ESA's Cometary Mission Rosetta-Re-Characterization of the COSAC Mass Spectrometry Results

Guillaume Leseigneur, Jan Hendrik Bredehöft, Thomas Gautier, Chaitanya Giri, Harald Krüger, Alexandra J Macdermott, Uwe J Meierhenrich, Guillermo M Muñoz Caro, François Raulin, Andrew Steele, et al.

A Journal of the Gesellschaft Deutscher Chemiker

Angewandte International Edition GDCh Chemie

www.angewandte.org

Accepted Article

Title: ESA's Cometary Mission Rosetta – Re-Characterization of the COSAC Mass Spectrometry Results

Authors: Guillaume Leseigneur, Jan Hendrik Bredehöft, Thomas Gautier, Chaitanya Giri, Harald Krüger, Alexandra J. MacDermott, Uwe J. Meierhenrich, Guillermo M. Muñoz Caro, François Raulin, Andrew Steele, Harald Steininger, Cyril Szopa, Wolfram Thiemann, Stephan Ulamec, and Fred Goesmann

To be cited as: *Angew. Chem. Int. Ed.* **2022**, e202201925

Link to VoR: https://doi.org/10.1002/anie.202201925

WILEY-VCH

10.1002/anie.202201925

RESEARCH ARTICLE

WILEY-VCH

ESA's Cometary Mission Rosetta – Re-Characterization of the COSAC Mass Spectrometry Results

Guillaume Leseigneur,[a] Jan Hendrik Bredehöft,[b] Thomas Gautier,[c,d] Chaitanya Giri,[e,f] Harald Krüger,[g] Alexandra J. MacDermott,[h] Uwe J. Meierhenrich,*[a] Guillermo M. Muñoz Caro,[i] François Raulin,[j] Andrew Steele,[k] Harald Steininger,[l] Cyril Szopa,[c] Wolfram Thiemann,[m] Stephan Ulamec[n] and Fred Goesmann[g]

In memory of Helmut Rosenbauer († 05 May 2016) who designed and developed the scientific program of Rosetta's lander Philae

[a] G. Leseigneur, Prof. Dr. U. J. Meierhenrich
Université Côte d'Azur, CNRS UMR 7272
Institut de Chimie de Nice
28 Avenue Valrose, 06108 Nice, France
E-mail: uwe.meierhenrich@univ-cotedazur.fr
[b] PD Dr. J.H. Bredehöft
University of Bremen, Department 02 Biology/Chemistry
Institute for Applied and Physical Chemistry
Leobener Str.5, 28359 Bremen, Germany
[c] Dr. T. Gautier, Prof. Dr. C. Szopa
Laboratoire Atmosphère, Milieux, Observations Spatiales (LATMOS)
LATMOS/IPSL, UVSQ Université Paris-Saclay, Sorbonne Université, CNRS
11 Bd d'Alembert, 78280 Guyancourt, France
[d] Dr. T. Gautier
LESIA, Observatoire de Paris,
Université PSL, CNRS, Sorbonne Université, Université de Paris
5 place Jules Janssen, 92195 Meudon, France
[e] Dr. C. Giri
Research and Information System for Developing Countries, India Habitat Centre
Lodhi Road, New Delhi-110 003, India
[f] Dr. C. Giri
Earth-Life Science Institute, Tokyo Institute of Technology
2-12-1-IE-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
[g] PD Dr. H. Krüger, Dr. F. Goesmann
Max Planck Institute for Solar System Research
Justus von Liebig Weg 3, 37077 Göttingen, Germany
[h] Prof. Dr. A. J. MacDermott
University of Houston-Clear Lake
2700 Bay Area Boulevard, Houston, TX 77058, USA
[i] Dr. G. M. Muñoz Caro
Centro de Astrobiología (CSIC-INTA)
Ctra. de Ajalvir, km 4, Torrejón de Ardoz, 28850 Madrid, Spain
[j] Prof. em. Dr. F. Raulin
Univ Paris Est Créteil and Université de Paris, CNRS
LISA, F-94010 Créteil, France
[k] Dr. A. Steele
Geophysical Laboratory, Carnegie
Institution of Washington, Washington, DC, USA
[l] Dr. H. Steininger
Design Assurance Department
OHB System AG
Universitätsallee 27, 28359 Bremen, Germany
[m] Prof. em. Dr. W. Thiemann
University of Bremen, Institute for Applied and Physical Chemistry
Leobener Strasse NW2, 28359 Bremen, Germany
[n] Dr. S. Ulamec
German Aerospace Center (DLR), Space Operations and Astronaut Training
Linder Höhe, 51147, Cologne, Germany

Supporting information for this article is given via a link at the end of the document.

Abstract: The most pristine material of the Solar System is assumed to be preserved in comets in the form of dust and ice as refractory matter. ESA's mission Rosetta and its lander Philae had been developed to investigate the nucleus of comet 67P/Churyumov-Gerasimenko *in situ*. Twenty-five minutes after the initial touchdown of Philae on the surface of comet 67P in November 2014, a mass spectrum was recorded by the time-of-flight mass spectrometer COSAC onboard Philae. The new characterization of this mass spectrum through non-negative least squares fitting and Monte Carlo simulations reveals the chemical composition of comet 67P. A suite of 12 organic molecules, 9 of which also found in the original analysis of this data, exhibit high statistical probability to be present in the grains sampled from the cometary nucleus. These volatile molecules are among the most abundant in the comet's chemical composition and represent an inventory of the first raw materials present in the early Solar System.

Introduction

The Philae lander, part of the ESA Rosetta space mission, made a non-nominal landing on comet 67P/Churyumov-Gerasimenko on November 12, 2014. The lander bounced several times on the surface of the comet before coming to rest in an unfortunately shadowed spot, where its solar arrays could not provide sufficient energy to recharge the onboard batteries.[1]

However, the first and most energetic impact excavated about 0.4 m^3 of surface material,[2] and led Philae to bounce hundreds of meters above the comet surface for about two hours in the low gravity environment of comet 67P, due to a malfunction of the anchoring harpoons.[1] As a result of the impact, nucleus material had been deposited in the exhaust port of the COmetary SAmpling and Composition (COSAC) instrument on board of Philae.[3] COSAC was a gas chromatograph coupled to a mass spectrometer that could also be used independently in what has been referred to as "sniffing" mode, where the mass spectrometer ionizes and detects molecules that passively entered the chamber.[4] Twenty-five minutes after the first touchdown, one such mass spectrum was obtained by COSAC that showed much higher peak diversity and intensity than the blanks taken beforehand (or spectra taken several hours later on the surface of the comet). The temperature in the exhaust pipes was 12° to 15°C, which would have allowed volatile molecules to sublimate and get detected by the mass spectrometer in a measurement that took 2 minutes and 20 seconds. After that, several other sniffing mass spectra were taken and showed a fast decrease in overall intensity,[5] proving that these represented excavated material.

