MS/MS spectra fragmented in HCD-mode using orbital ion traps and compared these to the 1,111 mass spectra in the positive ESI mode obtained by a Q-TOF mass spectrometer. In contrast to the overall analysis in Figure 2 that focused on the best MS/MS match across all experimental and in silico collision energies, here we kept all individual MS/MS dot-score similarities separate that matched each experimental spectrum against the simulated CFM-ID spectra for each of the three CFM-ID predictions. We binned all experimental collision energies into 1 eV bins, ranging from 1 to 45 eV for Q-TOF spectra and 1−70 eV for orbital ion trap mass spectra (Figure 3). For orbital ion traps, energy data differed within the NIST20 library, and we therefore selected only one specific instrument type (the Thermo Finnigan Elite Orbitrap data) to be able to utilize uniform energy descriptors. For the full range of energies calculated for this instrument type, we generated 200 bins but found a dramatic dip in the number of spectra beyond the first 50 bins (up to 70 eV) to which we therefore limited the analyses. We conclude that CFM-ID performs poorly for the Q-TOF mass spectra from the NIST20 library that were not publicly available during CFM-ID 4.0 software development. We did not find any relationship of dot-score similarities of predicted versus experimental spectra, neither with respect to the experimental energies nor when analyzed for the different simulated energies at 10–40 eV.

For the Elite Orbitrap mass spectra, we yielded a more nuanced result. While averaged MS/MS dot-score similarities remained well below the mark of 600 scores, a threshold that is often used to annotate compounds in experimental MS/MS investigations, we still saw an increase in higher-ranking dot-score similarities depending on the collision energies. For simulated low collision energies at 10–20 eV in CFM-ID (orange and purple graphs in Figure 3b), much better dot scores were achieved for experimental spectra at <10 eV or <20 eV than at >40 eV collision energies. Vice versa, CFM-ID spectra simulated for 40 eV collision energy showed the best dot-score similarities around 40 eV experimental collision energies. Based on these observations, we conclude that CFM-ID is best used for Orbitrap spectra that match in silico with experimental collision energies. However, very often experimental MS/MS spectra at 10–20 eV showed very simplistic mass spectra with very little fragmentation, which we interpret as the main reason why average dot-score similarities reached higher maxima than experimental versus predicted MS/MS spectra at 40 eV. In practice, low energy MS/MS spectra only yield uninformative neutral losses such as water or ammonium losses. Hence, for the purpose of annotating unknown compounds with in silico libraries, experimental and in silico spectra at 40 eV should be more useful.

Orbital ion trap collision energies are often given in relative normalized collision energies (%NCE). To refer %NCE values to energies given in eV, we used information from metadata given in the NIST20 library for collision energies for the Thermo Finnigan Elite Orbitrap instrument containing both eV and %NCE information. Applied Orbitrap energies are represented as proportions of an optimal energy that scales (linearly) with the precursor mass. This proportion is typically written as "%NCE".

$$(\text{Applied eV}) = (\text{Optimal eV}) \times (\%\text{NCE})$$

and

$$(\text{Optimal eV}) \propto (\text{Precursor mass})$$

therefore

$$(\text{Applied eV}) \propto (\text{Precursor mass}) \times (\%\text{NCE})$$

The applied eV was used as the *x*-axis in Figure 3b. Hence, histograms give very similar results if eV values are known of if they are displayed as $\text{Precursor mass} \times \%\text{NCE}$ (Supplement S1).

For other instrument types, such as the Thermo Fisher Lumos instrument, a different constant C in the proportionality would be needed. For this reason, we did not include all Orbitrap NIST20 spectra but only spectra from this specific instrument type. Overall, it is clear that one cannot simply use %NCE values that are typically reported for orbital ion trap instruments and report definitive eV values across all instrument types.

We wondered why most spectra predictions gave either excellent results at >900 similarity or dismal results at <100 similarity. We used the best scoring CFM-ID energy for each molecule and analyzed the percentage of all 86,747 molecules for $[M + H]^+$ adducts for the Thermo Finnigan Elite orbital ion trap mass spectrometer that yielded acceptable dot-score similarities between CFM-ID predictions and HCD-experimental MS/MS spectra (Figure 4). In this analysis, it becomes clear that very good predictions were found for a comparatively large population of very low experimental collision energies, while very poor MS/MS predictions consisted of a comparatively large population of very high experimental collision energies. The “best predictions” (>950) were bolstered by experimental collision energies close to 1 eV. Hence, the vast majority of the “best predicted spectra” resulted from a systematic bias of matching very simple MS/MS fragmentation spectra with simple predictions.

