Taming Combinatorial Explosion of the Formose Reaction *via* Recursion within Mineral Environments

*Stephanie Colón-Santos, Geoffrey J. T. Cooper and Leroy Cronin**

School of Chemistry, University of Glasgow, University Avenue, Glasgow, G12 8QQ, UK.
E-mail: lee.cronin@glasgow.ac.uk

Abstract *One-pot reactions of simple precursors, such as those found in the formose reaction or formamide condensation, continuously lead to combinatorial explosions in which simple building blocks capable of function exist, but are in insufficient concentration to self-organize, adapt, and thus generate complexity. We set out to explore the effect of recursion on such complex mixtures by 'seeding' the product mixture into a fresh version of the reaction, with the inclusion of different mineral environments, over a number of reaction cycles. Through untargeted UPLC-HRMS analysis of the mixtures we found that the overall number of products detected reduces as the number of cycles increases, as a result of recursively enhanced mineral environment selectivity, thus limiting the combinatorial explosion. This discovery demonstrates how the involvement of mineral surfaces with simple reactions could lead to the emergence of some building blocks found in RNA, Ribose and Uracil, under much simpler conditions that originally thought.*

The mechanism which led to the first genetic-machine, an adaptive, chemical system that uses a genetic code to organise metabolic function, and propagate that code, is one of the most important outstanding questions in science.[1,2] Modern organisms are genetic machines that take part in open-ended information transfer using biopolymers such as RNA and DNA, which are ubiquitous to all known life forms. *De novo* DNA and RNA synthesis has been accomplished from isolated ribose and nucleobases in a multi-step synthesis, but not from a simple or prebiotic route to sugars or purines.[5,6] The one-pot synthesis of all the required compounds can be achieved through a diverse set environmental conditions,[7–10] but they always result in a convoluted, and analytically intractable, complex mixture of products.[11-15] Identification of the direct transition of such units into polymers from these mixtures is very challenging analytically, and complete chemical characterization is nearly impossible, as in the case of tholins.[16] As such, the combinatorically large number of products can justify employing a less product-explosive process involving a multi-step synthesis approach. However, the interaction of simple molecules with the environment has been proven to steer the chemical networks into different outcomes or product populations, giving them a higher level of order as a result of environmental constraints (such as inorganic catalysts).[17-20] In particular, the presence of mineral surfaces is known to sometimes truncate the combinatorial explosions generated by one-pot reaction of simple compounds. Two relevant examples are the preferential formation of ribose when borate minerals were added to the formose reaction,[21] a system known for the incredible complexity of its product distribution, and the clear selectivity towards the production of certain nucleobases when formamide condensation was carried out on different mineral surfaces.[22] Notably, these previous results were obtained in batch reactions, leaving the possibility that this effect could be amplified if the reaction mixture was cycled over a given environment.

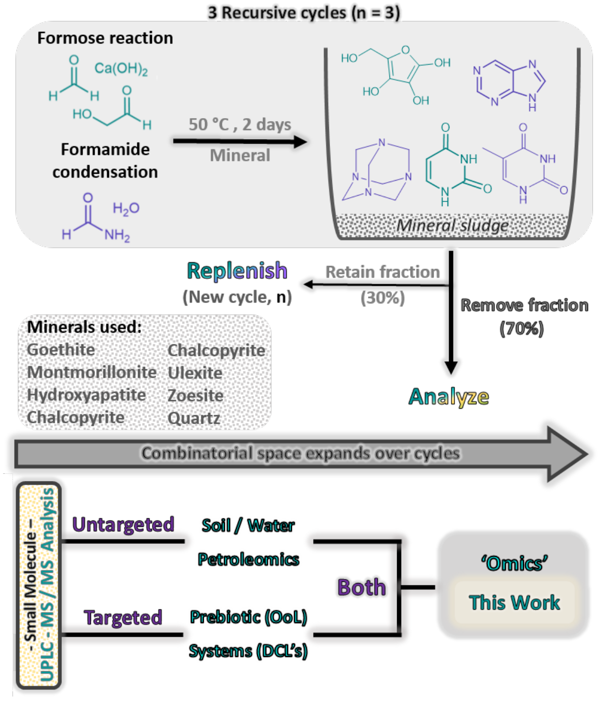


Figure 1. *Recursive cycles:* A formose reaction (green) in formamide/water (purple) is carried out in the presence of a mineral. After each cycle of 48 hours at 50 ºC, a fraction of the total volume (70%, from the top) is removed and the vial is replenished with fresh starting materials to start the next cycle. *UPLC-MS/MS analysis*: An untargeted analysis, followed by a targeted data processing was conducted in order to explore the resulting product distribution with an 'omics' approach.

To investigate this, we set out to explore the effect of reaction cycling by seeding with the products of the previous reaction cycle (recursion) on well-known combinatorial explosions. We carried out the formose reaction in formamide with different mineral environments, see **Figure 1**, to assess whether the selectivity imparted by the environment can be amplified through recursion, whilst truncating the combinatorial explosion by reducing the overall number of products. We found that the recursive action resulted in a lower number of individual products, with or without a mineral surface, demonstrating that reaction cycling has a significant effect on the product distribution. We also observed a significant increase in the yields of certain species when minerals were present, showing that selection by the environment also plays a role in determining the product mixture, see **Figure 2**.

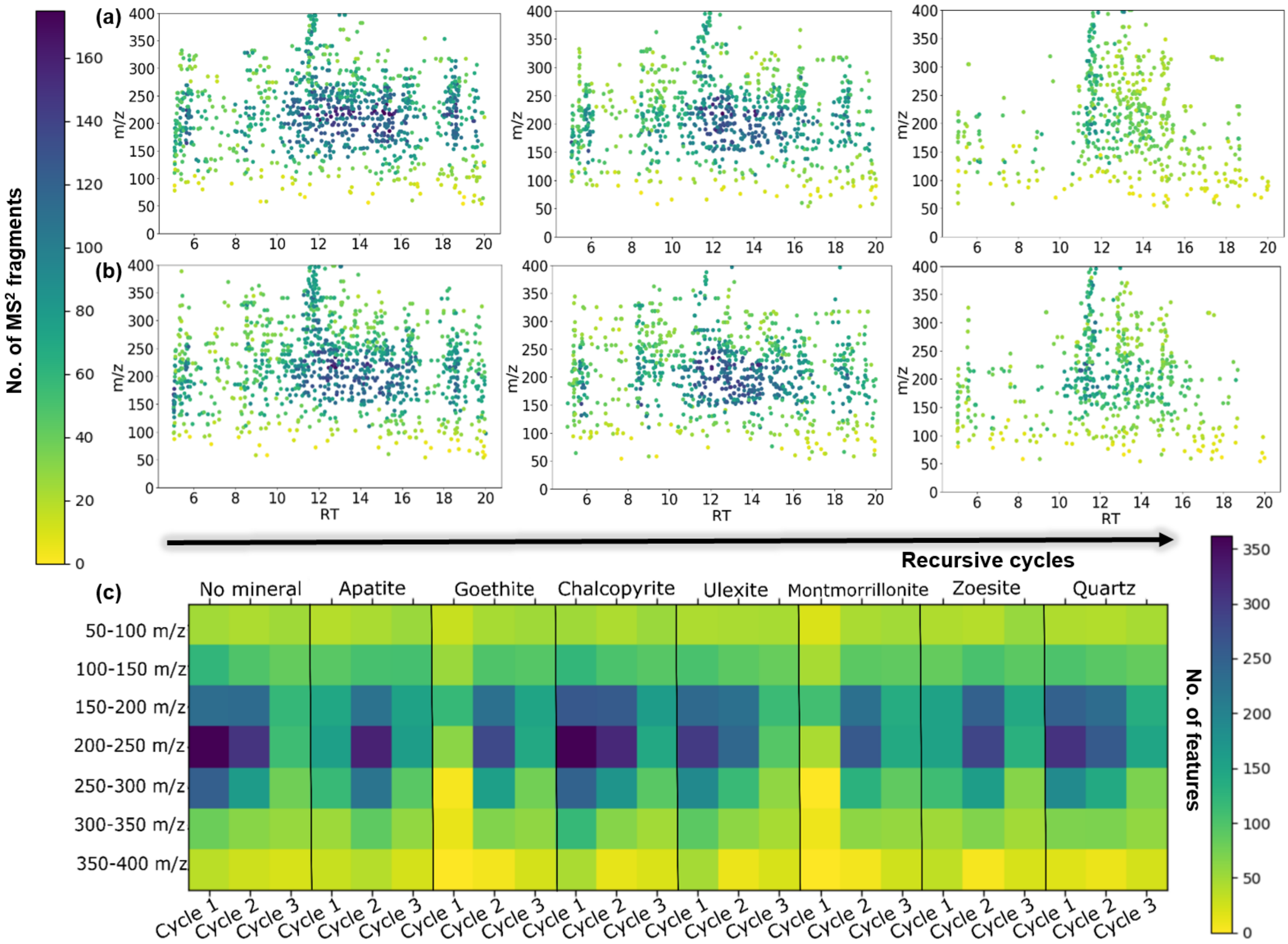


Figure 2. *Mass spectral features:* Features are based on unique exact mass (m/z) and retention time (RT). Over recursive cycles, differences in the number of features and MS^2 fragments (of each feature) can be observed for both (a) the recursive formose reaction in formamide control (no mineral) and (b) the recursive reaction in the presence of a mineral surface (Chalcopyrite, Cu_2FeS). A heatmap of the features (c) was generated by grouping the features into 50m/z bins, resulting in a unique pattern for each reaction environment over the three recursive cycles.