The first mass spectrum (MS) was analyzed by the COSAC science team (CT), with a convincing fit that, however, fails to cover a large part of the MS signal observed for mass/charge ratio (*m/z*) 15 and a fraction of the *m/z* 29 peak.[6] Goesmann *et al.* (2015) clearly stated that a single MS of mixed compounds is inherently degenerate. This is true and even more so in this particular MS, as the low count number adds complexity and degeneracy due to low signal to noise ratio (SNR). Furthermore, this is a unique *in situ* measurement from a cometary nucleus, likely not to be repeated in the next decade, so we do not have a precise idea on which molecules to exclude from the initial pool, and we cannot be sure of how many molecules have significant contributions to this spectrum. In addition, the very limited amount of data makes noise characterization challenging.

Finally, the non-nominal sampling of cometary material may induce an unknown instrumental transfer function. Therefore, definite mixing ratios of molecules cannot be given as a certainty, and we aim to give a broader view of the possible molecules found in this MS, and assess confidence in specific identifications. It is important to note that, as in the original analysis, the mixing ratios provided correspond to those in the ion source only. To extrapolate to mixing ratios of the cometary material we would need to consider the transport mass-dependent fractionation from the moment the grain sublimates to the moment the gas arrives in the ionization chamber. This information is complicated to model and currently unavailable.

The final fit from the CT used 16 molecules (Table 1), ranging in mass from 16 u (methane) to 62 u (ethylene glycol).[6] We started our analysis from the already binned and background-subtracted spectrum from the CT, shown here in Figure 1. The intensity of the *m/z 18* peak attributed to water is normalized to 100, which represents 2366 detector counts. All other peaks are then represented as a percentage of this count number.

Already at first glance, the MS shows a very interesting peak distribution (see Figure 1). Indeed, peaks observed at *m/z* 56–61 are of almost equal intensity to more common peaks such as *m/z* 26–32 and 42–46, with still a sharp cut off at *m/z* 62 after which no significant signal is found for higher masses. The peak observed at *m/z* 15 (likely related to a NH or CH_3 fragment) is unexpectedly high and is discussed extensively in the Supporting Information (SI), subsection "Molecular contributors to *m/z* 15". These observations led the CT to consider molecules of mass no higher than 62 u. For the present analysis, we have taken the original list of about 110 candidate molecules and added a few hand-picked ones (SI, "Higher mass molecules"), up to 86 u, that have significant intensities mainly on peaks below *m/z* = 62 to try to explain this peculiar higher mass peak distribution, especially for *m/z* 57 and 59. A list of the 120 molecules total that have been considered is shown in Table S1.

The approach we use is the same as that utilized by the CT: find the best fit to this spectrum by a superposition of standard National Institute of Standards and Technology (NIST) mass spectra of candidate cometary molecules. The differences lie in the method: the CT started with all molecules with a mass up to 62 u with a reference mass spectrum in the NIST database (Table S1 of Goesmann *et al.*) and manually trimmed down this list by removing from consideration molecules with a strong peak at an *m/z* where the taken MS shows low, zero or even negative intensity (Table S2 of Goesmann *et al.*). The final list was obtained manually through trial and error, with the aim of finding a chemically consistent set of organics that was realistic to be found in the environment of a cometary nucleus. Finally, to further constrain this list, standard boiling points were used as a proxy for volatility (Table S3 of Goesmann *et al.*). However, in the extremely low pressures of the cometary environment these are not trivial to evaluate.

In the present work, we used non-negative least squares fitting[7] coupled with a Monte Carlo iteration method,[8] starting from the same raw data excel file as the previous study. We also consider recent developments in cometary and interstellar chemistry in order to further constrain the original pool of molecules to feed our algorithm. After validating our model and its stability with respect to the different input parameters (see SI), a major part of the work was to adjust the initial pool of potential target molecules and compare the output list of these molecules and their abundances.

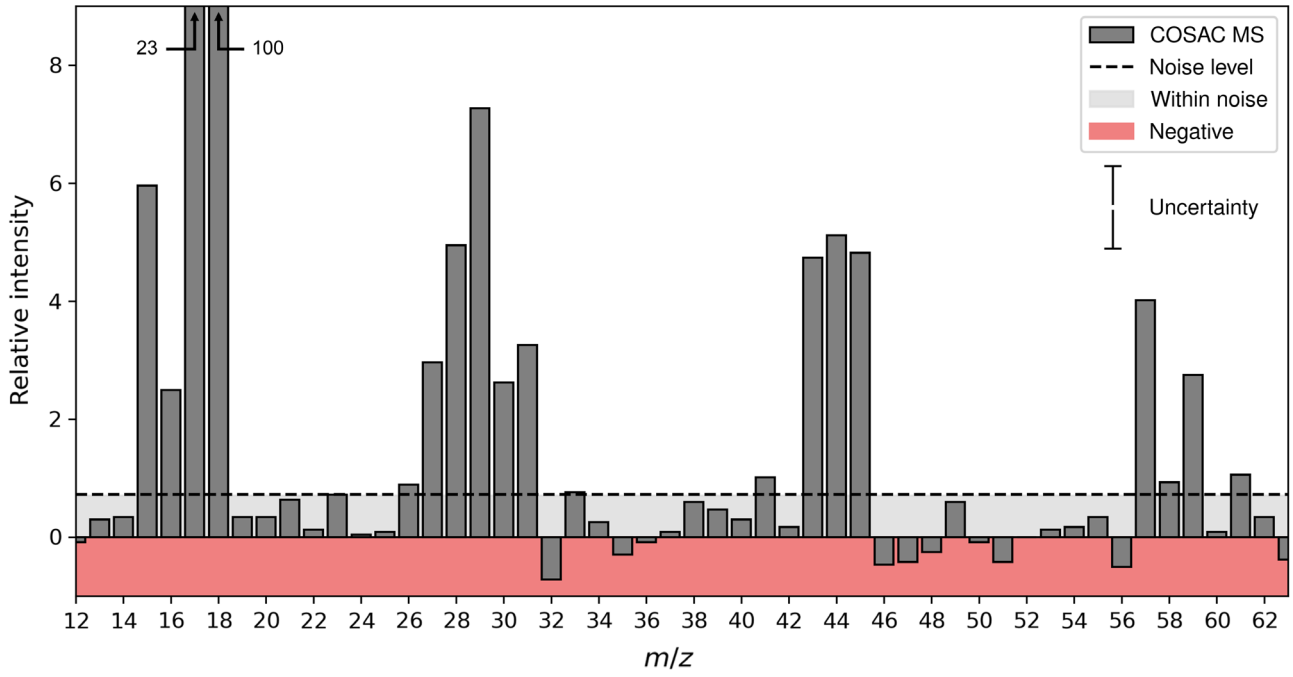


Figure 1. Binned COSAC mass spectrum shown with negative intensities after background subtraction, normalized to peak 18 that represents 2366 counts. Uncertainty on every peak due to low count statistics is shown to scale and represents ±17 counts or about 0.7% relative intensity. This is also the magnitude of our defined noise level, as it is both the intensity of the highest unexplainable peak (*m/z* 23) and of the most negative peak (*m/z* 32).

