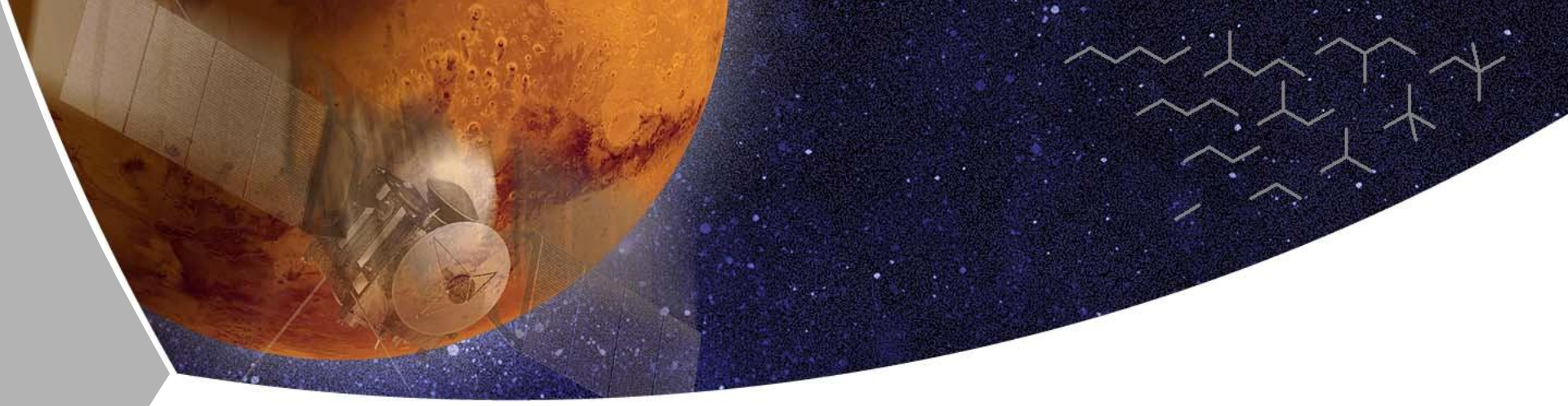


**Markus Meringer**

# Computational Approaches towards Life Detection by Mass Spectrometry

**International Workshop on Life Detection Technology: For Mars, Encheladus and Beyond**

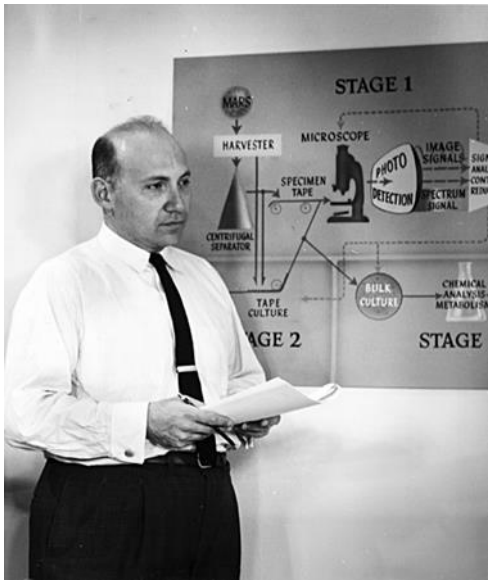
**Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo, Japan**

**October 5-6, 2017**

Deutsches Zentrum
für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

# Outline

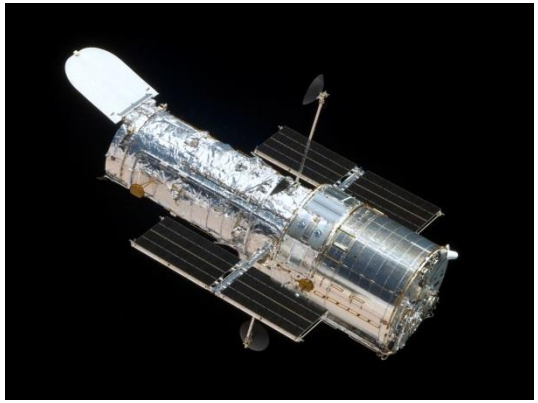
- The Past
  - DENDRAL
  - COSAC
- The Present
  - from mass to formula
  - from spectrum to structure
- The Future
  - from structure to life
  - from spectrum to life
  - from masses to life
- Conclusion