In order to investigate and establish the nature of any differences in the product distribution without bias, untargeted analysis of the mixtures was conducted. Ultra-Performance Liquid Chromatography (HILIC) coupled to tandem mass-spectrometry (UPLC-MS/MS) was carried out in a data dependent fashion, which allowed us to investigate the resulting chemical space without having to target any particular compound. By generating features based on exact mass (m/z) and retention time (RT), we were able to achieve a meaningful representation of the product distribution from mass-spectral data. The features represent unique reaction products and their number maps to the number of individual species, providing a way to gauge the complexity of the mixture.

To make the large volume of data more accessible, detected features were binned by their m/z values as a means to fingerprint the product distribution (**Figure 2**). The number of features in each range of molecular weight changes as an effect of recursive action, with a general trend that the number decreases from Cycle 1 to Cycle 3 (**Figure 2 a & b**). Differences in the distribution also arise as an effect of the environment, as observed in **Figure 2 c**, between the reaction with no mineral and with the inclusion of a mineral surface. For each environment, the features generate a different pattern which also changes across recursive cycles. The number of detected features decreases from Cycle 2 to Cycle 3 for all reactions, demonstrating that the action of recursive cycles is limiting the combinatorial explosion expected from these reactions. In the case of Chalcopyrite, Quartz and in the absence of any mineral surface (control), the number of features reduces linearly from Cycle 1 to Cycle 3, while all other mineral environments see the number of features peak in Cycle 2 and decrease again in Cycle 3. This suggests that the reactions proceed along different trajectories, towards different product distributions, as a direct result of the mineral environment. To validate our in-house feature generator and 'omics' based approach to complex mixture analysis, we processed the data using CompoundDiscoverer™ (Thermo Scientific),[23] a conventionally used software for processing untargeted mass-spectral data, which also enabled the extraction of ion chromatograms (EIC's) in a targeted fashion. While this method generated fewer features overall, the trends were consistent throughout the experiments (see **Figure 2** and **Figure S4).**

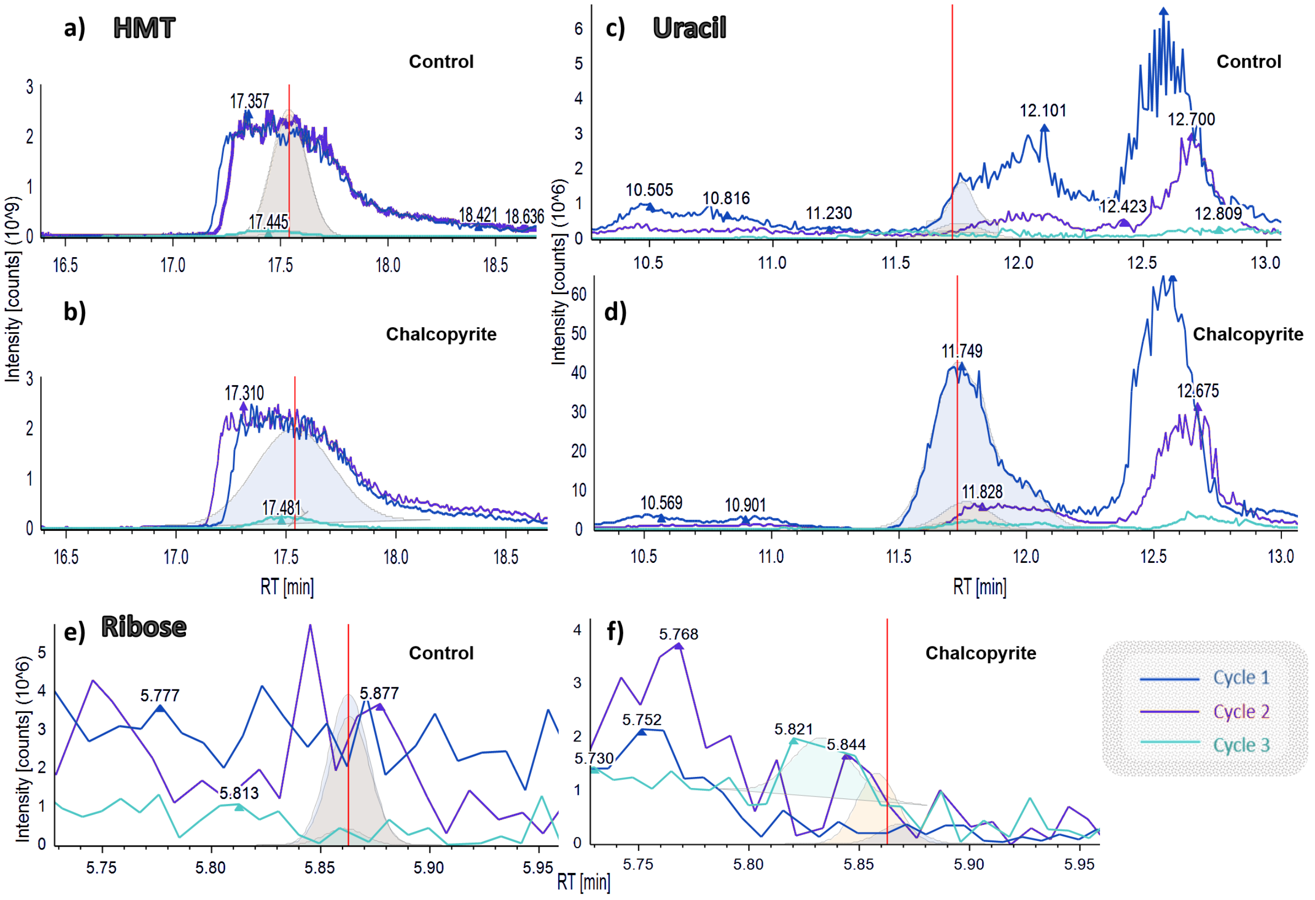


Figure 3. *The formaldehyde sink, Hexamethylenetetramine (HMT):* Extracted Ion Chromatograms for HMT (m/z: 141.11, Adduct: [M+H]) for **(a)** the control reaction and **(b)** in the presence of a mineral surface (Chalcopyrite). *RNA building blocks, Ribose and Uracil:* Extracted Ion Chromatograms for Ribose (m/z: 172.96, Adduct: [M+Na]) **(e)** in the control reaction and **(f)** in the presence of a mineral surface (Chalcopyrite). As well as, for Uracil (m/z: 113.03, Adduct: [M+H]) **(c)** in the control reaction and **(d)** in the presence of a mineral surface.

During the acquisition of the mass-spectral data, the most intense peaks were fragmented further to MS^2. This allowed us to identify some of the products using database matching and validation against pure standards to confirm chemical identities. By using the MS^2 data, we were able to identify some of the features as Ribose and Uracil, the building blocks of RNA. Extracted ion chromatograms (EICs) for Ribose and Uracil, for the reaction with and without the mineral chalcopyrite, are shown in **Figure 3 *c-f***. We found that conventionally analytically targeted products, such as nucleobases, where not only present in our product mixtures but also produced preferentially on mineral surfaces, as observed in the difference between intensity scales in **Figure 3 *c, d***. As well as, traces of nucleoside formation (Thymidine and Adenosine) for most samples in Cycle 3, including the non-mineral control reaction (see **Figure S11 – S13**).

In addition, we detected Hexamethylenetetramine (HMT) across all reactions. HMT was discovered by Aleksandr Butlerov in 1859 and is prepared industrially by combining formaldehyde and ammonia.[24] The significance of HMT in prebiotic chemistry has been discussed previously,[25] particularly in its role of incorporating formaldehyde (from its reaction with ammonia, which is generated *in-situ* by the decomposition of formamide) into a more stable compound, possibly allowing for it to be concentrated in a prebiotic, evaporative environment. The concentration of HMT changed across recursive cycles (**Figure 3**), with a significant drop being observed after the second cycle for all samples (including the control). We postulate that the HMT is depleted by reaction with the products of Cycle 2, but we currently have no definitive evidence for this, or a mechanism responsible.

In conclusion, we carried out the formose reaction and formamide condensation in a one-pot fashion, under milder conditions than previously reported,[2] while a recursive environment was applied to the resulting mixture in a series of cycles. We found that recursive cycles not only truncated the combinatorial explosion by reducing the number of individual products, but also successfully generated sugars and nucleobases from potentially prebiotic routes, in an integrated fashion. Traces of nucleoside formation were also detected after two recursive cycles, for the first time in this simple-precursor systems (e.g. Formose reaction/ Formamide condensation). Furthermore, we found a molecule with a strong connection to prebiotically-relevant compounds, hexamethylenetetramine (HMT), which might have a non-trivial relationship with the formation of these building blocks. We believe that recursive experiments bring us one step closer to a plausible 'real-life' scenario, and therefore provide an improved experimental regime for looking at the evolution of complex mixtures from simple precursors under non-equilibrium conditions.

Experimental

Experimental Methods: A formose reaction (formaldehyde, glycolaldehyde, calcium hydroxide) was carried out in in formamide-water (50:50 v/v) on seven different mineral surfaces (*see **SI**, Page 2*), as well as, in the absence of any mineral surface (*e.g.* control). The reactions were stirred at 1200rpm and heated at 50°C, for 48 hours. Then, about 70% (~3.5 mL) of the reaction volume (supernatant) was removed for analysis.

Recursive cycles: The remaining fraction (~1.5 mL) was used to seed the next reaction. Topping up with the same concentration of starting materials (3.5 mL), but conserving the total reaction volume (5 mL); we repeated the process.

Sample preparation: The removed fraction was allowed to cool to room temperature. Then, a 100µL aliquot was taken for each analysis; to which an ion-exchange resin was employed to remove excess cations in solutions (*e.g.* Ca^{2+}) and the supernatant transferred to glass vial, followed by a 1 in a 100 dilution with MS grade water. Finally, the solution was filtrated with a syringe filter (0.22µm cut-off)].