We also now have the added information from the ROSINA instrument onboard the Rosetta orbiter that published its results on a small grain impact believed to have come from the nucleus since the Goesmann *et al.* paper was published.[9] A direct comparison between both instruments' results is not possible but can still yield valuable information. Indeed, the grain ROSINA detected may have been ejected from a completely different location on the comet's nucleus and likely underwent sublimation processes during its "travel" to Rosetta. Whereas COSAC sampled freshly excavated grains likely much more pristine and loaded in volatiles. This is also evidenced by the very different peak distribution of the two mass spectra.[9]

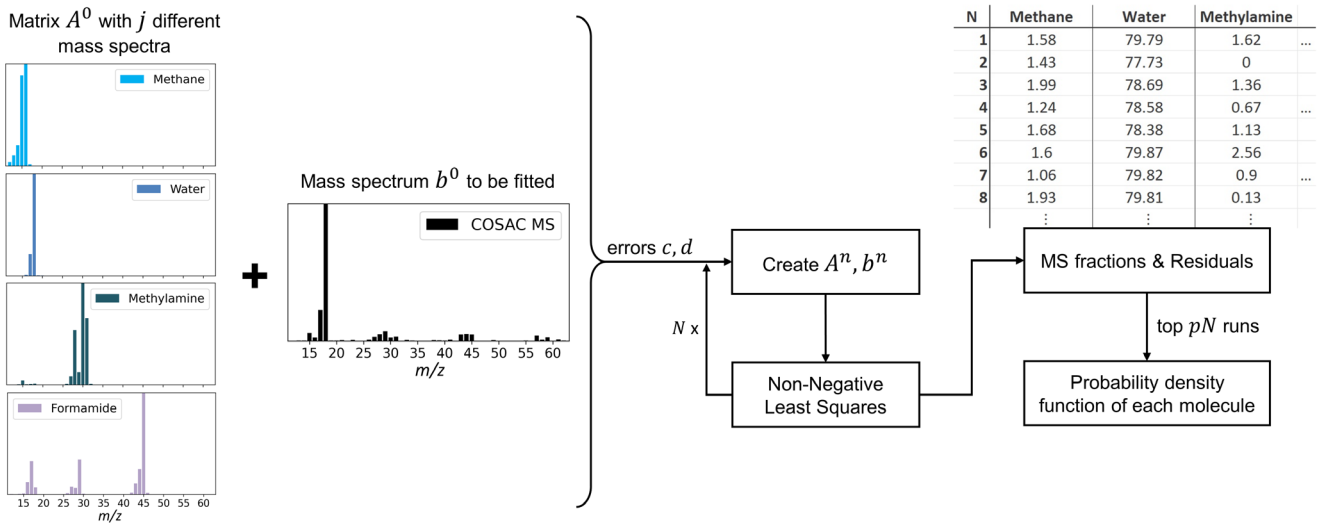
Non-Negative Least Squares fitting (NNLS) has already been used to fit the COSAC Mass Spectrum (CMS), but only on the proposed 16 molecules from the CT.[7] Meringer *et al.* used this method to obtain more mathematically rigorous abundances, under the hypothesis that the molecules in this spectrum are exactly the 16 proposed by the CT. Least squares fitting and single value deconvolution have been previously utilized in similar studies i.e. Wong *et al.* (2004) to analyze Jupiter's atmospheric composition with the Galileo Probe Mass Spectrometer,[10] Niemann *et al.* (2005) for Titan/GC-MS and Cui *et al.* (2009) for Titan/INMS.[11,12]

Gautier *et al.* (2020) introduced a Monte-Carlo approach to electron ionization (EI) mass spectra decomposition to take into account the 20–30% error bars on the peak intensities from the NIST database.[8] This method has been used on synthetic spectra of known composition, to retrieve accurate uncertainties on the relative abundances of the 7 molecules present in their spectrum. The method has also been successfully applied to Cassini/INMS data.[13] Gautier *et al.* also commented on the COSAC data, pointing out the need for counting statistics in addition to fragmentation pattern uncertainties, due to the spectrum's very low count rate and SNR.

Here we apply a randomization to the nominal fragmentation pattern of each molecule in our database and to the intensity of the peaks from the CMS, to create *N* (typically 10 000) different "possible realities" or runs. In each run we fit a modified COSAC spectrum with a modified EI fragmentation mass spectral database (see Scheme 1 and SI, "Computational method").

Previous calibration of the COSAC mass spectrometer showed good agreement between its produced fragmentation patterns and those from the NIST library.[3,14] Still, our model accommodates for appreciable differences between the two.

For a given set of original molecules we can then observe the behavior of their relative abundances under small variations in the initial conditions. This allows us to "lift" the degeneracy, in the way that we observe the abundance probability distribution for each molecule for a given set of initial molecules, and under the noise regime we input. Table 1 shows the comparison between the results obtained by the CT and our method after only NNLS fitting (N = 0), as well as when Monte Carlo iteration is added (N = 10000). For the latter, the mean, median and variance (MMV thereafter) of the MS fractions are shown. The biggest variable of this method is the set of molecules being fed to the algorithm. Different *a prioris* can be tested by removing or adding certain molecules to this initial pool. We can then observe deviations in the final list of molecules needed for the fit, and take note of the ones that are most often there or not there under slight changes in initial pool (we will talk about stable/unstable molecules in the fits). The quality of fits can be compared through their residuals and Ratio of Unexplained Intensities (RUI), both defined in the SI.



Scheme 1. Flowchart of the algorithm, with a visualization of the input and output files. The input errors *c* and *d* are the bounds used for calculating the randomized CMS intensities and NIST fragmentation patterns, respectively, at every iteration (see SI). After a full simulation the runs are ordered by residuals and only the top *p* percent "survive", from which all our data analysis follows.

Table 1. Results of a simulation using only the 16 COSAC molecules for *N* = 10000; *d* = 20%; *c* = 17 and *p* = 0.2 (see Scheme 1 and SI), as well as *N* = 0 (only NNLS fitting with original data) compared to the MS fractions found by the manual fit of the CT. The RUI is 1.5% better after NNLS fitting while using only 13 of the 16 molecules which is non-negligible: RUI (CT) = 12.1% and RUI (*N*=0) = 10.6%. The mean, median and variance (MMV) allow us to see the range of possible compositions under the hypothesis that no molecules other than these 16 are potentially present.