Impact of a Molecule Structure on CFM-ID Performance.

Next, we investigated the impact of a structure on CFM-ID predictability. To remove observed systematic bias from mismatched energies, we limited the analyses of MS/MS spectra to the 8,035 molecules that were assigned with explicit eV units in the NIST20 library between 35 and 45 eV for the Thermo Finnigan Orbitrap. Figure 5 shows that for >90% of these compounds, MS/MS similarity dot scores of <700 were yielded, even when choosing the optimal 40 eV setting in CFM-ID predictions for HCD-Orbitrap spectra. Yet, for about 10% of these molecules, decent MS/MS spectra could be simulated with dot scores > 600 and in some cases even >800 dot-score similarities.

We therefore used this subset of data to explore the impact of chemical structure on CFM-ID predictability of MS/MS spectra. We first hypothesized that compounds with a greater similarity to the CFM-ID training set might yield better dot-score MS/MS similarities. To this end, we acquired CACTVS fingerprints using the PubChem REST API for 4,040 molecules of the training set (that was disclosed by the authors of the CFM-ID software) and applied these to 8,298 chemical fingerprints for the 35–45 eV HCD spectra molecules for $[M + H]^+$ adducts in the Orbitrap NIST20 database.[26] With all chemical fingerprints combined, we created a 2-dimensional reduction embedding of fingerprints using Uniform Manifold Approximation and Projection (UMAP), Figure 6.[27] We also examined dimensionality reduction using PCA and t-SNE. Pairwise comparison of PCA’s dimensions as well as t-SNE projections yielded the same clustering of well-performing compounds (Supplements S2, S3). Chemical fingerprints of molecules with low dot-score MS/MS similarities were expected to be found far away from the training data. We found that compounds with very poor MS/MS dot scores (dark blue) showed UMAP structural overlaps to the same degree as compounds with good dot scores. Hence, chemical similarity to the training data itself did not predict the ability to correctly simulate MS/MS spectra in CFM-ID. Instead, we found clusters of good predictions (yellow dots), suggesting a success of CFM-ID for very specific chemical classes but not for others. To this end, we classified all 8,298 molecules by the ClassyFire algorithm into chemical SuperClasses and analyzed the proportion of dot-score similarities for the top-6 SuperClasses (Figure 7). It became clear that well-predicted compounds in CFM-ID at >900 dot-score similarities were very likely to be benzenoids, while the poorly predicted compounds at <600 dot scores were likely to be organoheterocyclics. The overall proportion of chemical compounds was heavily biased toward these two SuperClasses, precluding definitive comments about other chemical structures.

Intrigued by the notion that specific compound types were well-predicted and specific compound types were poorly predicted, we sought to achieve a higher-resolution view on chemical substructures. Here, we used a random forest approach to identify fingerprint bits with the capability to distinguish between well-predicted and poorly predicted compounds and then later, in an attempt to predict CFM-ID's capability to predict spectra, used a binary classification scheme with a dot-score similarity of 700 as a watershed mark between good and poorly predictable substructures.

This simplistic binary scheme was performed to allow the RF model to learn specific chemical features that had a high impact on overall good CFM-ID scoring, instead of using regression models that might focus on differentiating among the more sampled, lower MS/MS similarity dot scores. We chose the model that maximized precision because precision is most important for building libraries of predicted MS/MS spectra. To identify features, we selected the top-50 chemical fingerprint bits that showed the greatest capacity to distinguish between good and worse MS/MS predictions. We examined the distributions for compounds for each chemical fingerprint bit in heatmaps and give results for the top-substructure fingerprints in Table 2, Table 3, and Supplement S2. Using the chemical fingerprint bit 185 ("two rings of membership 6") and bit 143 ("at least 1 ring of size 5") explicitly reproduced the result of the superclass analysis. Hence, both the fingerprint analysis and the ClassyFire SuperClass analysis showed that CFM-ID maintained the trained ability to predict MS/MS spectra for simple aromatic molecules that consisted of carbon-only rings. However, this training did not extend to other cyclic structures such as small ring systems with heteroatoms for which CFM-ID predictions failed. Using a train/test split as 20%/80%, chosen randomly from the NIST20 data set, we found that more than 90% of the structures yielded <700 dot-score similarities to the corresponding experimental spectra (see the confusion matrix, Supplement S3). Yet, 20 of the 23 benzenoids included in this withheld testing set gave >700 dot score similarities of confidence that the model can be used to select subsets of proposed compounds for which one can generate an in silico library.