Ultra-performance Liquid Chromatography and tandem Mass Spectrometry: Chromatographic separation was achieved using a Thermo Vanquish UPLC with a ZIC-HILIC column, eluted in a linear gradient mixture of solvents A (water w/20 mM Ammonium Acetate, pH = 5) and B (100% acetonitrile w/0.1% v/v formic acid) over 25 min, coupled to a Thermo Fusion Orbitrap for mass-spectral analysis. Spectra were collected for 30 minutes in positive mode over a scan range of 50–500 m/z. Ion transfer tube was set to 275 °C, RF lens 60%, and acquisition was performed in a Data-dependent (DDA) manner. The Fragmentation data was collected at top speed (3 second window) with an intensity threshold of 5.0E4 and dynamic exclusion, after one time for 15 seconds, using the ion trap isolation at HCD collision energy of 35 eV and resolution 15000.

Interpretation of Raw Data: All raw files were converted to mzML and centroided using Proteowizard's [26] convert function (with a vendor-specific algorithm). The converted files (mzML) were processed in Python using Pymzml. In each file, (m/z, intensity, rt) features were extracted using pymzml feature detection algorithm, with default parameter values used for both the centroiding and mass trace detection. Performance of the feature detection and extraction algorithm was evaluated by comparing them with those generated in an analogous processing software, CompoundDiscoverer™, which was developed particularly for data acquired in Thermo-Orbitrap instruments and used to automatically detect features across samples; which were comparable with those obtained with Pymzml.

Data Analysis: After aligning the peaks detected across all samples and removing those present in the blanks, duplicate features were removed by eliminating values that had the same exact mass (to the third decimal value) and were within an acceptable retention time window (+/- 30 s) of each other. Filtering of the features was achieved by a 2-step procedure, with in-house scripts developed in python: (1) All detected features were filtered for those that had MS/MS spectra appended and (2) which were not present in any sample blanks. The DDA fashion in which the data was acquired, allows for this filtering to be possible without losing any of the most abundant compounds and allows for plausible chemical identification of the features.

Acknowledgements

We gratefully acknowledge financial support from the EPSRC (Grant Nos EP/P00153X/1, EP/J015156/1, EP/K021966/1, EP/K038885/1, EP/L015668/1, EP/L023652/1), BBSRC (Grant No. BB/M011267/1), ERC (project 670467 SMART-POM), and the The John Templeton Foundation Grant ID 60625 and Grant ID 61184. We thank Dario Caramelli, Dr. Davide Angelone and Graham Keenan for their help in writing the python scripts.(School of Chemistry, University of Glasgow)

Keywords: Systems Chemistry • Prebiotic Chemistry • Complex Mixtures

References

[1] L. Cronin, S. I. Walker, Science (80-.). **2016**, 352, 1174 LP – 1175.

[2] N. Guttenberg, N. Virgo, K. Chandru, C. Scharf, I. Mamajanov, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **2017**, 375.

[3] E. T. Parker, H. J. Cleaves, M. P. Callahan, J. P. Dworkin, D. P. Glavin, A. Lazcano, J. L. Bada, *Orig. Life Evol. Biosph.* **2011**, *41*, 201–212.

[4] R. Saladino, G. Botta, S. Pino, G. Costanzo, E. Di Mauro, *Biochimie* **2012**, DOI 10.1016/j.biochi.2012.02.018.

[5] M. W. Powner, B. Gerland, J. D. Sutherland, *Nature* **2009**, *459*, 239–242.

[6] S. C. Kim, D. K. O'Flaherty, L. Zhou, V. S. Lelyveld, J. W. Szostak, Proc. Natl. Acad. Sci. USA **2018**

[7] J. D. Sutherland, *Angew. Chemie - Int. Ed.* **2015**, *54*, 104–121.

[8] D. Ross, D. Deamer, S. Lower, **2019**, 19, 1–5.

[9] S. A. Benner, H. J. Kim, M. a. Carrigan, *Acc. Chem. Res.* **2012**, *45*, 2025–2034.

[10] N. V. Hud, *Synlett* **2017**, *28*, 36–55.

[11] K. Ruiz-Mirazo, J. Peretó, A. Moreno, *Orig. Life Evol. Biosph.* **2004**, *34*, 323–346.

[12] E. Wollrab, S. Scherer, F. Aubriet, V. Carré, T. Carlomagno, L. Codutti, A. Ott, *Orig. Life Evol. Biosph.* **2016**, *46*, 149–169.

[13] S. Scherer, E. Wollrab, L. Codutti, T. Carlomagno, S. G. da Costa, A. Volkmer, A. Bronja, O. J. Schmitz, A. Ott, *Orig. Life Evol. Biosph.* **2016**, 1–23.

[14] M. Ferus, A. Knížek, S. Civiš, *Proc. Natl. Acad. Sci.* **2015**, *112*, 7109–7110.

[15] W. Martin, M. J. Russell, *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2003**, *358*, 59–83; discussion 83–5.

[16] M. Ruiz-Bermejo, C. Menor-Salván, E. Mateo-Martí, S. Osuna-Esteban, J. Á. Martín-Gago, S. Veintemillas-Verdaguer, *Icarus* **2008**, 198, 232–241.

[17] D. A. Baum, K. Vetsigian, *Orig. Life Evol. Biosph.* **2016**, 1–17.

[18] L. Boiteau, R. Pascal, *Orig. Life Evol. Biosph.* **2011**, *41*, 23–33.

[19] R. M. Hazen, D. A. Sverjensky, *Cold Spring Harb. Perspect. Biol.* **2010**

[20] A. J. Surman,, M. R. Garcia, Y. M. Abul-Haija, G. Cooper, P. S. Gromski, R. Turk-MacLeod, M. Mullin, C. Mathis, S. Walker, L. Cronin, **2018**,

[21] A. Ricardo, Science (80-.). **2004**, 303, 196–196.

[22] R. Saladino, J. M. G. Ruiz, E. Di Mauro, Chem. - A Eur. J. **2018.**

[23] E. Hung, *ThermoFisher Scientific*, **2017**.

[24] D. A. Butlerow, Berichte der Dtsch. Chem. Gesellschaft **1860**, 8, 398–416.

[25] H. J. Cleaves, Precambrian Res. **2008**, 164, 111–118.

[26] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, et al., *Nat. Biotechnol.* **2012**, *30*, 918–

Taming Combinatorial Explosion of the Formose Reaction via Recursion within Mineral Environments.

Stephanie Colón-Santos, Geoffrey J. T. Cooper and Leroy Cronin*

*WestCHEM, School of Chemistry, The University of Glasgow, University Avenue, Glasgow G12 8QQ, UK, *Correspondence: Lee.Cronin@glasgow.ac.uk*

Table of Contents

Experimental Methods

Reagents

Formaldehyde (ASG reagent, 37% wt. in H_2O), Glycoaldehyde (97%), Formamide (Reagent Plus®, >99.0 (GC)), Calcium Carbonate (purity >96%) Formic Acid (reagent grade, >95%) and were purchased from Sigma Aldrich and Calcium Hydroxide (purity > 96.0%) were purchased from Fluka Analytical. Analytical solvents (Water and Acetonitrile, HPLC-MS grade) and Ammonium Acetate (Ambion® Molecular Biological Grade (5M), >98%) were purchased from ThermoFisher Scientific UK. Analytical standards of Hexamethylenetetramine (HMT), Ribose, Adenine, Guanine, Thymine, Cytosine, Uracil (Molecular Biological Grade, >98%), Adenosine and Thymidine were purchased from Tokyo Chemical Industry UK.

Minerals

Goethite, α-FeO(OH), Montmorillonite, $(Na, Ca)_{0.33}(Al, Mg)_2(Si_4O_{10})$ and Hydroxyapatite, $Ca_5(OH)(PO_4)_3$ were purchased from Sigma-Aldrich. Chalcopyrite, $CuFeS_2$ was purchased from Alpha-Aesar by Thermo Fisher Scientific. Ulexite, $NaCaB_5O_6(OH)_6{\cdot}5H_2O$, Zoesite, $Ca_2Al_3(SiO_2)_3(OH)$ and Quartz, SiO_2 were purchased from Richard Taylor Minerals, a private collection from the United Kingdom.

Experimental procedure

Formaldehyde (0.5 mL), Glycolaldehyde (0.0126 g), Water (2.25 mL), Formamide (2.25mL) and Calcium Hydroxide (0.0705g) was carried out on seven (7) different mineral surfaces (plus control) in 22mL borosilicate glass vials. It was stirred at 1200rpm with a magnetic stirrer and heated at 50°C, for 48 hours. Then, about 70% of the reaction volume (supernatant) was removed for analysis.

The remaining fraction was used to seed the next reaction. Topping up with the same concentration of starting materials, but conserving the total reaction volume; we repeated the process three times.

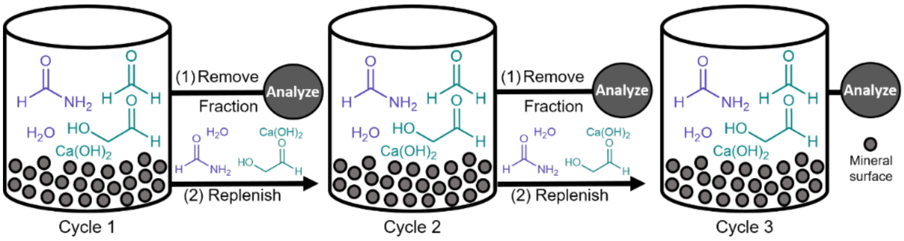


Figure S1. *Recursive cycles:* After each reaction, the supernatant is removed for analysis and a small fraction is left in the reaction vessel and used to seed a next reaction.

Analytical methods

Sample preparation

The removed fraction was allowed to cool to room temperature, then a 100μL aliquot was taken for each analysis. Followed by removal of excess cations in solution (e.g. Ca^{2+}) with Amberlite™ Ion-exchange resin, before the supernatant was diluted 1 in a 100 with MS grade water. Finally, the solution was filtrated with a syringe filter (0.22μm cut-off)] and placed in an HPLC sample vial, before analysis.