Molecule	Formula	Molar mass (u)	MS Fraction (CT)	MS Fraction (*N*=0)	Mean (*N*=10000[a])	Median (*N*=10000)	Variance (*N*=10000)
Water	H_2O	18	80.9	80.1	80.0	80.1	0.6
Methane	CH_4	16	0.7	1.9	2.0	2.0	0.1
Hydrogen cyanide	HCN	27	1.1	0.9	1.0	0.9	0.1
Carbon monoxide	CO	28	1.1	1.0	1.1	1.1	0.2
Methylamine	CH_3NH_2	31	1.2	2.0	1.7	1.8	0.3
Acetonitrile	CH_3CN	41	0.6	0.4	0.5	0.5	0.1
Isocyanic acid	HNCO	43	0.5	0.0	0.1	0.0	0.1
Acetaldehyde	CH_3CHO	44	1.0	3.0	2.9	2.9	0.3
Formamide	$HCONH_2$	45	3.7	3.5	3.5	3.5	0.2
Ethylamine	$C_2H_5NH_2$	45	0.7	0.0	0.2	0.0	0.2
Isocyanatomethane	CH_3NCO	57	3.1	2.7	2.6	2.5	0.1
Acetone	CH_3COCH_3	58	1.0	1.0	0.9	0.9	0.3
Propanal	C_2H_5CHO	58	0.4	0.8	0.8	0.8	0.1
Acetamide	CH_3CONH_2	59	2.2	1.3	1.4	1.4	0.1
Glycolaldehyde	$HOCH_2CHO$	60	1.0	0.0	0.1	0.0	0.0
Ethylene glycol	$(CH_2OH)_2$	62	0.8	1.3	1.4	1.3	0.3

[a] This means, since *p* = 20%, that the MMV values are from a set of 2000 data points.

Results and Discussion

After a first elimination process targeting only the most unlikely of molecules, our cleaned-up database (Table S2, Sheet "NIST_87") is composed of 87 EI mass spectra of molecules ranging from methane (16 u) to alanine (89 u). Then, after the first tests and the removal of an additional 4 molecules (nitromethane, nitrosomethane, methoxyethene and 2-propen-1-ol) we are left with 83 molecules. This process and the criteria for removing certain molecules are discussed extensively in the SI.

We then have 83 candidate molecules, but only about 20 peaks above noise in the CMS. Mathematically, to close the equation system for a single NNLS fit we cannot have more than 20 final candidates if we want the solution to be unique. Although Monte Carlo iteration allows us to see the extent of the degeneracy and has been used to trim the list down, the goal was to reduce our database to less than 20 compounds with a comfortable margin.

However, we aimed to give a full characterization of the CMS: for all 83 molecules in our database, we give a thorough analysis of the likelihood of each molecule being present at the comet and in what abundance, as well as the potential anti-correlations with other compounds. To do so, we progressively removed molecules from consideration based on the calculated likelihood of them being present in the CMS. The detailed results for individual molecules are found in Table S1, sheet "Results", where they are ranked in groups from least to most likely.

Starting from this database of 83 molecules, to be able to confidently remove candidates down to less than 20 molecules, we added more drastic initial condition variations in addition to the Monte Carlo randomization. The goal was to see, under different hypotheses, how much worse the fit became and what were the changes in the molecules used by the fit. For example: acetaldehyde is consistently one of the highest abundances under nominal conditions, but scaling down or completely removing the CMS peak at *m/z* 44 logically induces its progressive removal from the fit since its second highest peak is at this *m/z*. Interestingly, formaldehyde which was not required for the fit before, then becomes a core molecule with a consistently high abundance but the RUI is more than a percent worse. In effect, since *m/z* 44 is by far the most important peak of CO_2's MS, reducing it in the CMS is equivalent to forcing a certain amount of CO_2 in the fit. Consequently, this amounts to testing a different *a priori*. The absence of CO_2 in the CT fit has been the subject of debate, but we can see here that if we have no reason to not include acetaldehyde in the pool of molecules, its presence over CO_2 and formaldehyde is more likely by a non-negligible margin. However, to be as exhaustive as possible, molecules such as formaldehyde were not removed from the database at this point.

We created 4 different scenarios: nominal conditions (scenario 1), one where we reduced by up to 50% the intensity of the *m/z* 15 peak (scenario 2), one with an almost complete (80%) removal of the *m/z* 44 peak (scenario 3: CO_2 hypothesis explained above) and a last one where we only used the CT database and no molecules with a mass higher than 62 u (no addition of cyclopentanol, 3-pentanone, 2-methoxypropane, 2-methyl-2-propanol and neopentane as well as no glycine nor alanine, see SI subsection "Higher mass molecules"). To be even more thorough, for each scenario we also did additional sub-scenario runs where we removed from the initial pool different core molecules 1 by 1 to see which molecules would be "next in line", and how much worse the fit gets (Table 2). All these simulations are detailed in Table S3.

The insights of these results were two-fold: first, it allowed us to confidently remove from our database any molecule that is not used in any of the runs described above, which represents almost half of the database. Secondly, it allowed us to get an idea of how important each core molecule was (in each scenario) by looking at the increase in RUI after its removal from the initial pool. The results are shown in Table 2.

Table 2. Increase in RUI resulting from the removal of a given molecule for 4 different scenarios: **1** = nominal, **2** = *m/z* 15 halved, **3** = CO_2 (80% of *m/z* 44 removed), **4** = CT (only the molecules used by the COSAC team, none with a molar mass higher than 62 u). An increase in RUI of 0 means the molecule is not used in the given scenario. All these simulations were done using the 83 molecules database. Shown here are 19 molecules that are first choices in at least one scenario. The upper fitted *m/z* for all these simulations was 64 and not 87, hence the lower scenarios RUI compared to all other results in the article. Molecules are ordered from high to lower RUI in scenario 1 (nominal).

Scenarios	**1**	**2**	**3**	**4**		
RUI	6.67%	4.11%	6.98%	7.41%		
Molecule removed	**Increase in RUI (additive %)**				**Average[a]**	**Next best molecule(s)[b]**
Water	100+	100+	100+	100+	100+%	-
Cyclopentanol	0.69	0.20	0	0	0.22%	3-Pentanone, Neopentane
Acetaldehyde	0.50	0.42	0.01	0.93	0.47%	Alanine, Carbon dioxide, Propane
Methane	0.46	0.53	0.96	0.40	0.59%	Methoxyethane, Acetaldoxime, Ammonia
Acetone	0.40	0.45	0.66	0.29	0.45%	Butane, Isocyanic acid
Carbon monoxide	0.29	0.22	0.29	0.18	0.25%	Ethane
Ethylene glycol	0.28	0.33	0.35	0.40	0.34%	Ethanol
Hydrogen cyanide	0.25	0.23	0	0.20	0.17%	Ethane, Ethylene
Formamide	0.22	0.48	0.16	0.04	0.23%	2-Propanol, N-Methylformamide, Ammonia
Methylamine	0.14	0.31	0.23	0.22	0.23%	Monoethanolamine, Methyl nitrite
Methoxyethane	0.12	0.01	0.21	0.01	0.09%	2-Propanol, Formamide
N-Methoxy-methanamine	0.06	0.10	0.17	0.04	0.09%	-
2-Methoxypropane	0.03	0.04	0.34	0	0.10%	2-Methyl-2-propanol, N-Methylformamide
Isocyanatomethane	0.01	0	0	2.36[c]	0.59%[d]	Propanal, 2-Propen-1-amine
Ethane	0.01	0.01	0.03	0	0.01%	Ethylene
3-Pentanone	0	0.10	0.56	0	0.17%	Cyclopentanol, Neopentane
N-Methylformamide	0	0.08	0	0.59	0.17%	Acetamide
2-Propanol	0	0	0	0.09	0.02%	Methoxyethane, Formamide
Neopentane	0	0	0.03	0	0.01%	3-Pentanone