## The DENDRAL Project

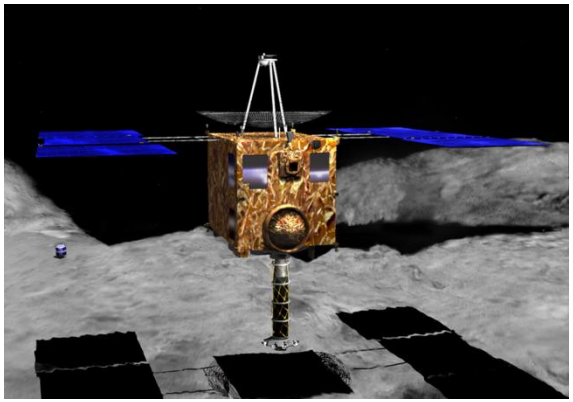
- driven by Joshua Lederberg
- initiated in the mid 1960's
- aim: identify unknown organic compounds by analyzing their mass spectra computationally
- perspective: processing of MS recorded on mars missions

# Detecting Life in Space: Types of Measurements and Missions

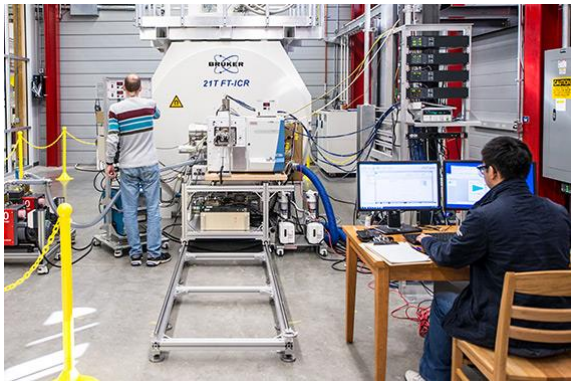


Remote sensing

In-situ measurements



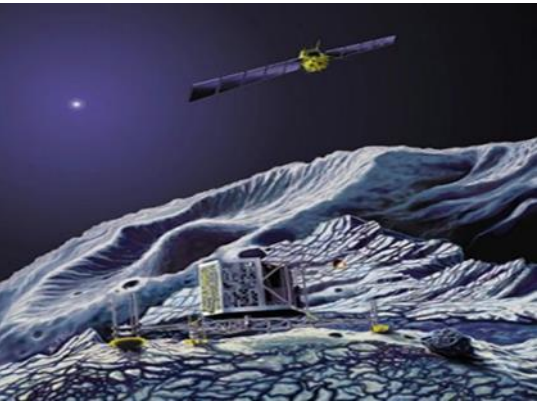
Sample return missions

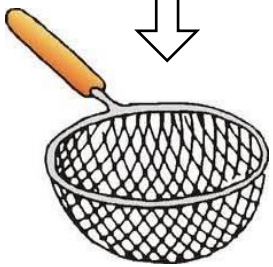



Mass spectrometry

# Processing of the COSAC Data

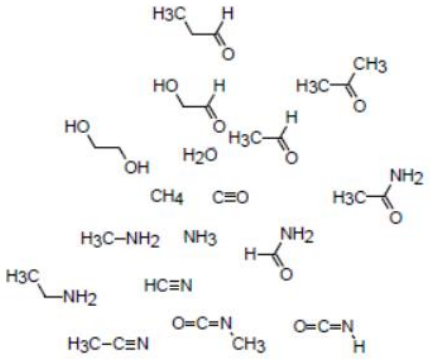
COSAC measurements on ROSETTA mission



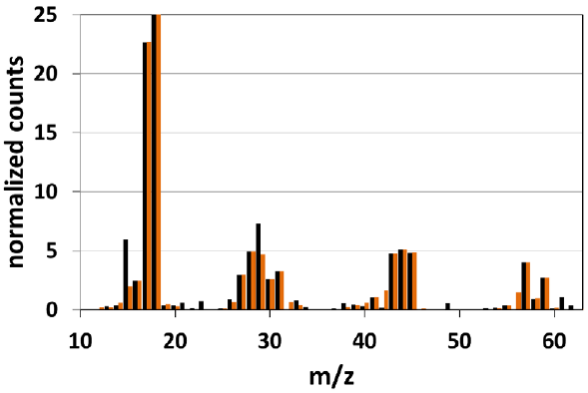


filter

NIST DB



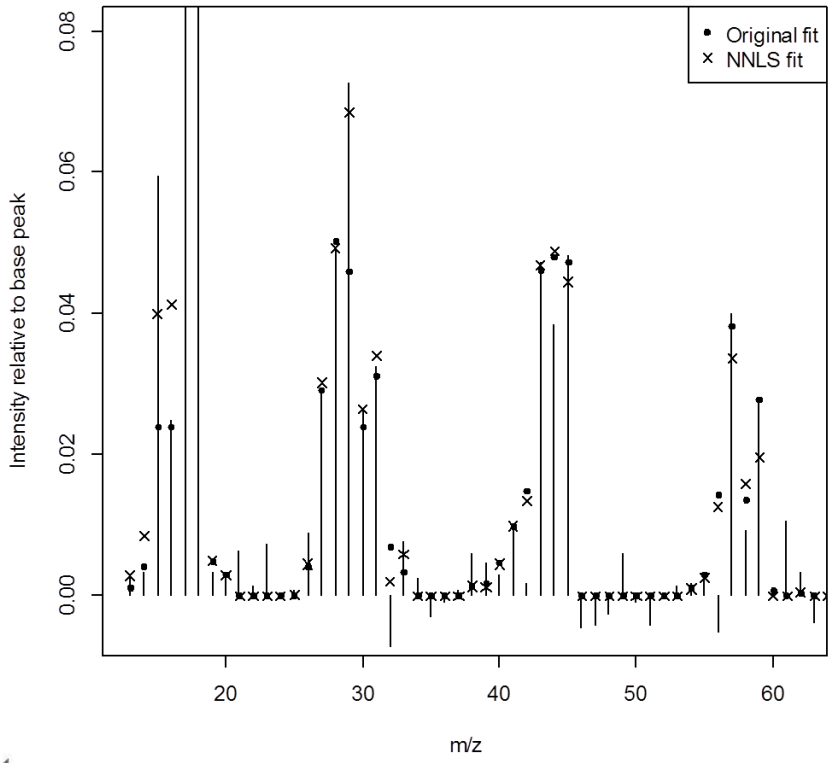
fit

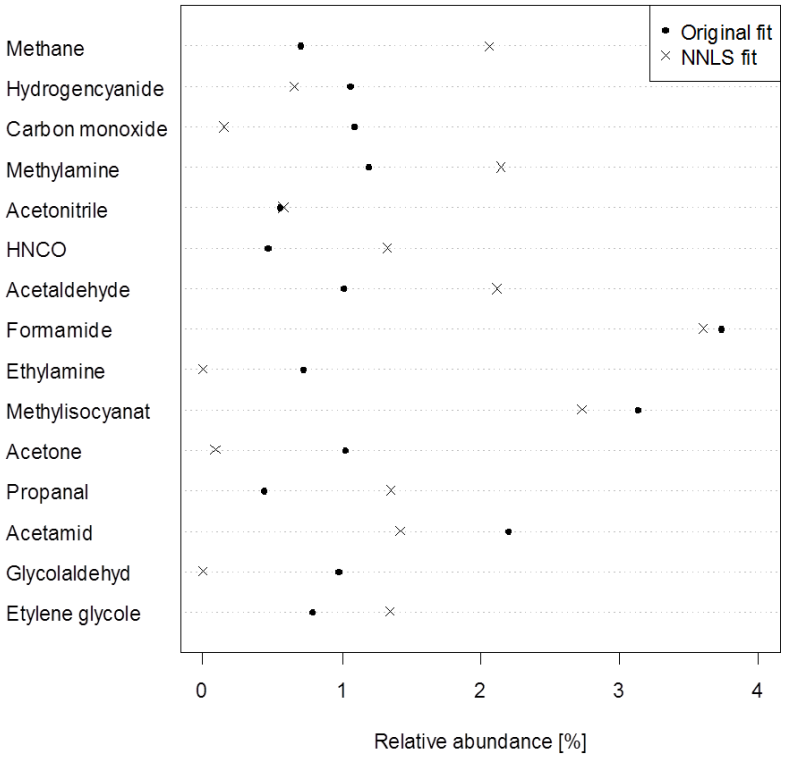



F. Goesmann et al, Organic compounds on comet 67P/Churyumov-Gerasimenko revealed by COSAC mass spectrometry, Sciene 349 (2015)
J. H. Bredehöft et al, The organics on the nucleus of 67P/C-G and how they might have gotten there, XVIIIth International Conference on the Origin of Life (2017)

# Revisiting the COSAC Data

- use non-negative linear least squares (NNLS) fit
- unexplained part: 13.5 % of TIC instead of 14.4 %
- similar results…



# From Mass to Formula

- Higher mass resolution enables better assignment of molecular formulas
- **Which mass resolution is required to enable unambiguous assignment of molecular formulas?***

Procedure:

1. Generate all molecular formulas up to a certain nominal mass (151)
2. Sort formulas by increasing exact masses: $m_1 \leq m_2 \leq m_3 \leq \ \ldots$
3. Compute for each mass $m_i$ the minimum distance to the neighbored masses: delta($m_i$) = min($m_{i+1}$-$m_i$, $m_i$-$m_{i-1}$)
4. Plot $m_i$/delta($m_i$) vs. $m_i$