UPLC-MS/MS analysis

Ultra-Performance liquid chromatography and tandem mass spectrometry (UPLC-MS/MS) analysis was performed with a Thermo Vanquish Ultra-performance liquid chromatography system coupled to a Thermo Orbitrap Fusion Mass-Spectrometer. Samples were injected directly (no splitting) in 10 µl aliquots and chromatographic separation was achieved with a ZIC-HILIC C18 (4.6 × 150 mm, 2.7 µm) column, eluted in a linear gradient mixture of solvents A (water w/20 mM Ammonium Acetate, pH = 5) and B (100% acetonitrile w/0.1% v/v formic acid) over 25 min as follows: 0 min, 100% A; 4min, 100% A; 19 min, 100% B; 23 min, 100% A; 25min, 100% A; in a method adjusted from H. Idborg, *et.al.* [1] The column was maintained at 30 °C and the MS spectra was collected for 30 minutes in positive mode over a scan range of 50–500 m/z. Ion transfer tube was set to 275 °C, RF lens 60%, and acquisition was performed in a Data-dependent (DDA) manner.

The Data-Dependent Acquisition (DDA) was performed by prioritizing the top most intense fragments in a 3 second window with an intensity threshold of 5.0E4 and dynamic exclusion, after one time for 15 seconds (in order to avoid the selection of the same fragments), using the ion trap isolation with a HCD collision energy of 35 eV and a resolution 15000.

Feature Generation

The raw data was extracted from its vendor format (.raw) with MSConvert from ProteoWizard [2], into an .mzML format before introducing it to Python. The package PymzML [3] was used to extract MS^1 values, Retention Time (RT), Intensity (in counts) and MS^2 values (with their corresponding RT and Intensity). Detected MS^1 values where then filtered by selecting those which were taken for MS^2 fragmentation by the DDA method, which we found were representative of the most important features within the complex mixture.

The features generated are based on unique Retention Time (RT), Exact Mass (m/z) and MS^2 fragments; which (as previously mentioned) we achieved by filtering for all fragments which were selected in a DDA fashion for MS^2 (Top most intense). Each one of the points (**Figure S2**) represents a compound within the product mixture; which gives a representation of the reaction products, even if we cannot identify each one of the compounds.

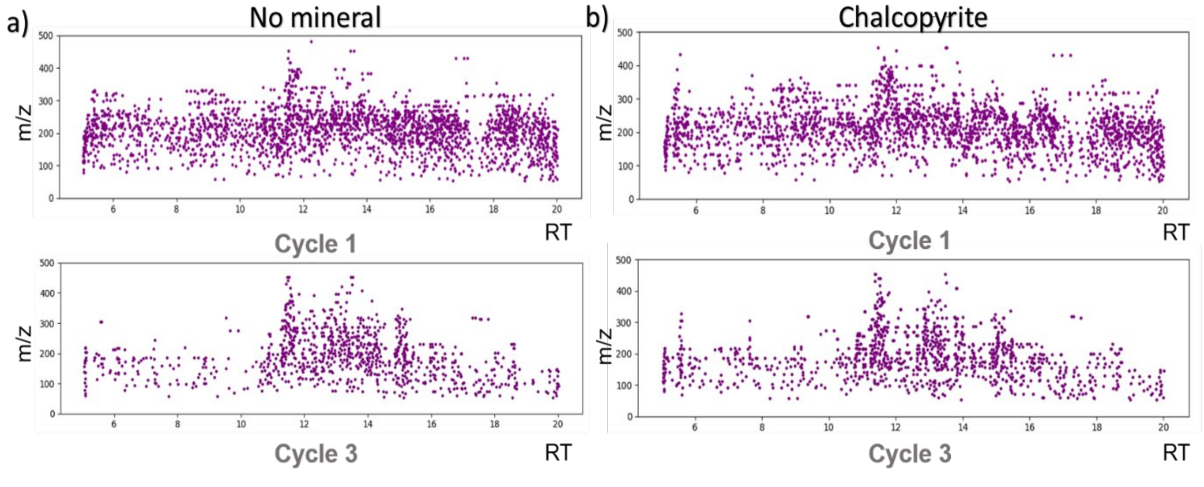


Figure S2. The feature generation based on unique retention time (RT) and *m/z* combinations for Cycle 1 and Cycle 3, in the reaction control (a) and in the presence of the mineral Chalcopyrite **(b)**. Each point represents a compound within the product distribution ensemble.

Duplicate (MS^1) values where filtered further (**Figure S3**), by eliminating values that had same exact mass to the third decimal value, besides being within an acceptable retention time window (+/- 30 s) of each other. Furthermore, the number of MS^2 fragments obtained for each feature (MS^1) was calculated, as a generic way to access the overall complexity of the molecules within the complex mixture.

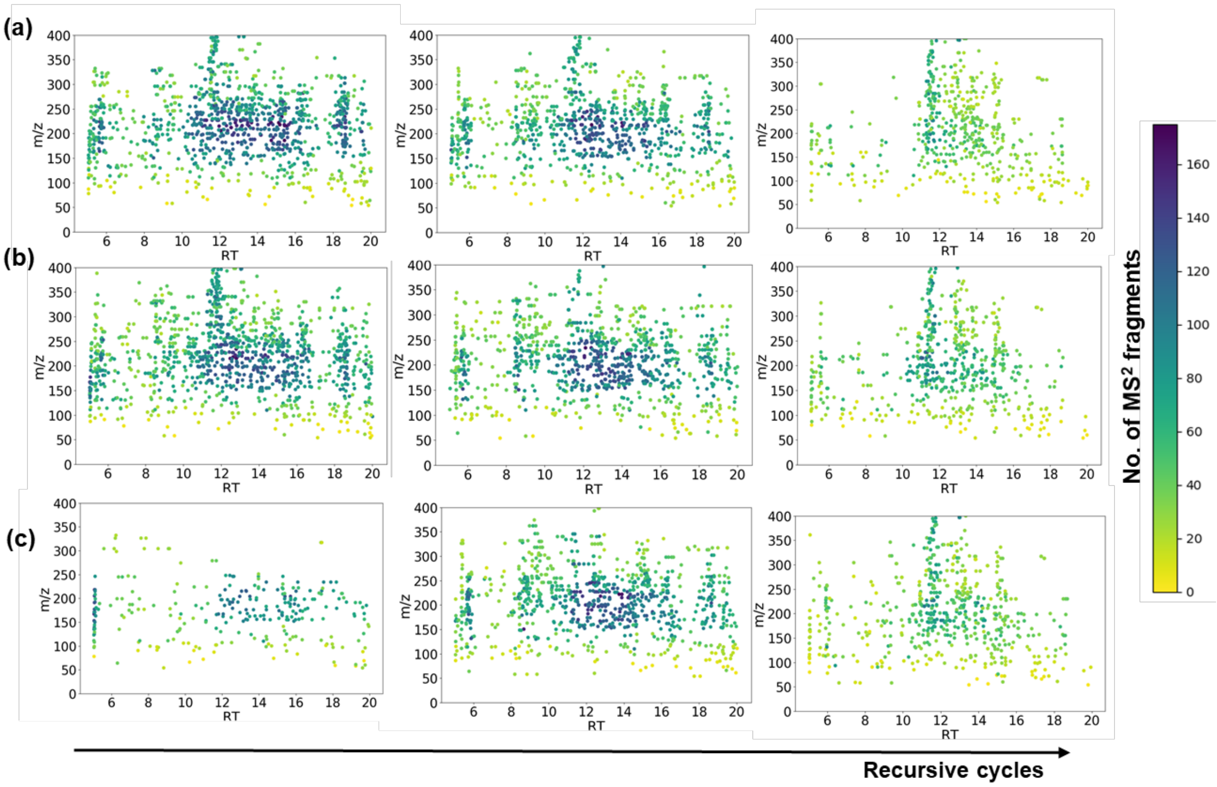


Figure S3. Filtered features for Cycle 1, Cycle 2 and Cycle 3 in the **(a)** control reaction and with **(b)** Chalcopyrite and **(c)** Goethite as a mineral surface. Differences across the profile of the products and number of MS^2 fragments of each feature can be observed.

Also, as seen in **Figure S4**, the changes in the product distribution are consistent with the features generated in CompoundDiscoverer™ (Thermo Scientific) [4], a conventionally used software to process untargeted mass-spectral data, which also enabled the extraction of ion chromatograms (EIC's) in a targeted fashion. The number of features generated by the software's processing method where far less than the ones generated by our in-house feature extraction method, but preserved the same trend (as **Figure S3**). This comparison provided a good validation of the bespoke feature generator and our 'omics' based approach to complex mixture analysis.