[a] This value assumes that all 4 scenarios are given the same weight, which is most likely false, and therefore should only be thought of as an indicator. [b] In order of decreasing mass. [c] As discussed in the SI subsection "Candidate molecules for *m/z* 57", without including higher mass molecules (Scenario 4), isocyanatomethane is the only possible contributor to *m/z* 57 with virtually no "next best molecules", therefore leaving the peak completely unfitted, and hence the huge increase in RUI after its removal. [d] This mean value specifically is probably overestimated.

We then removed molecules with consistent negligible contributions, before gradually removing molecules with minor abundances only in some sub-scenarios, down to 18 candidates. These are, adding carbon dioxide and barring isocyanatomethane and N-methoxy-methanamine, the ones shown in Table 2. This step-by-step trimming process is detailed in the SI.

For the final step we removed molecules that were consistently present in one of the alternate scenarios but not scenario 1 (nominal). This includes 3-pentanone, neopentane and most importantly carbon dioxide. CO_2 is a special case since scenario 3 is built around forcing it to cover 80% of *m/z* 44, therefore the RUI increase after its addition can be seen as simply the difference in RUI between scenario 3 (CO_2) and scenario 1 (nominal), which is 0.31%. After removing these three molecules from the database we were left with 15 candidates.

For the top 15 molecules, 2-propanol was a likely secondary choice to methoxyethane and formamide but is extremely poorly constrained. N-methylformamide is a likely contributor to *m/z* 59 and an almost perfect secondary choice to 2-methoxypropane. Depending on how much a lower molecular mass is valued over a slight reduction in RUI, either of these two could be chosen. Ethane is often present, but its removal had almost no impact on RUI as hydrogen cyanide and carbon monoxide easily compensate *m/z* 27 and 28 respectively. Figure 2 shows this effect visually and goes into further depth as to the reasoning why these 3 molecules did not make the final cut.

Finally, after removing these 3 from consideration, we were left with our top 12 shortlist of compounds (Table 3). In order of most important to least important (Table 2), and only counting molecules that are core in at least 3 out of 4 scenarios: water, methane, acetaldehyde, acetone, ethylene glycol, carbon monoxide, formamide, methylamine, hydrogen cyanide, 2-methoxypropane, methoxyethane. Cyclopentanol is in its own category and is discussed in the next paragraph. A simple least squares fit ($N = 0$) using these 12 molecules is shown in Figure 3. From Table 3, we note that formamide and methoxyethane have higher variances due to both molecules having a base peak at *m/z* 45 and therefore compete for the fit at this *m/z*. The variance is an error bar on the MS fraction of each molecule under the hypothesis that these and only these 12 molecules compose the CMS. As is evidenced in Figure 2, adding molecules in the database will cause the variances of certain molecules to increase significantly. Even though it might appear surprising from the low MS fraction and molecular abundance, ethylene glycol is found to be one of the most important and constant molecules present in the CMS. Methoxyethane is one of the least important core molecules, as evidenced by the relatively low increase in RUI after its removal in most scenarios. It is also dependent on its *m/z* 15 contribution, making it less reliable, as discussed in the SI. To a lesser extent this is also true for 2-methoxypropane.

The last column of Table 3 shows that for larger molecules like cyclopentanol and 2-methoxypropane, a large MS fraction very quickly diverges from meaning a high molecular fraction. Table S3 details all simulations made for Table 2 and gives a broader view of all possible results. Whether there is nitromethane or not, under the assumption that there is no methoxyethene and no 2-propen-1-ol, the 12 molecules listed in Table 3 are statistically the most likely to compose the CMS. Table S1 shows the likelihood of presence for every molecule in our database, represented by a color (from dark orange for least likely to dark green for most likely) which shows at which step of the trimming process this molecule was removed from consideration.

The presence of cyclopentanol in such quantities is a symptom of the problems faced when interpreting this MS, of which the most important is the virtual cut-off at *m/z* 61 while the *m/z* 57 and 59 peaks are so pronounced. While this “cut-off” heavily limits the number of potential higher mass molecules present, there are still some heavier than 62 u to consider, such as cyclopentanol and the other molecules proposed here. Other heavier molecules with great fragmentation pattern accordance with the CMS are for example: propylene glycol (76 u), dimethyl carbonate (90 u) and 2-oxobutanoic acid (102 u; this molecule is already in Table S2 of Goesmann *et al.* as an example of a higher mass molecule to explain the *m/z* 57 peak).

This is the reason why, even though cyclopentanol is a perfect fit for *m/z* 57 and the rest of the spectrum, we cannot confidently say that it is part of this MS, but it is the lowest mass molecule that can perfectly explain the *m/z* 57 peak. This is in contrast with nitromethane, which as discussed in the SI, does not lead to a perfect fit. While our database is virtually complete for molecules with $m < 62$ u, the ones with a higher mass are hand-picked for our purpose.

Larger polymer molecules could perhaps be proposed in a manner similar to that adopted by the Ptolemy team during interpretation of their mass spectrum, although they had clear peaks up to *m/z* higher than 100.[15] Due to their much higher 70 eV electron ionization cross sections, a smaller amount of these bigger molecules will still create a significant signal. This is already the case for cyclopentanol (86 u) with a cross section of 16.29 Å² compared to 6.73 Å² for acetaldehyde (44 u) or 4.35 Å² for methane (16 u) for example (see SI). This infers that cyclopentanol is more than 2 and 3 times more likely (than acetaldehyde and methane, respectively) for cyclopentanol to form a fragment in the mass spectrometer chamber due to the size of its electron cloud. Therefore, its fragmentation pattern will be seen in the MS with that much more intensity (Table 3).

The absence of any sulfur-bearing species in this fit to the COSAC data is surprising, especially considering that the ROSINA instrument onboard the Rosetta orbiter has identified a number of such compounds.[16] The absence of any appreciable amounts of ammonia is similarly puzzling, as compounds with amino- or amid- functions, which are formed in reactions with ammonia, are used in the fit. We hypothesize that the depletion of these compounds in the gas phase is caused by their very efficient adsorption on metal surfaces, here the walls of the pipes and the inside of the instrument. Both ammonia and thiols are notorious for sticking to steel surfaces in vacuum vessels where they can only be removed by heating the walls and pumping for a long time. Icy dust entering the instrument and slowly warming up to 12-15°C are almost perfect conditions for ensuring maximum coverage of container walls. It could be that a small abundance of ammonia and/or thiol compounds in the ice was thus lost to adsorption after transition to the gas phase. It is also important to remind that ROSINA most likely sampled material coming from a very different site of the nucleus of the comet.