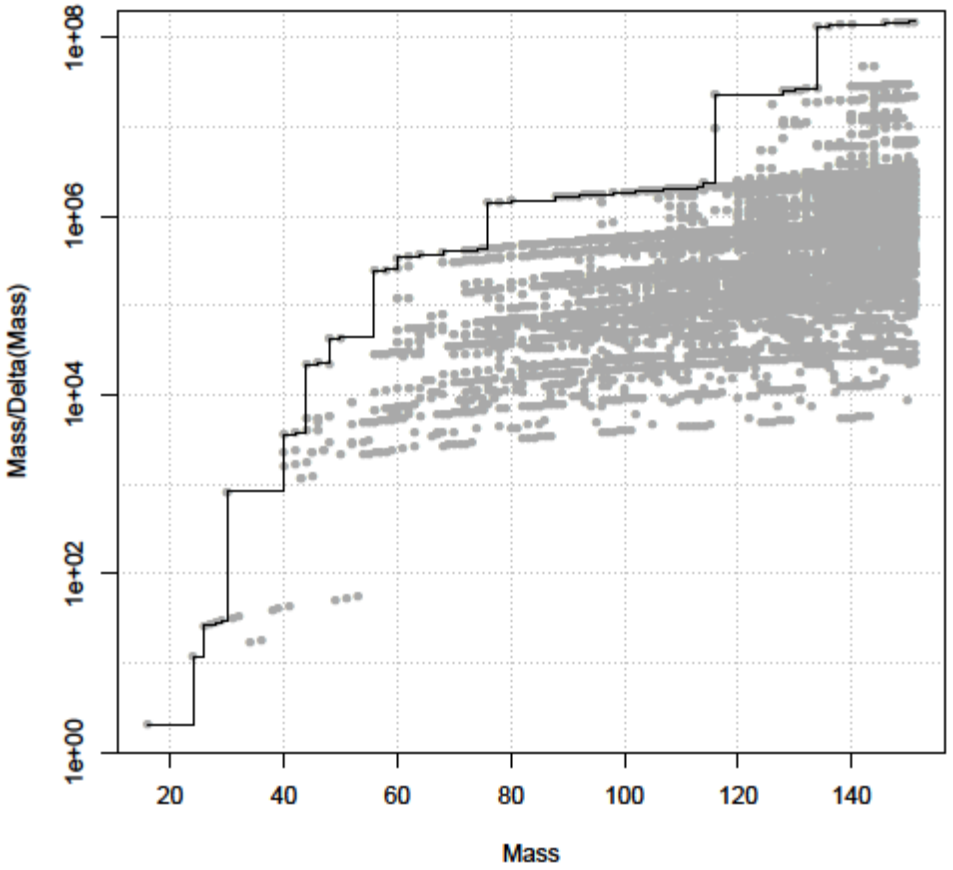
*question received from the Europa lander science definition team

# From Mass to Formula

Settings:

- Element set:
  C, H, N, O, F, P,
  I, S, Si, Br, Cl
- Constraint:
  atom-valence rules
  (LEWIS and SENIOR
  check of the
  "7 Golden Rules")