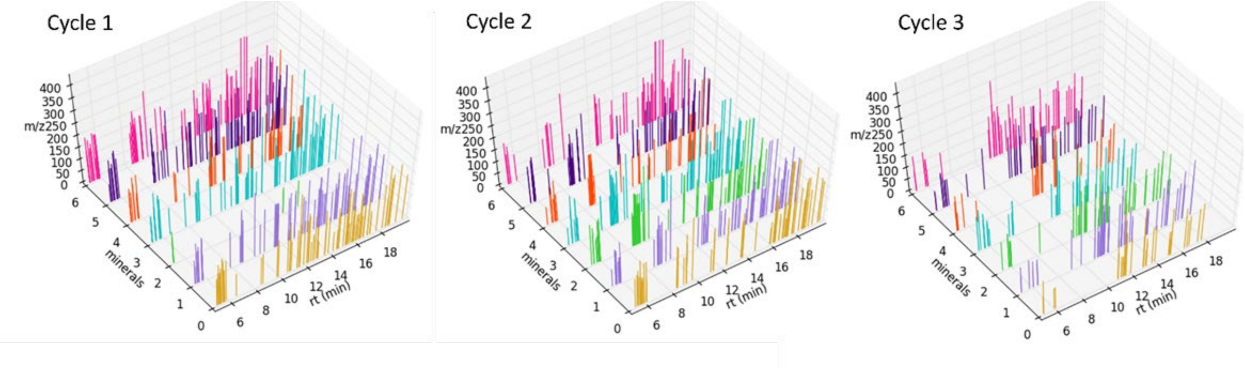


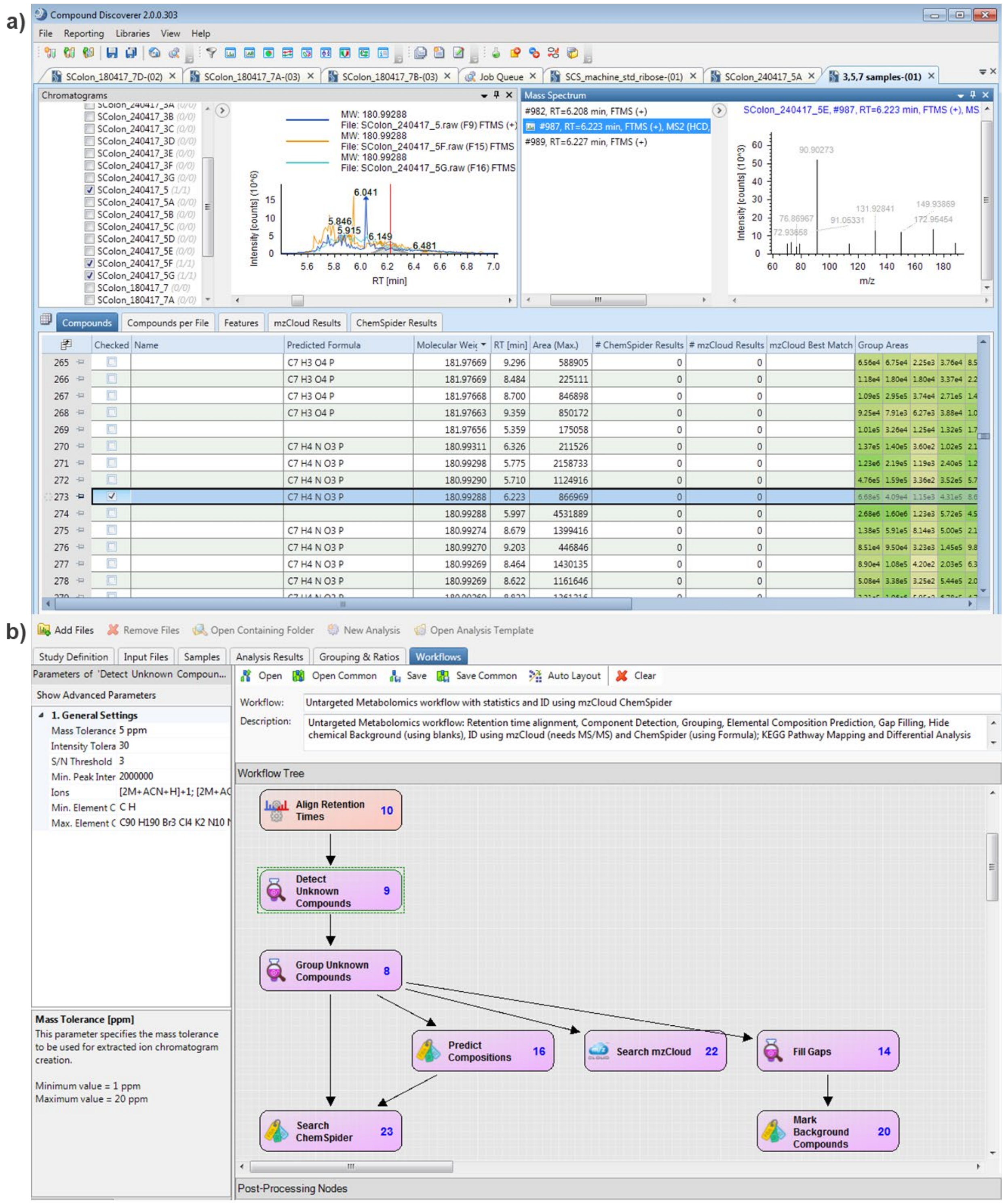
Figure S3. *Multi-colour plots:* Differences in the feature distribution are displayed across mineral surfaces (1-6) and the reaction with no mineral (0) and the number of detected features reduces over the recursive cycles, in features extracted by the CompoundDiscoverer™ software suite.

Molecular weight distribution of features

The filtered features were manually grouped into 50m/z bins, in the range of 50 to 400 m/z, as seen in **Figure 2**. The number of features detected for each bin are displayed in a heatmap, generated in matplotlib through Python. The heatmap produces a unique pattern for each reaction, in which the number of features decreases over recursive cycles consistently, in the absence *and* presence of a mineral surface.

Compound identification and validation

Identification of the compounds in the reaction mixture was performed by Compound Discoverer 2.0 *(3)* by matching the exact mass and the resulting MS^2 spectra with the all the available databases, through MZcloud® search (**Figure S5**). This was further validated with pure standards, where a match in retention time, exact mass and a robust correlation with the MS/MS mass-spectral pattern was used to confirm the identity of the compounds. The validation preformed done manually through ThermoScientific™ Mass Frontier™ spectral interpretation software (**Figure S6**, **S7** and **S8**) and CompoundDiscoverer™ **(Figure S9 and S10)**.



	Checked	Name	Predicted Formula	Molecular Weig	RT [min]	Area (Max.)	# ChemSpider Results	# mzCloud Results	mzCloud Best Match	Group Areas				
265	☐		C7 H3 O4 P	181.97669	9.296	588905	0	0		6.56e4	6.75e4	2.25e3	3.76e4	8.5
266	☐		C7 H3 O4 P	181.97669	8.484	225111	0	0		1.18e4	1.80e4	1.80e4	3.37e4	2.2
267	☐		C7 H3 O4 P	181.97668	8.700	846898	0	0		1.09e5	2.95e5	3.74e4	2.71e5	1.4
268	☐		C7 H3 O4 P	181.97663	9.359	850172	0	0		9.25e4	7.91e3	6.27e3	3.88e4	1.0
269	☐			181.97656	5.359	175058	0	0		1.01e5	3.26e4	1.25e4	1.32e5	1.7
270	☐		C7 H4 N O3 P	180.99311	6.326	211526	0	0		1.37e5	1.40e5	3.60e2	1.02e5	2.1
271	☐		C7 H4 N O3 P	180.99298	5.775	2158733	0	0		1.23e6	2.19e5	1.19e3	2.40e5	1.2
272	☐		C7 H4 N O3 P	180.99290	5.710	1124916	0	0		4.76e5	1.59e5	3.36e2	3.52e5	5.7
273	☑		C7 H4 N O3 P	180.99288	6.223	866969	0	0		6.68e5	4.09e4	1.15e3	4.31e5	8.6
274	☐			180.99288	5.997	4531889	0	0		2.68e6	1.60e6	1.23e3	5.72e5	4.5
275	☐		C7 H4 N O3 P	180.99274	8.679	1399416	0	0		1.38e5	5.91e5	8.14e3	5.00e5	2.1
276	☐		C7 H4 N O3 P	180.99270	9.203	446846	0	0		8.51e4	9.50e4	3.23e3	1.45e5	9.8
277	☐		C7 H4 N O3 P	180.99269	8.464	1430135	0	0		8.90e4	1.08e5	4.20e2	2.03e5	6.3
278	☐		C7 H4 N O3 P	180.99269	8.622	1161646	0	0		5.08e4	3.38e5	3.25e2	5.44e5	2.0

Parameters of 'Detect Unknown Compoun...

Show Advanced Parameters

1. General Settings

Mass Tolerance	5 ppm
Intensity Tolera	30
S/N Threshold	3
Min. Peak Inter	2000000
Ions	[2M+ACN+H]+1; [2M+AC
Min. Element C	C H
Max. Element C	C90 H190 Br3 Cl4 K2 N10 N

Mass Tolerance [ppm]
This parameter specifies the mass tolerance to be used for extracted ion chromatogram creation.

Minimum value = 1 ppm
Maximum value = 20 ppm

Workflow: Untargeted Metabolomics workflow with statistics and ID using mzCloud ChemSpider

Description: Untargeted Metabolomics workflow: Retention time alignment, Component Detection, Grouping, Elemental Composition Prediction, Gap Filling, Hide chemical Background (using blanks), ID using mzCloud (needs MS/MS) and ChemSpider (using Formula); KEGG Pathway Mapping and Differential Analysis

Figure S5. General overview of the analytical workflow in *CompoundDiscoverer 2.0™* **(a)** Automated extraction of features and their MS/MS fragmentation. Also, Extracted Ion Chromatograms (EIC's) are generated for all features in each sample. **(b)** Integrated processing workflow does adduct calculations, predicts compositions and conducts a search through the MZcloud® database.