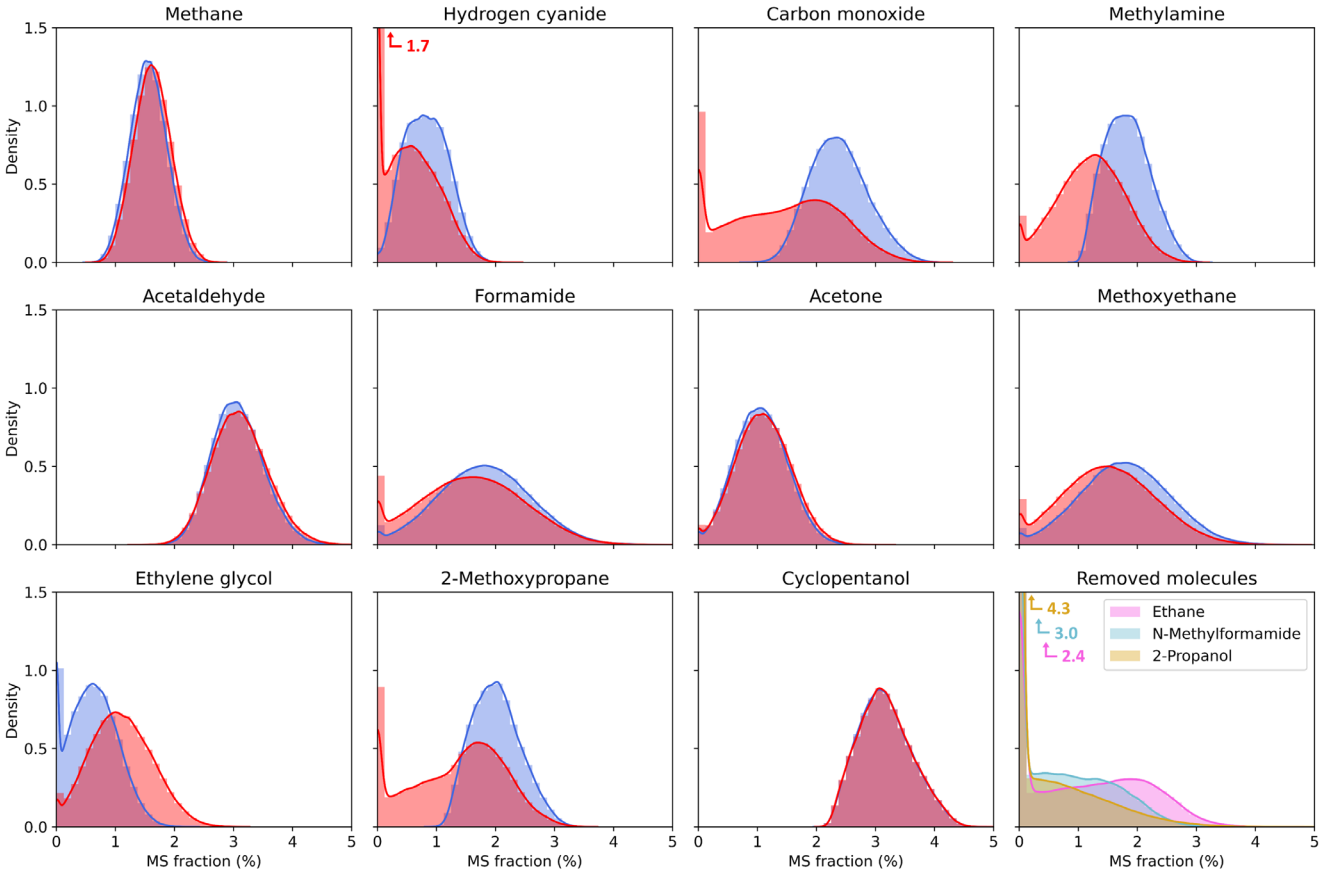


Figure 2. Probability density function estimations of the 15 (red curves and bottom right plot) and 12 (blue curves) most likely molecules composing the CMS as found by our trimming method. In slight transparency are the raw histogram from the simulations from which the kernel density estimation was made. For better visualization the simulations shown here were done using $N = 1000000$. The bottom right panel shows the 3 molecules (ethane, N-methylformamide and 2-propanol) that are removed to go from the 15- to 12-compound database. As evidenced here, these are very poorly constrained with a high fraction of non-utilization (26%, 32% and 48% respectively). This makes them less likely candidates when compared to the final 12 molecules. The rest of the panels show the effect that adding these 3 unstable compounds (fit-wise) has on the final 12 molecules by comparing their probability density functions under the two hypotheses of initial pool (red with and blue without). Detailed analysis of these results can be found in the SI subsection "Additional comments on final results and comparisons".

Table 3. Results of a simulation with only the final 12 molecules of our trimming process**.** RUI (N=0) = 8.56%. Removing ethane and N-methylformamide from the pool of molecules only cost 0.13% in RUI (Figure S1), hence why we assume they are not necessary to our final list of compounds to best fit the COSAC Mass Spectrum (CMS).

Molecule	Molar mass (u)	Formula	MS Fraction (N=0)	Mean (N=10000)	Median (N=10000)	Variance (N=10000)	Impact cross-section at 70 eV (Å²)[a]	Molecular fraction relative to water[b]
Water	18	H_2O	79.9	79.8	79.8	0.5	2.43	100
Methane	16	CH_4	1.6	1.6	1.6	0.1	4.35	1.1
Hydrogen cyanide	27	HCN	0.8	0.8	0.8	0.1	3.40	0.8
Carbon monoxide	28	CO	2.4	2.4	2.4	0.2	2.68	2.7
Methylamine	31	CH_3NH_2	1.8	1.8	1.8	0.1	6.63	0.8
Acetaldehyde	44	CH_3CHO	3.1	3.1	3.1	0.2	6.73	1.4
Formamide	45	$HCONH_2$	1.8	1.8	1.8	0.6	6.37	0.9
Acetone	58	CH_3COCH_3	1.1	1.1	1.0	0.2	9.67	0.3
Methoxyethane	60	$C_2H_5OCH_3$	1.7	1.8	1.8	0.5	11.39	0.5
Ethylene glycol	62	$(CH_2OH)_2$	0.6	0.6	0.6	0.2	9.14	0.2
2-Methoxypropane	74	$C_3H_7OCH_3$	2.0	2.0	2.0	0.2	16.36	0.4
Cyclopentanol	86	C_5H_9OH	3.1	3.2	3.1	0.2	17.31	0.5

[a] Details can be found in the Supporting Information. [b] In the ionization chamber, not of the cometary material. Calculated from MS fraction (N=0) and impact cross-section at 70 eV.