To confirm how generalized this model is, we sought an orthogonal test set for which we used the Vaniya-Fiehn Natural Product Library within the public MassBank.us repository. Because our collision energy analysis for CFM-ID strongly suggested that matching the %NCE for Orbital Ion Trap instrument was extremely important, we removed all compounds for which we could not obtain or calculate an equivalent %NCE to match the CFM-ID "40 eV collision energy". This constraint left 226 compounds to be tested using the CFM-ID 40 eV prediction. When removing all ions within 2 Da of the precursor ion, only 6 of the 226 tested natural product compounds yielded a >700 dot score (Supplement S3), confirming that CFM-ID has very limited prediction ability for correct MS/MS spectra beyond simple benzenoid structures.

CONCLUSIONS

It is important that machine learning-based prediction models are tested and benchmarked by independent analyses on data sets that were not available during model building. Here, we tested mass spectra from NIST20 and MassBank.us (MassBank of North America) to probe the accuracy for which CFM-ID 4.0 was able to predict spectra from the structure, a holy grail in tools for use in untargeted metabolomics or exposome research. As a standalone too, CFM-ID's performance provides only a few spectra with high MS/MS similarity scores when validated against experimental spectra. However, even with low dot-score similarities, tools like CFM-ID might be worthwhile to be used in the context of compound identification workflows to boost some structures over alternative chemicals, as has been shown in the CASMI 2016 contest.[16] For example, CFM-ID could be used to predict fragmentation at 40 eV at which richer fragmentations occur that are useful for compound identification. For HCD spectra in orbital ion trap mass spectrometers, we observed some structural clusters of good MS/MS predictability. While it is not possible to match CFM-ID to a specific %NCE, CFM-ID collision energies in eV are proportional to the product of %NCE and precursor mass of the compound. Based on these results, it seems reasonable that for improvement of MS/MS in silico prediction from structures, Q-TOF and HCD experimental spectra may be combined to expand the space of training sets. During our benchmarking tests, we found that the accuracy of CFM-ID 4.0 predictions depended on specific chemical substructures but not on the similarity of tested structures to the structural space in the training set. Hence, we can conclude that currently, machine learning for direct MS/MS predictions in CFM-ID did not work for most compound classes, except for the ClassyFire SuperClass of benzenoids. Nevertheless, if CFM-ID 4.0 is cautiously used in conjunction with compound-identification workflows, it may improve overall compound ID scores.[16,28] We hope that in the coming years the standardization of metabolomics repositories will enable massive data sets to drive the progress of machine learning methods to predict mass spectra from chemical structures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank the David Wishart research group (University of Alberta) for early access to the CFM-ID 4.0 program.

Funding

Funding was provided by the National Institutes of Health by NIH U2C ES030158.

DATA AND SOFTWARE AVAILABILITY

The code used in this manuscript is available at https://github.com/plbremer/cfmid_2. The CFM-ID docker images are available at https://hub.docker.com/repository/docker/wishartlab/cfmid. The NIST20 and NIST17 data sets are available for purchase at https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries. The VFNPL is freely available at https://massbank.us/.

ABBREVIATIONS

MS/MS	tandem mass spectrometry
NIST	National Institute of Standards and Technology
HMDB	Human Metabolome Database

LC	liquid chromatography
QTOF	quadrupole time of flight
MSRB	Mass Spectrum Rule-Based Fragmenter
VFNPL	Vaniya/Fiehn Natural Product Library
HCD	higher-energy collisional dissociation
CID	collision-induced dissociation
NCE	normalized collision energy
UMAP	Uniform Manifold Approximation/Projection

REFERENCES

(1). Wishart DS; Feunang YD; Marcu A; Guo AC; Liang K; Vázquez-Fresno R; Sajed T; Johnson D; Li C; Karu N; Sayeeda Z; Lo E; Assempour N; Berjanskii M; Singhal S; Arndt D; Liang Y; Badran H; Grant J; Serra-Cayuela A; Liu Y; Mandal R; Neveu V; Pon A; Knox C; Wilson M; Manach C; Scalbert A HMDB 4.0: The Human Metabolome Database for 2018. Nucleic Acids Res. 2018, 46 (D1), D608–D617. [PubMed: 29140435]

(2). Rai A; Saito K; Yamazaki M Integrated Omics Analysis of Specialized Metabolism in Medicinal Plants. Plant J. Cell Mol. Biol. 2017, 90 (4), 764–787.