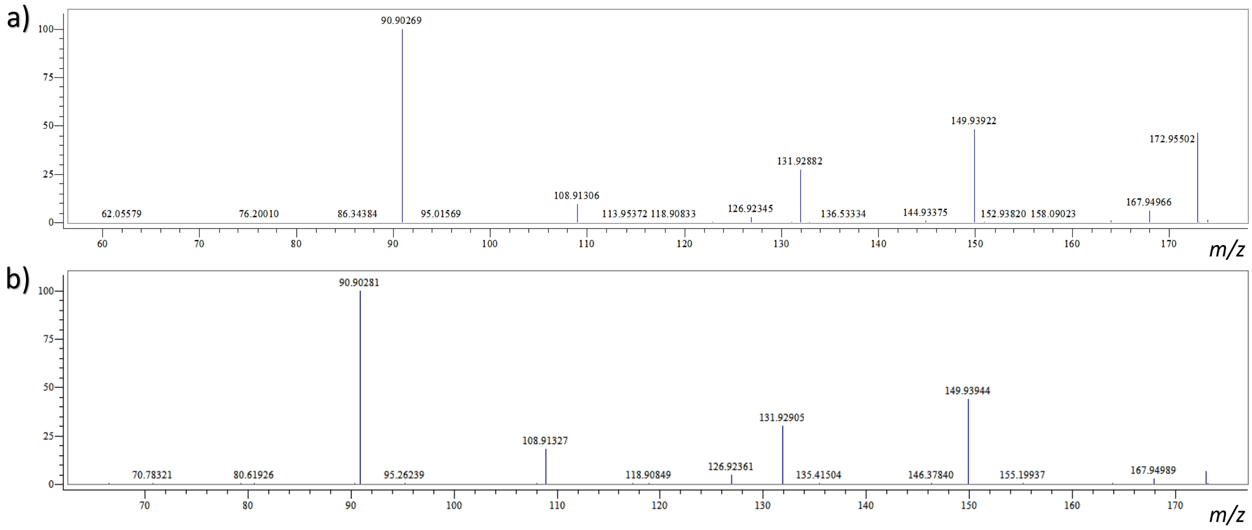


Figure S6. UPLC-MS/MS spectrum for Ribose (*m/z*: 172.96, Adduct: [M+Na]) in (a) a pure standard and (b) real sample (Chalcopyrite, Cycle 3).

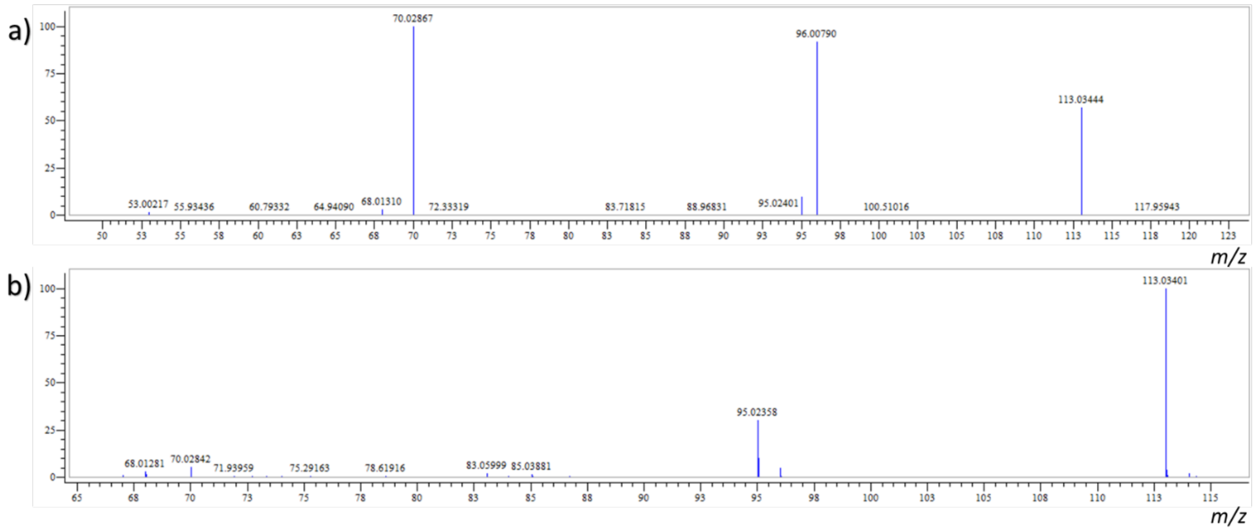


Figure S7 UPLC-MS/MS spectrum for Uracil (*m/z:* 113.03, Adduct: [M+H]) in **(a)** a pure standard and **(b)** a real sample, (Chalcopyrite, Cycle 3).

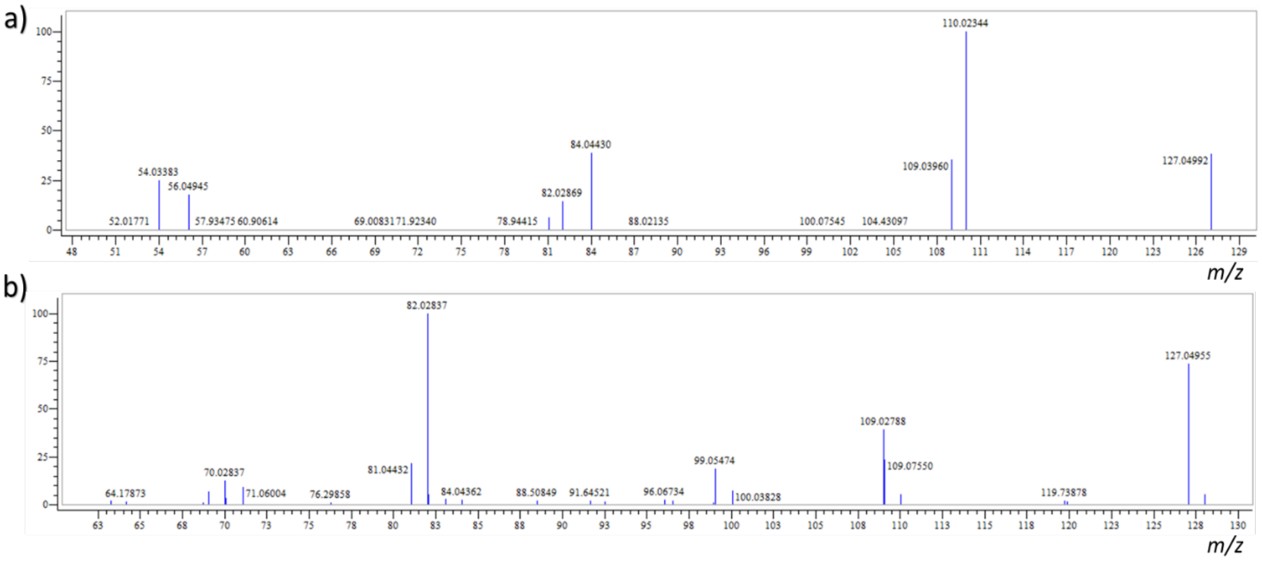


Figure S8 UPLC-MS/MS spectrum for Thymine (*m/z:* 127.05, Adduct: [M+H]) in **(a)** a pure standard and **(b)** a real sample, (Chalcopyrite, Cycle 3).

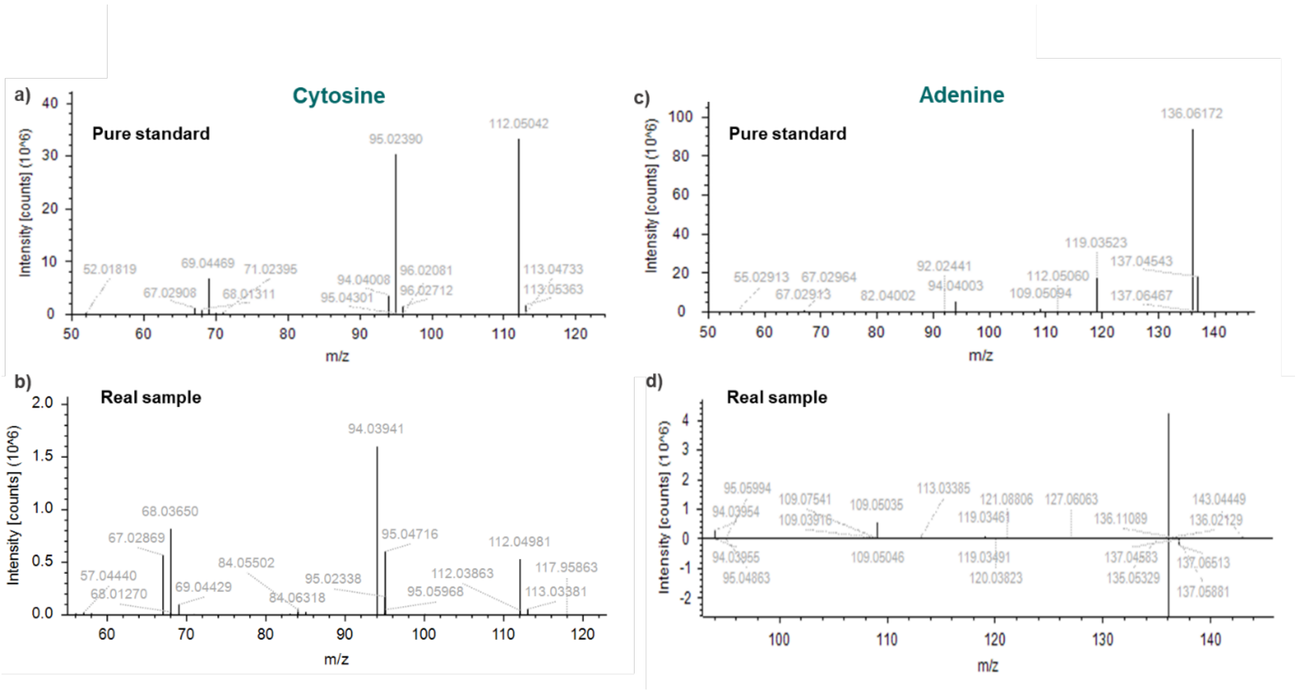


Figure S9 UPLC-MS/MS spectrum for Cytosine (*m/z*: 112.05, Adduct: [M+H]) in **(a)** a pure standard and **(b)** a real sample, (Control, Cycle 3); and Adenine (*m/z*: 136.06, Adduct: [M+H]) in **(c)** a pure standard and **(b)** a real sample, (Chalcopyrite, Cycle 3)