14138
molecular formulas

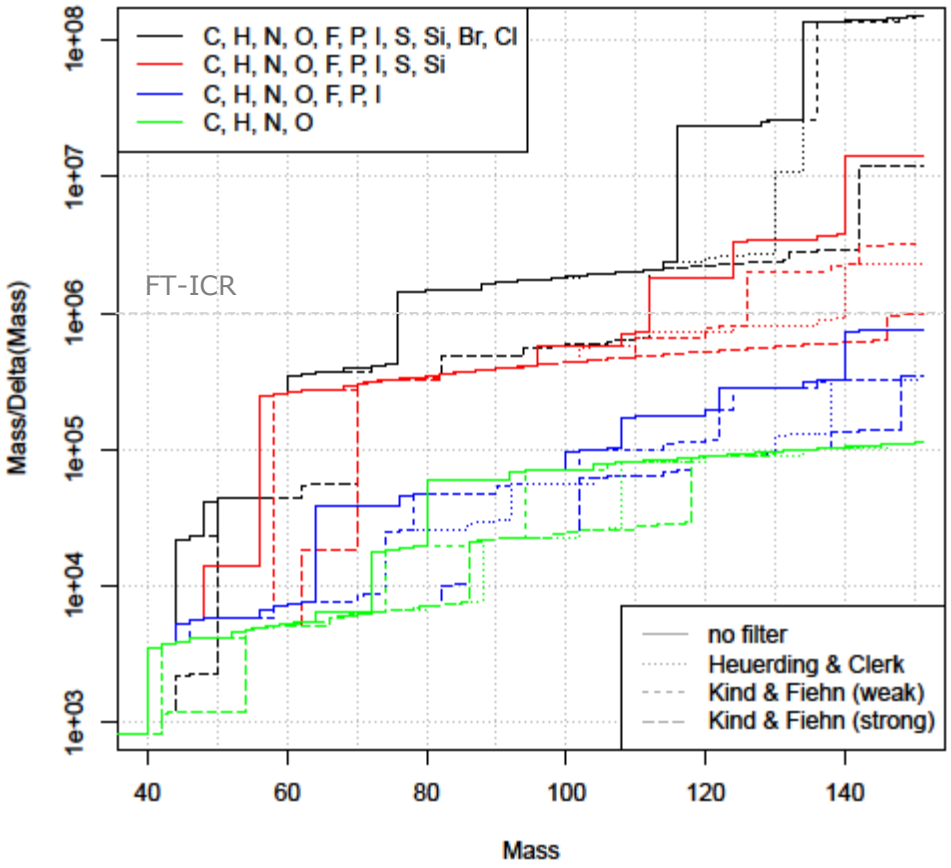



T. Kind and O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinf. 8 (2007), 105.

# From Mass to Formula

More settings:

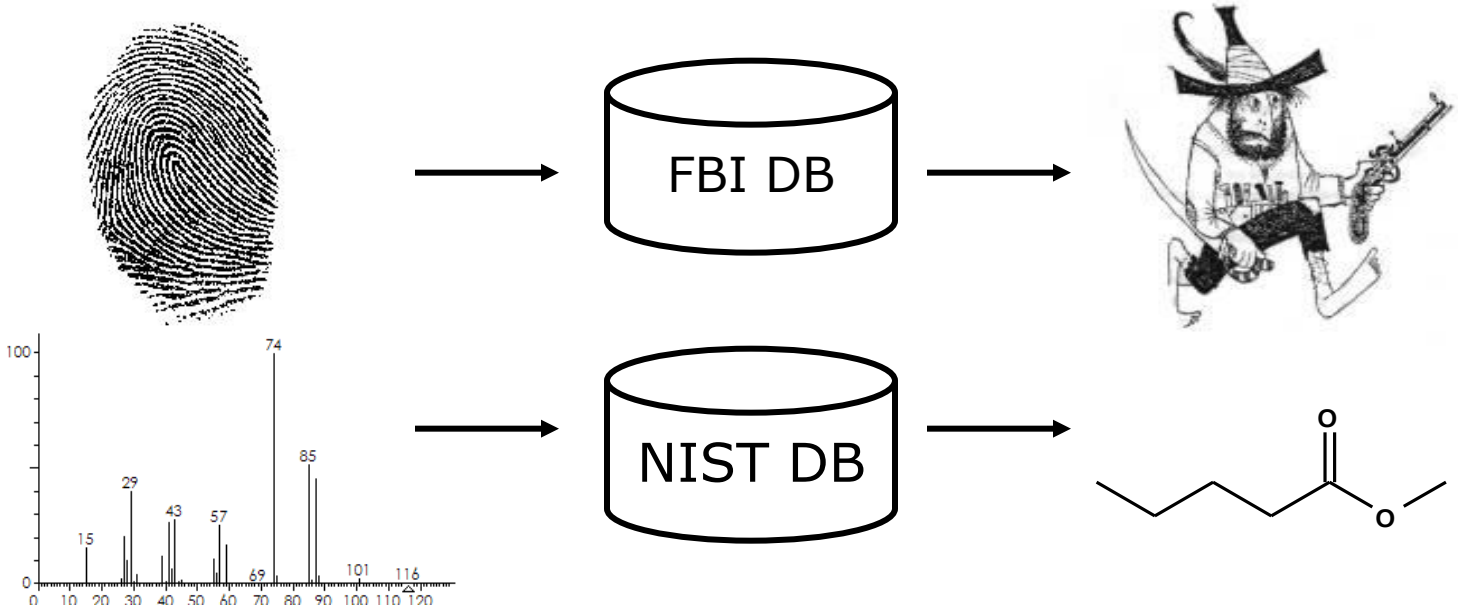
- Other element sets:
  C, H, N, O, F, P, I, S, Si
  C, H, N, O, F, P, I
  C, H, N, O
- Additional heuristic filters:
  element ratios by Heuerding & Clerc, Kind & Fiehn




S. Heuerding and J. T. Clerc, Simple tools for the computer-aided interpretation of mass spectra, Chemom. Intell. Lab. Syst. 20 (1993), 57{69.