(3). Cajka T; Fiehn O Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. Anal. Chem. 2016, 88, 524. [PubMed: 26637011]

(4). Djoumbou-Feunang Y; Pon A; Karu N; Zheng J; Li C; Arndt D; Gautam M; Allen F; Wishart DS CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. Metabolites 2019, 9 (4), 72. [PubMed: 31013937]

(5). Go Y-M; Walker DI; Liang Y; Uppal K; Soltow QA; Tran V; Strobel F; Quyyumi AA; Ziegler TR; Pennell KD; Miller GW; Jones DP Reference Standardization for Mass Spectrometry and High-Resolution Metabolomics Applications to Exposome Research. Toxicol. Sci. 2015, 148 (2), 531–543. [PubMed: 26358001]

(6). Schrimpe-Rutledge AC; Codreanu SG; Sherrod SD; McLean JA Untargeted Metabolomics Strategies – Challenges and Emerging Directions. J. Am. Soc. Mass Spectrom. 2016, 27 (12), 1897–1905. [PubMed: 27624161]

(7). Krettler CA; Thallinger GG A Map of Mass Spectrometry-Based in Silico Fragmentation Prediction and Compound Identification in Metabolomics. Brief. Bioinform. 2021, 22 (6), bbab073. [PubMed: 33758925]

(8). Allen F; Greiner R; Wishart D Competitive Fragmentation Modeling of ESI-MS/MS Spectra for Putative Metabolite Identification. Metabolomics 2015, 11 (1), 98–110.

(9). Tsugawa H; Kind T; Nakabayashi R; Yukihira D; Tanaka W; Cajka T; Saito K; Fiehn O; Arita M Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. Anal. Chem. 2016, 88 (16), 7946–7958. [PubMed: 27419259]

(10). Thermo Fisher Scientific. Powering Confident Insights - Explore Your Small-Molecule Data to Its Core; 12pp.

(11). Ásgeirsson V; Bauer CA; Grimme S Quantum Chemical Calculation of Electron Ionization Mass Spectra for General Organic and Inorganic Molecules. Chem. Sci. 2017, 8 (7), 4879–4895. [PubMed: 28959412]

(12). Wang S; Kind T; Tantillo DJ; Fiehn O Predicting in Silico Electron Ionization Mass Spectra Using Quantum Chemistry. J. Cheminformatics 2020, 12 (1), 63.

(13). Borges RM; Colby SM; Das S; Edison AS; Fiehn O; Kind T; Lee J; Merrill AT; Merz KM; Metz TO; Nunez JR; Tantillo DJ; Wang L-P; Wang S; Renslow RS Quantum Chemistry Calculations for Metabolomics. Chem. Rev. 2021, 121 (10), 5633. [PubMed: 33979149]

(14). Wei JN; Belanger D; Adams RP; Sculley D Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. ACS Cent. Sci. 2019, 5 (4), 700–708. [PubMed: 31041390]

(15). Liebal UW; Phan ANT; Sudhakar M; Raman K; Blank LM Machine Learning Applications for Mass Spectrometry-Based Metabolomics. Metabolites 2020, 10 (6), 243. [PubMed: 32545768]

(16). Blaženović I; Kind T; Torbašinović H; Obrenović S; Mehta SS; Tsugawa H; Wermuth T; Schauer N; Jahn M; Biedendieck R; Jahn D; Fiehn O Comprehensive Comparison of in Silico MS/MS Fragmentation Tools of the CASMI Contest: Database Boosting Is Needed to Achieve 93% Accuracy. J. Cheminformatics 2017, 9 (1), 32.

(17). Dührkop K; Shen H; Meusel M; Böcker S Searching molecular structure databases with tandem mass spectra using CSI:FingerID. PNAS 2015, 112, 12580. [PubMed: 26392543]

(18). Dührkop K; Nothias L-F; Fleischauer M; Reher R; Ludwig M; Hoffmann MA; Petras D; Gerwick WH; Rousu J; Dorrestein PC; Böcker S Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra. Nat. Biotechnol. 2021, 39 (4), 462–471. [PubMed: 33230292]

(19). Laponogov I; Sadawi N; Galea D; Mirnezami R; Veselkov KA ChemDistiller: An Engine for Metabolite Annotation in Mass Spectrometry. Bioinformatics 2018, 34 (12), 2096–2102. [PubMed: 29447341]

(20). NIST 20 MS/MS Library (2020); https://www.sisweb.com/software/nist-msms.htm#2 (accessed 2021-03-04).

(21). Stein SE NIST 17 MS/MS LIbrary; 2017; DOI: 10.18434/T4H594.