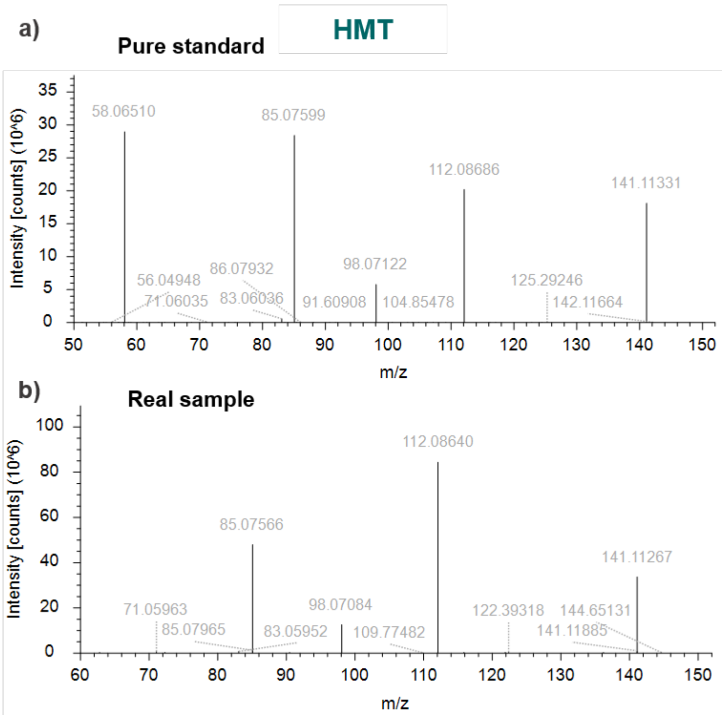


Figure S10 UPLC-MS/MS spectrum for Hexamethylenetetramine, HMT (*m/z*: 141.11, Adduct: [M+H]) in **(a)** a pure standard and **(b)** a real sample, (Chalcopyrite, Cycle 3)

Detection of nucleosides (traces): Validation by external standards and MS/MS

Traces of Adenosine and Thymidine nucleosides where found in the last cycle (3) of most samples (including the control reaction). The presence of Ribose and several nucleobases in Cycle 1 and 2, lead us to believe there was a possibility for nucleoside formation in Cycle 3. To address this, we looked for the corresponding nucleosides of Adenine and Thymine, in Cycle 3. **Figure S11** illustrates the cross validation with an external standard of Adenosine, by *(***1, 2***)* having a chromatographic match in retention time (+/-40s), (**1a, 2a**) the same exact mass and (**1b, 2b**) a matching MS/MS (MS^2) fragmental pattern. Furthermore, by running the pure standards through the same chromatographic method, we found that preferred adducts for Thymidine were not necessarily the $[M+H]^+$, but rather the charged (z = 2, for m/z) and sodium adducts predominated (as seen in **Figure S12** and **Figure S13**). The exact isomer of the nucleosides was not assessed, since the pure standards used for validation were not differentiated by their isomeric position. However, we believe this to be satisfactory evidence towards the presence of nucleosides in the product distribution ensemble; *particularly,* within a mixture of this complexity.

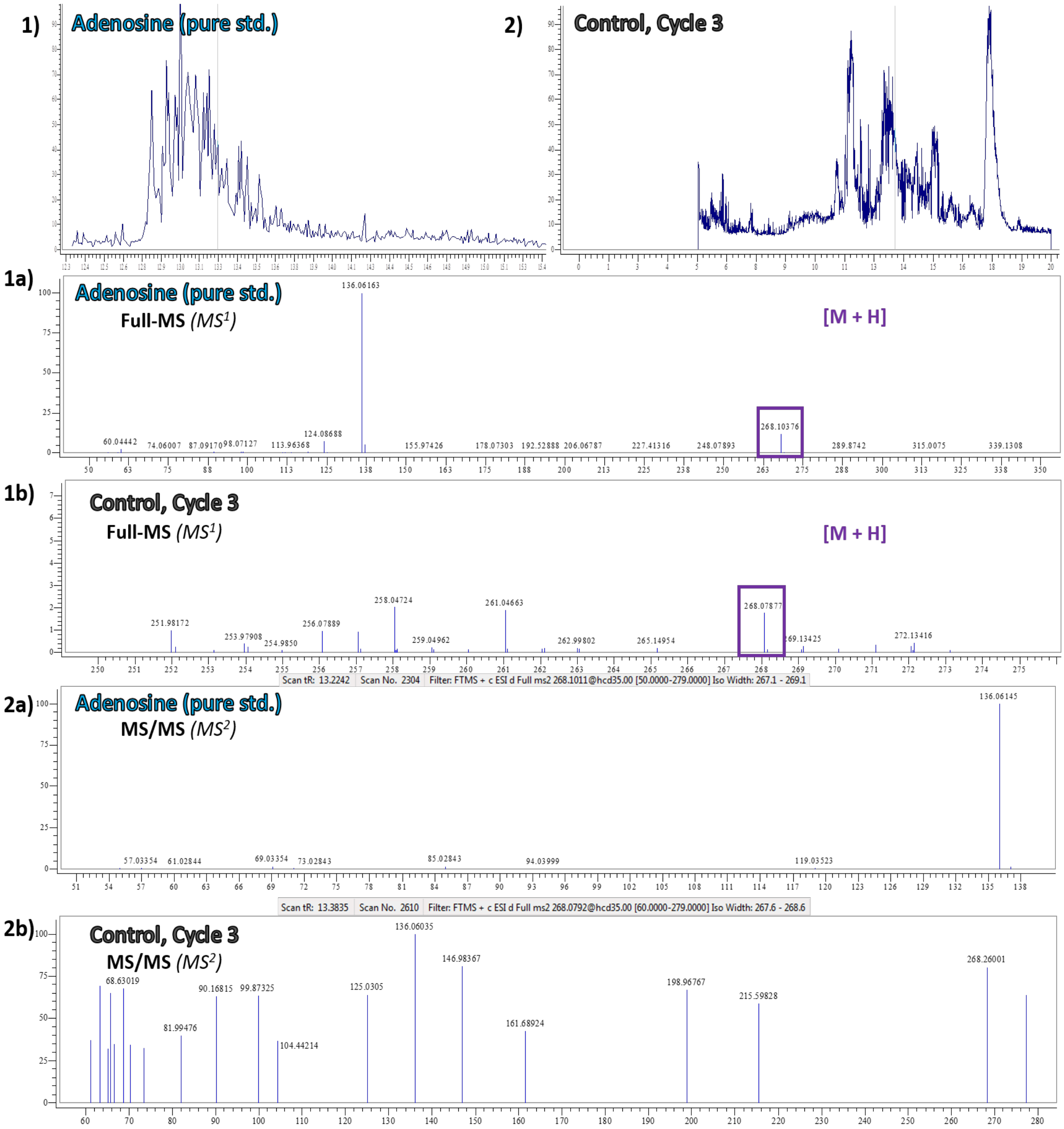


Figure S11. UPLC-MS/MS analysis and comparison of (**1-1b**) a pure standard of Adenosine (m/z: 268.1), Adduct: [M+H]) with (**2-2b**) a real sample (Chalcopyrite, Cycle 3).

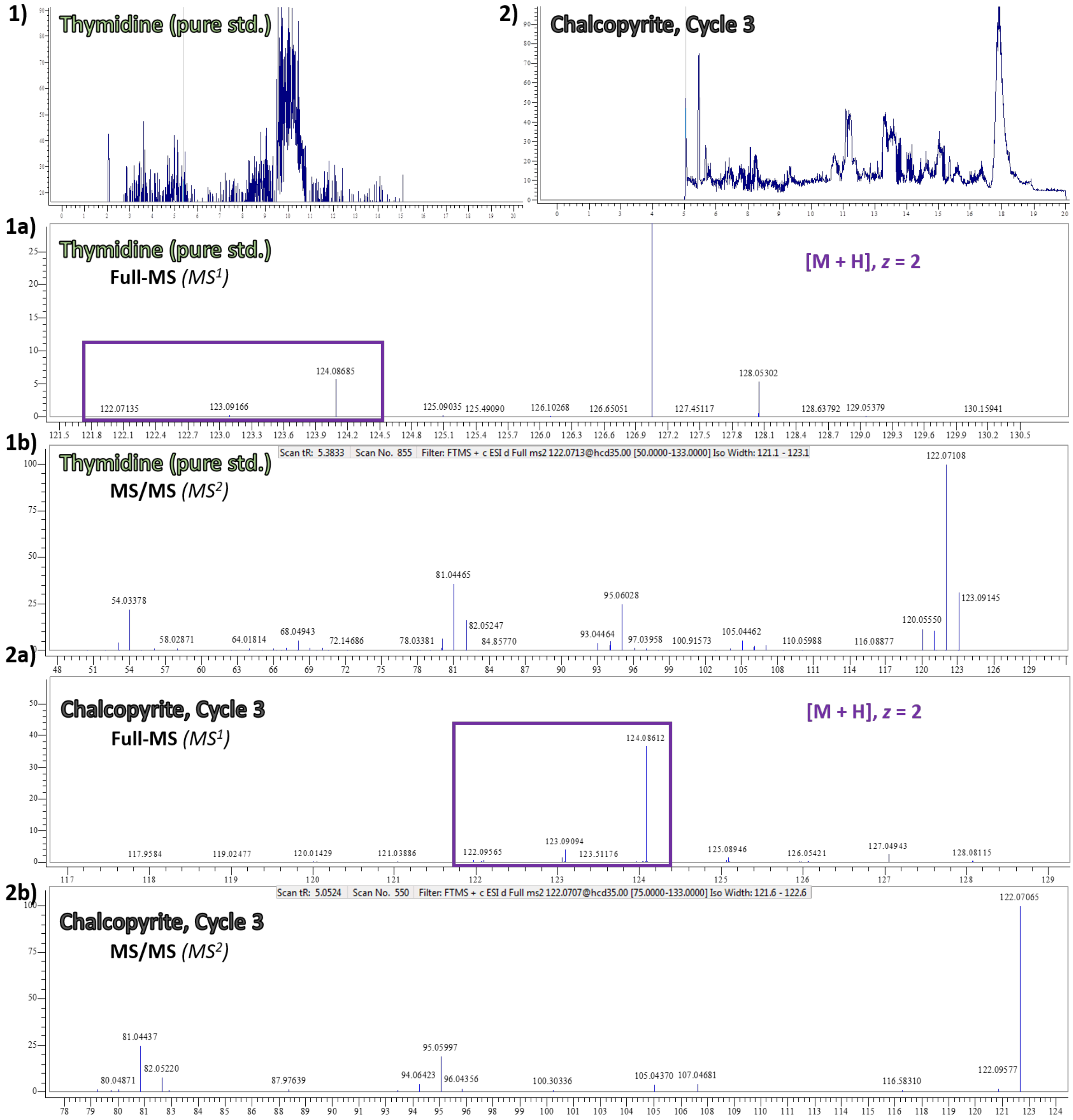


Figure S12. UPLC-MS/MS analysis and comparison of (**1-1b**) a pure standard of Thymidine (m/z: 122.07), Adduct: [M+H] z = 2, with (2-2b) a real sample (Chalcopyrite, Cycle 3).