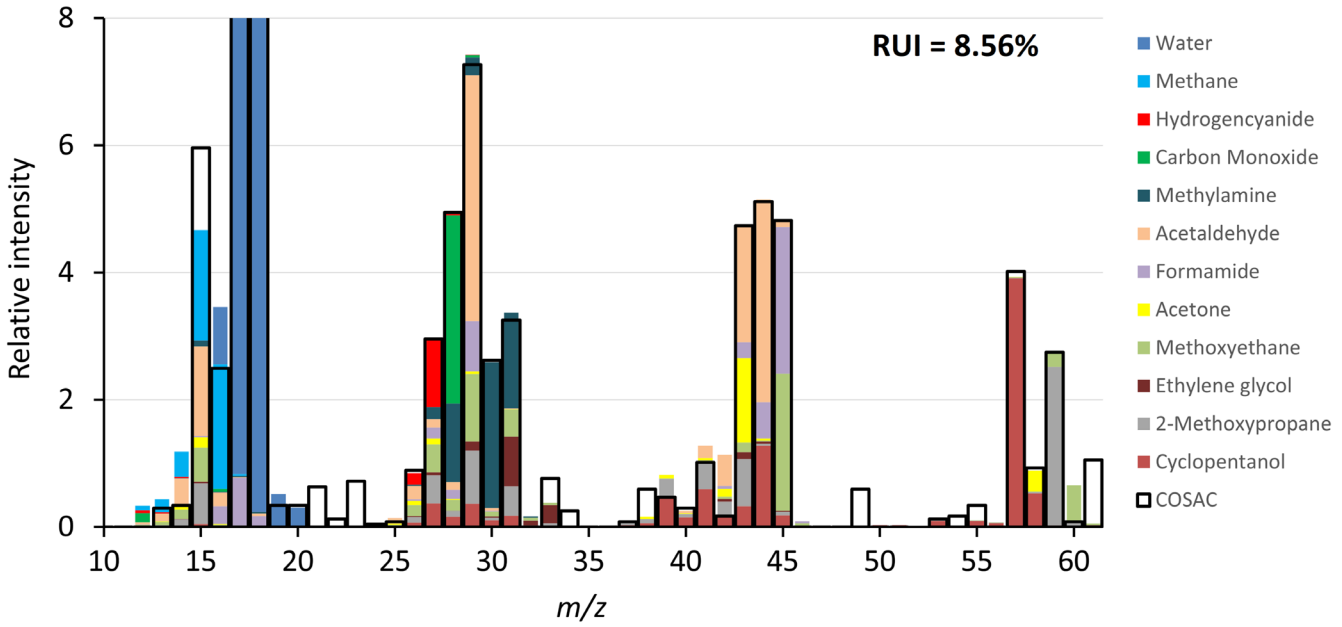


Figure 3. Individual color-coded contributions of molecules to the fitting of the CMS (black outline) when using our shortlist of 12 molecules. This is the fit without Monte Carlo iteration ($N = 0$), meaning this is the exact CMS fitted by exact NIST mass spectra. The same plot comparing the CT fit, this figure as well as the same one with the top 14 molecules (adding ethane and N-methylformamide) is shown in Figure S1.

Conclusion

This new study based on data from the Philae lander of the Rosetta space mission unveils a suite of 12 organic molecules originating from the nucleus of comet 67P. Starting from a NIST mass spectra database of 120 compounds, we gradually removed molecules from the most to the least obvious non-candidates to appear in the CMS, by testing of exhaustive initial condition variations. NNLS fitting and Monte Carlo simulations allowed us to observe the range of possible compositions from the CMS under the hypothesis that this mixed mass spectrum is the sum of individual contributions. Previous competence evaluation done on the flight spare model of the COSAC mass spectrometer allows us to be confident in its fragmentation patterns accordance with the NIST standard database.[14]

Our model is applicable to any mass spectral deconvolution problem. The CMS is an extreme case due to its very low count and almost no prior constraint on possible molecular detections, but the algorithm manages to discern chemically consistent results, consolidating the detection of more than half of the 16 molecules proposed by the CT. Indeed, 9 out of our final 12 molecules were also found in the original fit: water, methane, hydrogen cyanide, carbon monoxide, methylamine, acetaldehyde, formamide, acetone and ethylene glycol. The 3 that are not are methoxyethane, 2-methoxypropane and cyclopentanol. For the first two, 2-propanol and N-methylformamide respectively are the next best candidates by a very small margin and likely co-contributors. Cyclopentanol is the lowest mass molecule capable of fitting *m/z* 57 while in perfect accordance with the rest of the CMS. Our solution is not unique; however, from our thorough trimming process we show that these are the most likely candidates from our database to comprise the CMS. Glycolaldehyde, propanal and isocyanatomethane were not found by ROSINA and also rejected in the present analysis of the COSAC spectrum, consequently the disagreement between the results of the two instruments was reduced.

It has been assumed that when a cosmic cloud of gas and dust condensed into the solar system, its molecular inventory was largely preserved in comets. The *in situ* investigation of the cometary nucleus by the COSAC instrument onboard Rosetta's Philae lander and data analyses through non-negative least squares fitting and Monte Carlo simulations now confirm this assumption and reveal the presence of volatile molecules such as water, carbon monoxide, methane, and hydrogen cyanide. These detections indicate moreover that the cometary chemical inventory includes molecules issued from carbon-carbon and carbon-oxygen bond formations yielding a variety of oxygenated organics including acetaldehyde, ethylene glycol, and others. A lower involvement of nitrogen-bearing compounds such as methylamine and formamide was identified as well. Sulfur-containing species and amino acids have not been detected.

Future experimental and theoretical studies, including space missions,[17] will follow to investigate the mechanisms of formation of these pristine molecules and their further evolution towards higher complexity.

Acknowledgements

G.L and U.M acknowledge the financial support of the ANR (ANR-15-IDEX-01 and ANR-18-CE29-0004). T.G. acknowledges the financial support from the Programme National de Planétologie (PNP) of CNRS/INSU co-funded by CNES and of the ANR (ANR-20-CE49-0004-01) for the development of the Monte-Carlo inversion method. G.M.M.C. acknowledges the Spanish MICINN under project PID2020-118974GB-C21 (AEI/FEDER, UE) and the Unidad de Excelencia 'María de Maeztu' MDM-2017-0737–Centro de Astrobiología (INTA-CSIC).

Keywords: Analytical Methods • Comet • Mass spectrometry • Philae • Rosetta

[1] S. Ulamec, C. Fantinati, M. Maibaum, K. Geurts, J. Biele, S. Jansen, O. Küchemann, B. Cozzoni, F. Finke, V. Lommatsch, A. Moussi-Soffys, C. Delmas, L. O´Rourke, *Acta Astronaut.* **2016**, *125*, 80–91.