# From Spectrum to Structure

- Established approach: use spectral data as molecular fingerprint for a database search



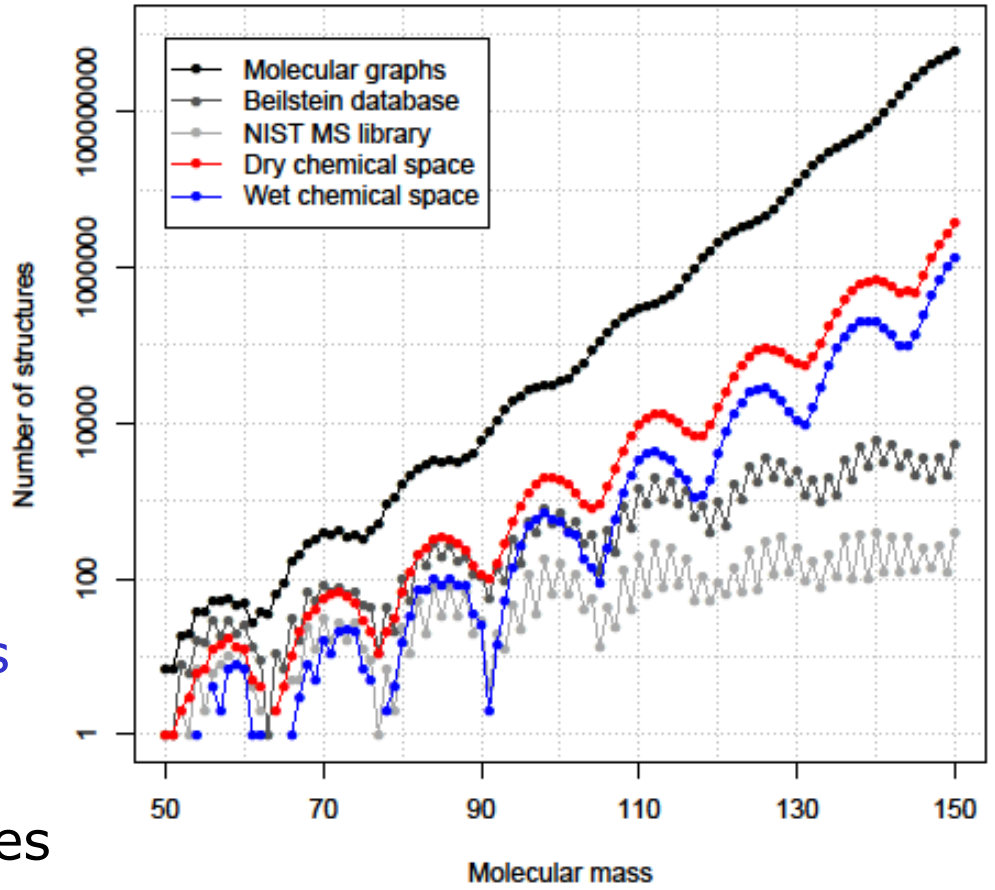

- Problem: only such data can be found that is stored in the database

# The Database – Chemical Space Gap

Structures:

- elements C, H, N, O
- at least 1 C-atom
- standard valences
- no charges, no radicals
- no stereoisomers
- only connected structures

Plus lists of forbidden sub-structures for chemical spaces with and without hydrolysis

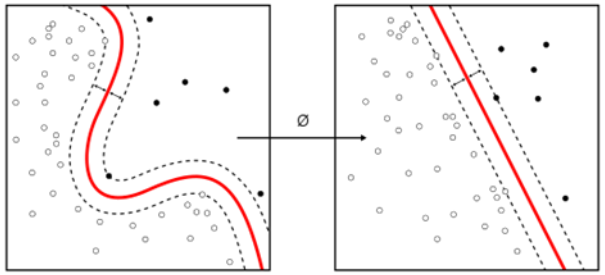



Structure generator: MOLGEN

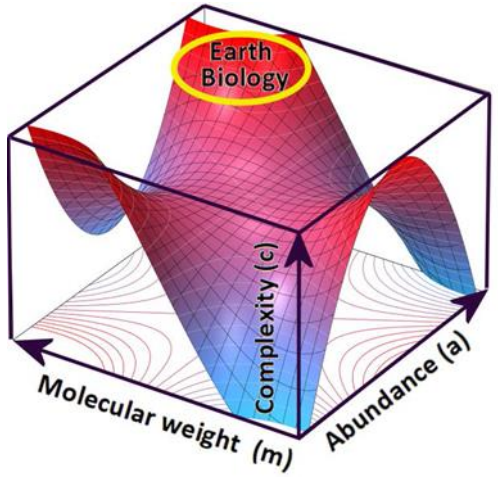
A. Kerber, R. Laue, M. Meringer, C. Rücker: Molecules in Silico: Potential versus Known Organic Compounds. MATCH 54 (2), 301-312, 2005.