(22). Wang F; Liigand J; Tian S; Arndt D; Greiner R; Wishart DS CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. Anal. Chem. 2021, 93 (34), 11692–11700. [PubMed: 34403256]

(23). Djoumbou Feunang Y; Eisner R; Knox C; Chepelev L; Hastings J; Owen G; Fahy E; Steinbeck C; Subramanian S; Bolton E; Greiner R; Wishart DS ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. J. Cheminformatics 2016, 8 (1), 61.

(24). PubChem/CACTVS Fingerprints. https://pubchemdocs.ncbi.nlm.nih.gov/data-specification (accessed 2021-09-08).

(25). Chao A; Al-Ghoul H; McEachran AD; Balabin I; Transue T; Cathey T; Grossman JN; Singh RR; Ulrich EM; Williams AJ; Sobus JR In Silico MS/MS Spectra for Identifying Unknowns: A Critical Examination Using CFM-ID Algorithms and ENTACT Mixture Samples. Anal. Bioanal. Chem. 2020, 412 (6), 1303–1315. [PubMed: 31965249]

(26). Ihlenfeldt WD; Takahashi Y; Abe H; Sasaki S Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. J. Chem. Inf. Comput. Sci. 1994, 34 (1), 109–116.

(27). McInnes L; Healy J; Melville J UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2020, ArXiv180203426. ArXiv Preprint. Cs Stat. https://arxiv.org/abs/1802.03426#:~:text=UMAP%20(Uniform%20Manifold%20Approximation%20and,applies%20to%20real%20world%20data (accessed 2022-08-29).

(28). Schymanski EL; Ruttkies C; Krauss M; Brouard C; Kind T; Dührkop K; Allen F; Vaniya A; Verdegem D; Böcker S; Rousu J; Shen H; Tsugawa H; Sajed T; Fiehn O; Ghesquière B; Neumann S Critical Assessment of Small Molecule Identification 2016: Automated Methods. J. Cheminformatics 2017, 9 (1), 22.

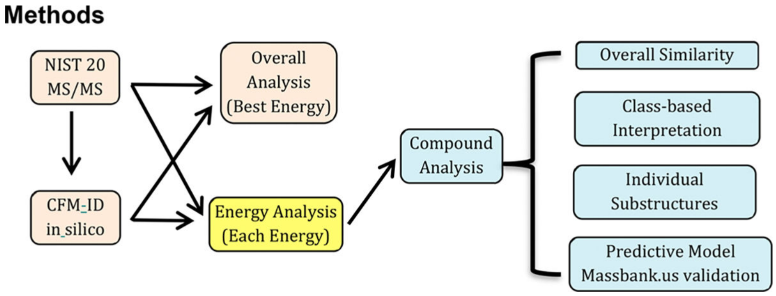

Methods
NIST 20 MS/MS
CFM-ID in silico
Overall Analysis (Best Energy)
Energy Analysis (Each Energy)
Compound Analysis
Overall Similarity
Class-based Interpretation
Individual Substructures
Predictive Model Massbank.us validation

Figure 1.

Overall method workflow.

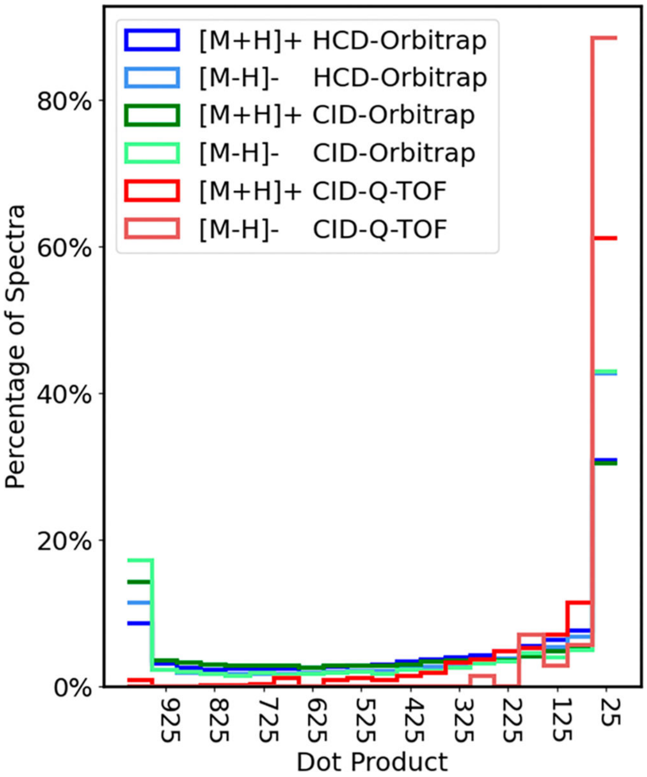


Figure 2.