[2] J. Biele, S. Ulamec, M. Maibaum, R. Roll, L. Witte, E. Jurado, P. Munoz, W. Arnold, H.-U. Auster, C. Casas, C. Faber, C. Fantinati, F. Finke, H.-H. Fischer, K. Geurts, C. Guttler, P. Heinisch, A. Herique, S. Hviid, G. Kargl, M. Knapmeyer, J. Knollenberg, W. Kofman, N. Komle, E. Kuhrt, V. Lommatsch, S. Mottola, R. Pardo de Santayana, E. Remetean, F. Scholten, K. J. Seidensticker, H. Sierks, T. Spohn, *Science (80-.).* **2015**, *349*, aaa9816–aaa9816.

[3] F. Goesmann, H. Rosenbauer, R. Roll, C. Szopa, F. Raulin, R. Sternberg, G. Israel, U. J. Meierhenrich, W. H.-P. Thiemann, G. M. Muñoz Caro, *Space Sci. Rev.* **2007**, *128*, 257–280.

[4] F. Goesmann, F. Raulin, J. H. Bredehöft, M. Cabane, P. Ehrenfreund, A. J. MacDermott, S. McKenna-Lawlor, U. J. Meierhenrich, G. M. Muñoz Caro, C. Szopa, R. Sternberg, R. Roll, W. H.-P. Thiemann, S. Ulamec, *Planet. Space Sci.* **2014**, *103*, 318–330.

[5] H. Krüger, F. Goesmann, C. Giri, I. P. Wright, A. D. Morse, J. H. Bredehöft, S. Ulamec, B. Cozzoni, P. Ehrenfreund, T. Gautier, S. McKenna-Lawlor, F. Raulin, H. Steininger, C. Szopa, *Astron. Astrophys.* **2017**, *600*, A56.

[6] F. Goesmann, H. Rosenbauer, J. H. Bredehöft, M. Cabane, P. Ehrenfreund, T. Gautier, C. Giri, H. Krüger, L. Le Roy, A. J. MacDermott, S. McKenna-Lawlor, U. J. Meierhenrich, G. M. M. Caro, F. Raulin, R. Roll, A. Steele, H. Steininger, R. Sternberg, C. Szopa, W. H.-P. Thiemann, S. Ulamec, *Science (80-.).* **2015**, *349*, aab0689–aab0689.

[7] M. Meringer, C. Giri, H. J. Cleaves, *ACS Earth Sp. Chem.* **2018**, *2*, 1256–1261.

[8] T. Gautier, J. Serigano, J. Bourgalais, S. M. Hörst, M. G. Trainer, *Rapid Commun. Mass Spectrom.* **2020**, *34*, DOI 10.1002/rcm.8684.

[9] K. Altwegg, H. Balsiger, J.-J. Berthelier, A. Bieler, U. Calmonte, S. A. Fuselier, F. Goesmann, S. Gasc, T. I. Gombosi, L. Le Roy, J. De Keyser, A. D. Morse, M. Rubin, M. Schuhmann, M. G. G. T. Taylor, C.-Y. Tzou, I. P. Wright, *Mon. Not. R. Astron. Soc.* **2017**, *469*, S130–S141.

[10] M. H. Wong, P. R. Mahaffy, S. K. Atreya, H. B. Niemann, T. C. Owen, *Icarus* **2004**, *171*, 153–170.

[11] H. B. Niemann, S. K. Atreya, S. J. Bauer, G. R. Carignan, J. E. Demick, R. L. Frost, D. Gautier, J. A. Haberman, D. N. Harpold, D. M. Hunten, G. Israel, J. I. Lunine, W. T. Kasprzak, T. C. Owen, M. Paulkovich, F. Raulin, E. Raaen, S. H. Way, *Nature* **2005**, *438*, 779–784.

[12] J. Cui, R. V. Yelle, V. Vuitton, J. H. Waite, W. T. Kasprzak, D. A. Gell, H. B. Niemann, I. C. F. Müller-Wodarg, N. Borggren, G. G. Fletcher, E. L. Patrick, E. Raaen, B. A. Magee, *Icarus* **2009**, *200*, 581–615.

[13] J. Serigano, S. M. Hörst, C. He, T. Gautier, R. V. Yelle, T. T. Koskinen, M. G. Trainer, *J. Geophys. Res. Planets* **2020**, *125*, DOI 10.1029/2020JE006427.

[14] C. Giri, F. Goesmann, A. Steele, T. Gautier, H. Steininger, H. Krüger, U. J. Meierhenrich, *Planet. Space Sci.* **2015**, *106*, 132–141.

[15] I. P. Wright, S. Sheridan, S. J. Barber, G. H. Morgan, D. J. Andrews, A. D. Morse, *Science (80-.).* **2015**, *349*, aab0673–aab0673.

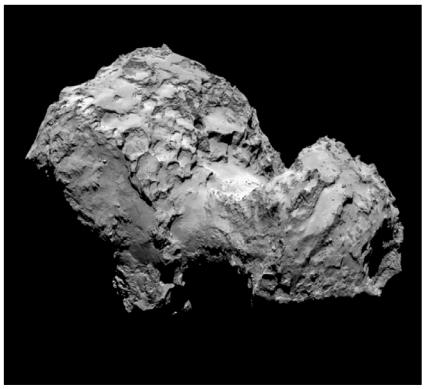
[16] M. Rubin, K. Altwegg, H. Balsiger, J.-J. Berthelier, M. R. Combi, J. De Keyser, M. Drozdovskaya, B. Fiethe, S. A. Fuselier, S. Gasc, T. I. Gombosi, N. Hänni, K. C. Hansen, U. Mall, H. Rème, I. R. H. G. Schroeder, M. Schuhmann, T. Sémon, J. H. Waite, S. F. Wampfler, P. Wurz, *Mon. Not. R. Astron. Soc.* **2019**, *489*, 594–607.

[17] N. Thomas, S. Ulamec, E. Kührt, V. Ciarletti, B. Gundlach, Z. Yoldi, G. Schwehm, C. Snodgrass, S. F. Green, *Space Sci. Rev.* **2019**, *215*, 47.

10.1002/anie.202201925

WILEY-VCH

Entry for the Table of Contents



Credit ESA/Rosetta/MPS for OSIRIS Team MPS/UPD/LAM/IAA/SSO/INTA/UPM/DASP/IDA

ESA's comet rendezvous mission Rosetta investigated the nucleus of comet 67P/Churyumov-Gerasimenko to reveal information about the most pristine material preserved in the Solar System. The analysis of the mass spectrum recorded twenty-five minutes after touchdown by non-negative least squares fitting and Monte Carlo simulations shows the presence of twelve organic molecules that represent the inventory of cometary nuclei and the early Solar System.

Institute and/or researcher Twitter usernames: @MhenriU, @NiceChemistry