# Approach: From Structure to Life

- Use machine learning methods to classify structures of biotic and abiotic compounds, e.g.
  - linear discriminant analysis
  - neural networks,
  - support vector machines,
  - decision trees,…



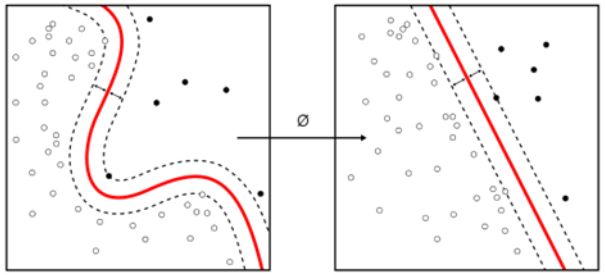

- Input variables: molecular descriptors, representing structural properties, e.g.
  - size (atom count, nominal weight,…),
  - complexity (e.g. information content),
  - degree of branching (e.g. walk counts),…
- Target variable:
  - biotic compound (Y/N)



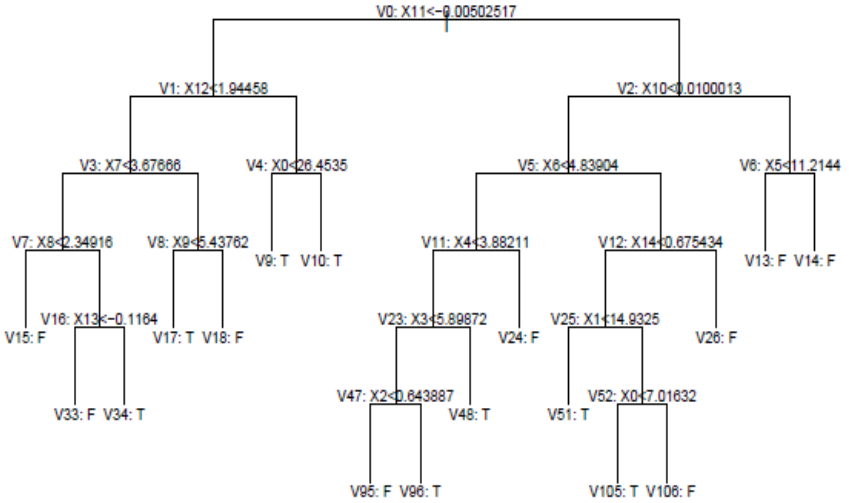

L. R. Cronin, Towards the Universal Life Detection System, Astrobiology Science Conference 2015

# Approach: From Spectra to Life

- Use machine learning methods to classify spectra of biotic and abiotic compounds, e.g.
  - linear discriminant analysis
  - neural networks,
  - support vector machines,
  - decision trees,…



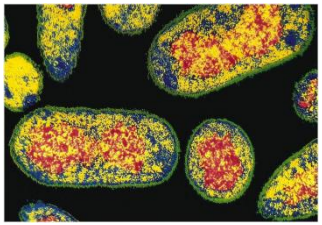

- Input variables: spectral descriptors, e.g.
  - ion series descriptors,
  - autocorrelation descriptors,
  - spectra shape descriptors,
  - logarithmic intensity ratios, …
- Target variable:
  - biotic compound (Y/N)

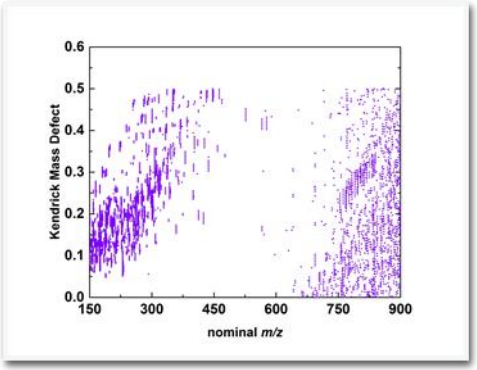


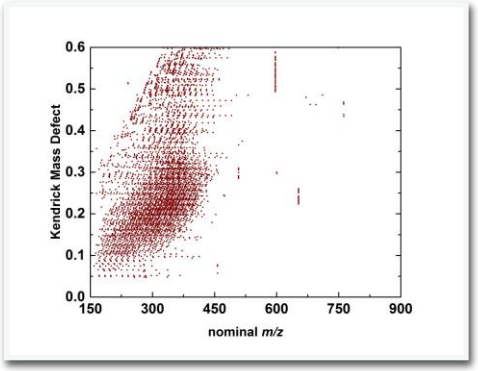

W. Werther et al, Classification of mass spectra, Chemom. Intell. Lab. Syst. 22, 63-76, 1994

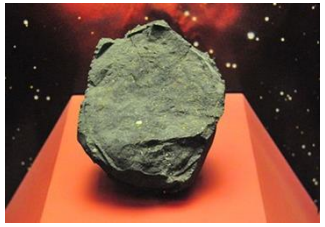
# Approach: From Masses to Life

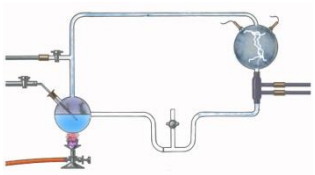
## Pattern recognition via Kendrick plots