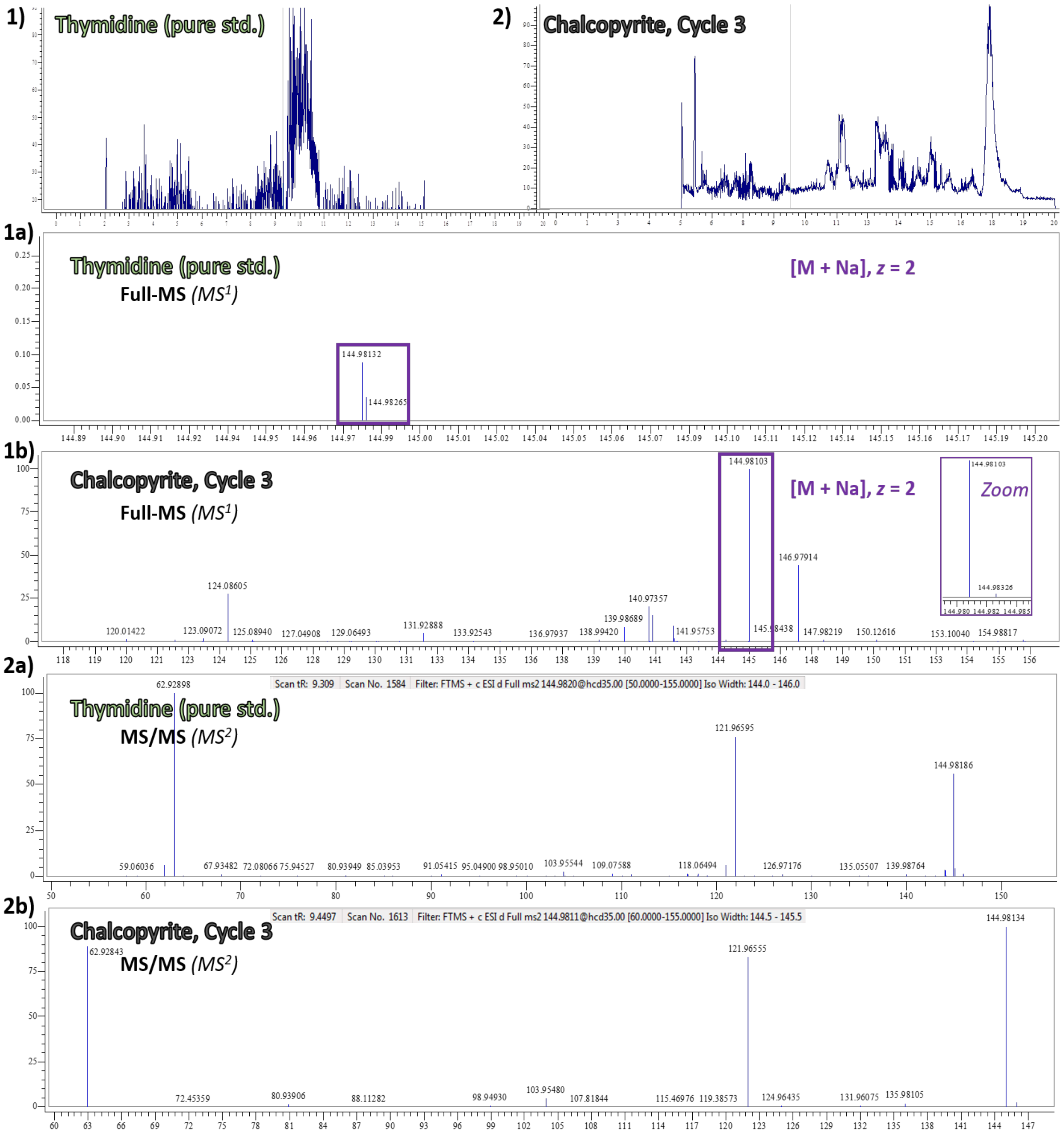


Figure S13. UPLC-MS/MS analysis and comparison of (**1-1b**) a pure standard of Thymidine (m/z: 122.07), Adduct: [M+H] z = 2, with (**2-2b**) a real sample (Chalcopyrite, Cycle 3).

Extracted Ion Chromatograms (EIC's) for selected compounds on different mineral surfaces, across recursive cycles.

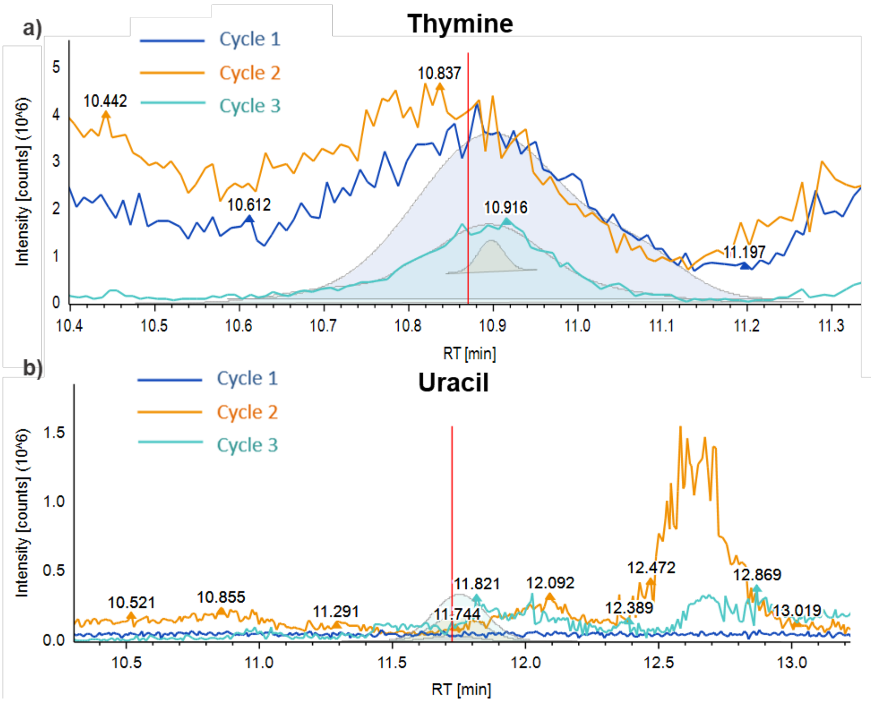


Figure S14. Extracted Ion Chromatogram for **(a)** Thymine (Molecular Weight: 127.05, Adduct: [M+H]), in Cycle 3 with Chalcopyrite; and Uracil (*m/z:* 113.03, Adduct: [M+H]), in Cycle 3 with Goethite.

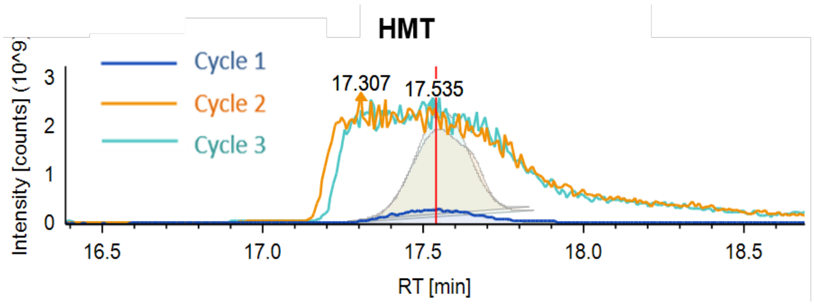


Figure S15. Extracted Ion Chromatogram for Hexamethylenetetramine, HMT (m/z: 141.11, Adduct: [M+H]) in (a) a pure standard and (b) a real sample, (Goethite, Cycle 3)

References

[1] H. Idborg, L. Zamani, P. Edlund, I. Schuppe-Koistinen, S. P. Jacobsson, Metabolic fingerprinting of rat urine by LC/MS: Part 1. Analysis by hydrophilic interaction liquid chromatography–electrospray ionization mass spectrometry, *Journal of Chromatography B*, **2005**, Volume 828, Issues 1–2, Pages 9-13.

[2] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, et al., Nat. Biotechnol. **2012**, 30, 918–920

[3] T. Bald, J. Barth, A. Niehues, M. Specht, M. Hipple, and C. Fufezan PymzML - Python module for high throughput bioinformatics on mass spectrometry data, *Bioinformatics*, **2012**

[4] E. Hung, ThermoFisher Scientific, **2017**.