Overall CFM-ID performance measured by dot products between experimental NIST20 MS/MS spectra and CFM-ID predictions for the same compound and adduct. The dot product was taken between experimental spectra and the three CFM-ID predictions, regardless of the fragmentation method or settings. The best scoring dot product among the three comparisons was recorded, and the total list was partitioned into six groups according to fragmentation conditions and adduct.

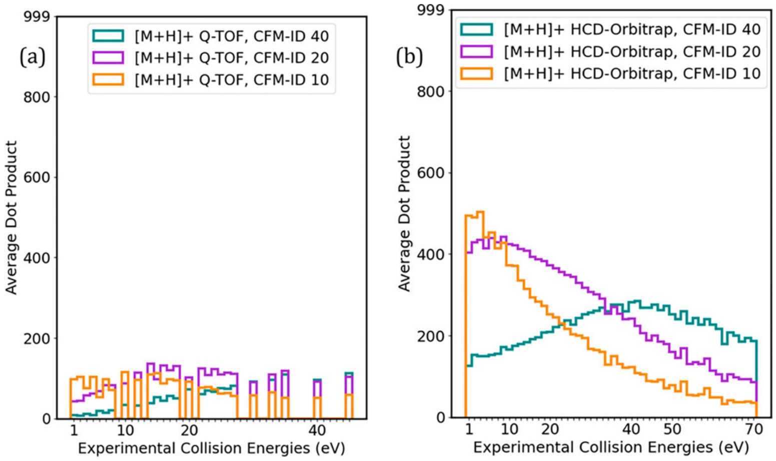


Figure 3.

Histograms of dot-score similarities for [M + H]+ molecules between experimental versus predicted MS/MS spectra, by experimental collision energies. *Left (a):* 1,111 experimental Q-TOF spectra from the NIST20 library. *Right (b)*: 86,747 Thermo Finnigan Elite Orbital Ion Trap spectra from the NIST20 library.

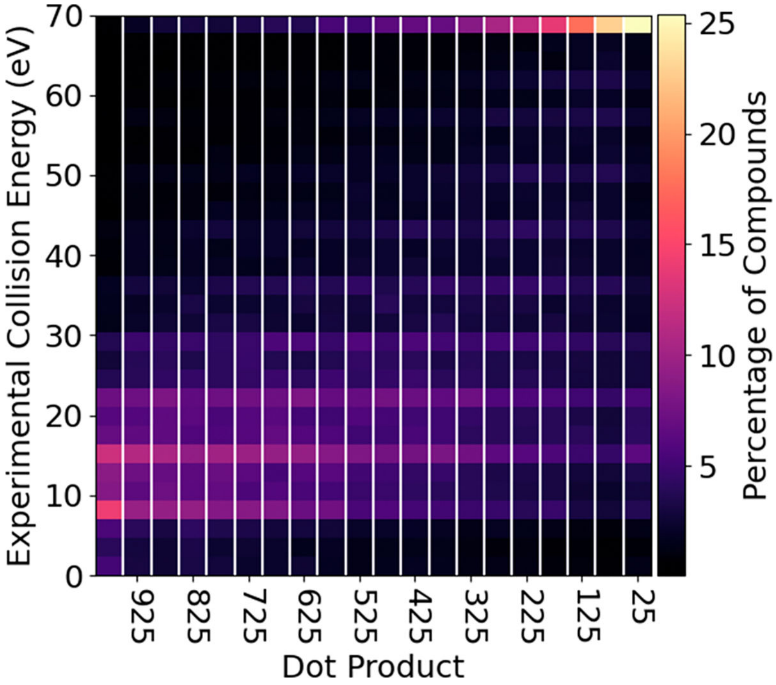


Figure 4.

Histogram of [M + H]$^+$/HCD-Orbitrap collision energy against CFM-ID predictions. Each normalized to the sum of spectra in that bin of dot product experimental column was scores.