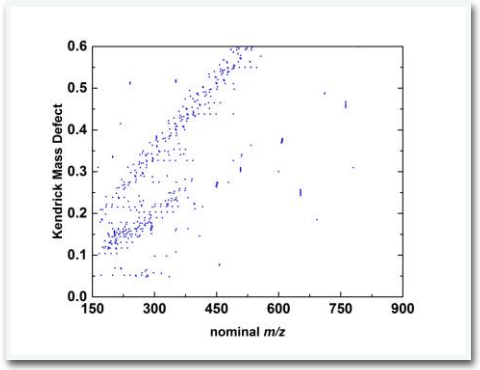


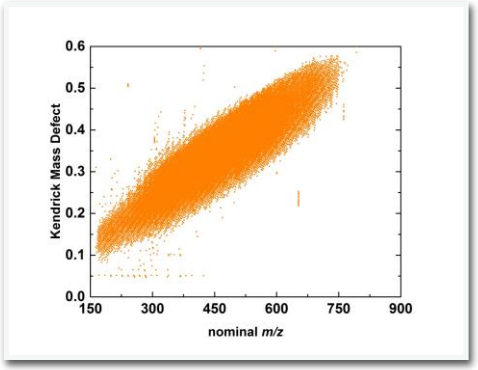






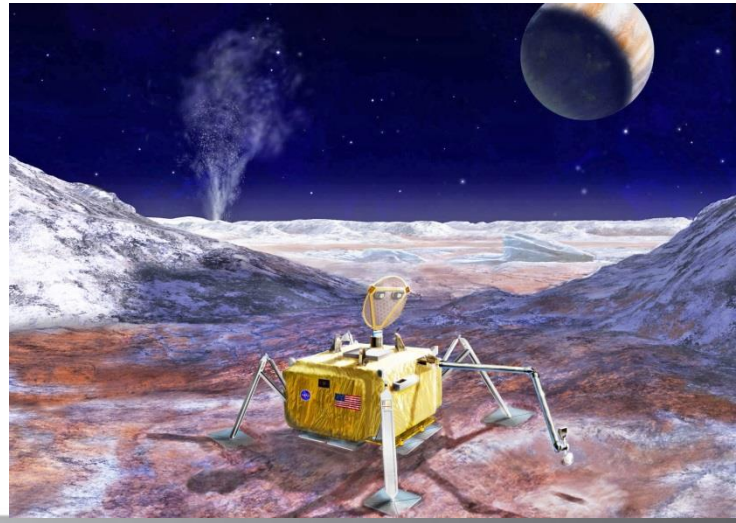

H. J. Cleaves, C. Giri, Universal Mass Spectrometry-Based Life Detection, Planetary Science Vision 2050 Workshop 2017

# Conclusions / Recommendations

- Find a good trade-off between compound characterization and life detection, provide
  - soft ionization mode to preserve the molecular ion
  - EI ionization mode the get a fragmentation spectrum
  - chromatography methods for compound separation

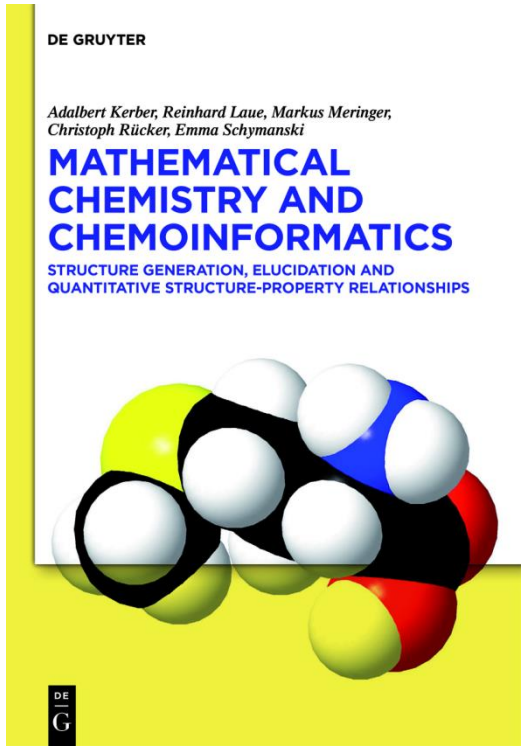
- Use simulations to define requirements (e.g. regarding mass resolution)

- For biotic/abiotic classification use
  - data mining
  - pattern recognition
  - machine learning



# Acknowledgements

Jim Cleaves
Caitanya Giri




MOLGEN Team
former Mathematics II
University of Bayreuth
www.molgen.de

next talk: Tue Oct 10, 13:30
Generation of Molecules
EON WS on Comput. Chem.

THANKS FOR YOUR ATTENTION!