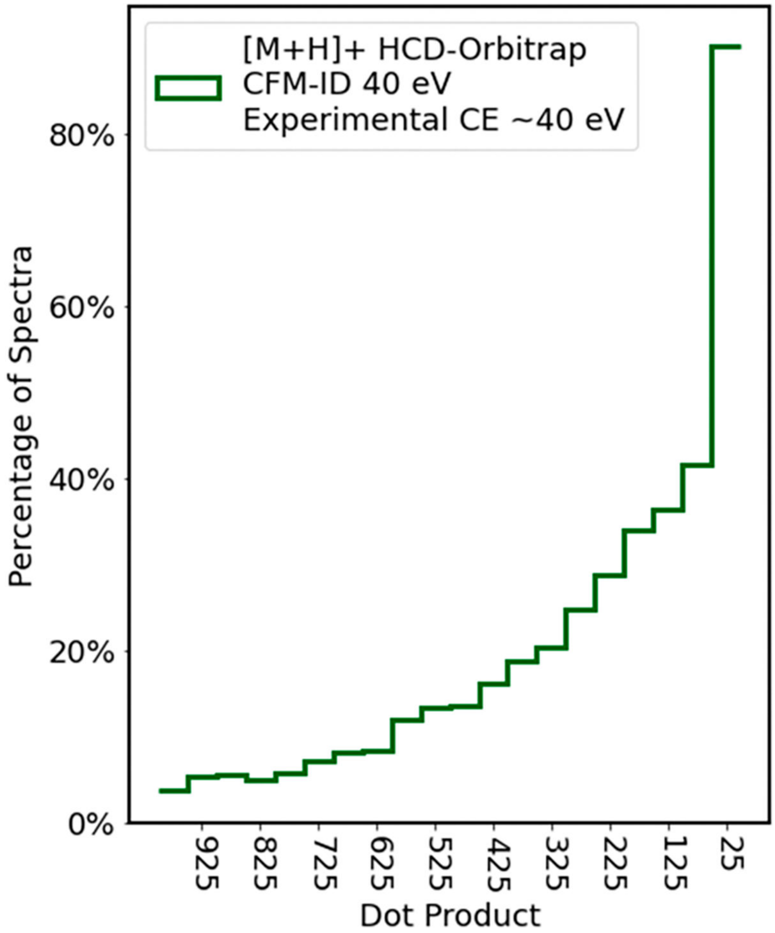


Figure 5.
Histogram of 8,035 [M + H]$^+$/HCD-Orbitrap compounds with experimental collision energies of 35–45 eV and simulated CFM-ID energy of 40 eV.

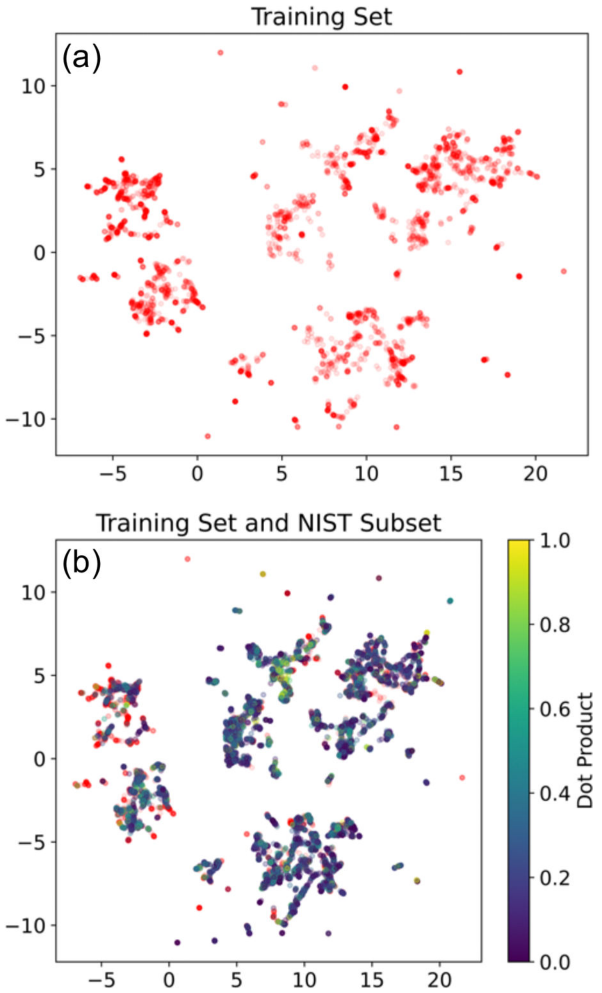


Figure 6.

2D UMAP embedding of CFM-ID positive training fingerprints and $[M + H]^+$/HCD-Orbitrap fingerprints. Upper panel (a) training data set. Lower panel (b) 8,298 molecules with 35–45 eV $[M + H]^+$ MS/MS spectra superimposed onto the training data (red dots). The yellow/blue color scheme indicates the normalized dot product values 0–1000 between 0 and 1.

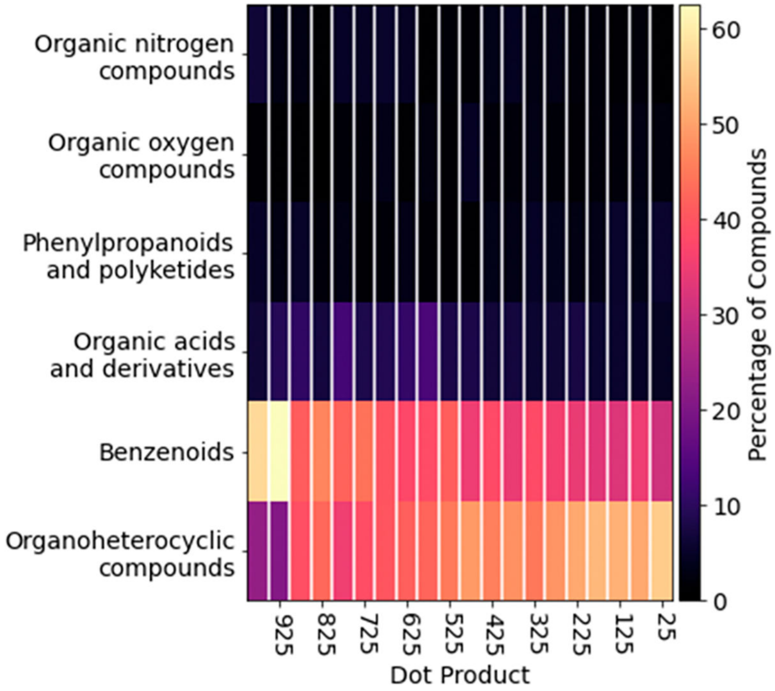


Figure 7.

ClassyFire-defined chemical superclasses vs the MS/MS dot product similarity for HCD-Orbitrap spectra $[M + H]^+$ between 35 and 45 eV. Each binned column of the dot product is sum-normalized.

Table 1.

MS/MS Spectra from the NIST20 Library Used to Benchmark CFM-ID Software

adduct and type of fragmentation	number of tested spectra
$[M + H]^+$, Orbitrap HCD	157,407
$[M - H]^-$, Orbitrap HCD	71,026
$[M + H]^+$, Orbitrap CID	12,295
$[M - H]^-$, Orbitrap CID	6,333
$[M + H]^+$, Q-TOF MS/MS	1,111
$[M - H]^-$, Q-TOF MS/MS	35

Table 2.

Substructures Associated with >700 Dot-Score Similarities by CFM-ID

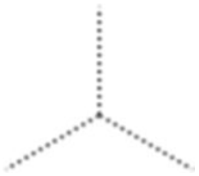
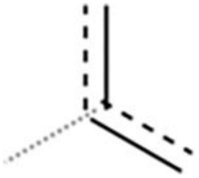
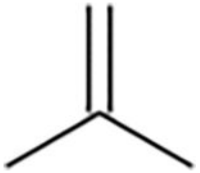
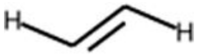
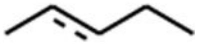
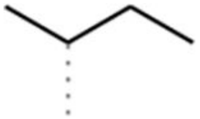
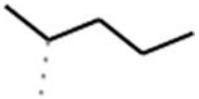
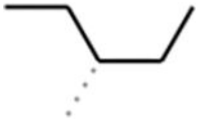
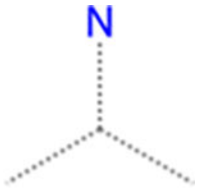
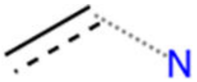
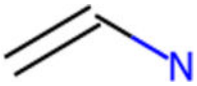
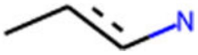
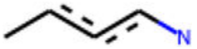
Bit Number	SMILES/SMARTS	Visualization
185	At least 2 rings of size 6	N/A
333	C(~C)(~C)(~C)	
345	C(~C)(~H)(~N)	N/A
356	C(~C)(:C)(:C)	
365	C(~H)(~N)	N/A
430	C(-C)(-C)(=C)	
516	[#1]-C=C-[#1]	
688	C-C:C-C-C	
708	C-C(C)-C-C	
709	C-C(C)-C-C-C	
710	C-C-C(C)-C-C	

Table 3.

Substructures Associated with <700 Dot-Score Similarities by CFM-ID

Bit Number	SMILES/SMARTS	Visualization
19	>=2 O	N/A
143	At least 1 ring of size 5	N/A
340	C(~C)(~C)(~N)	
374	C(~H)(~H)(~H)	N/A
376	C(~N)(:C)	
449	C(-N)(=C)	
545	N-C:C-C	
600	N-C:C:C-C	
665	N-C:C-C-C