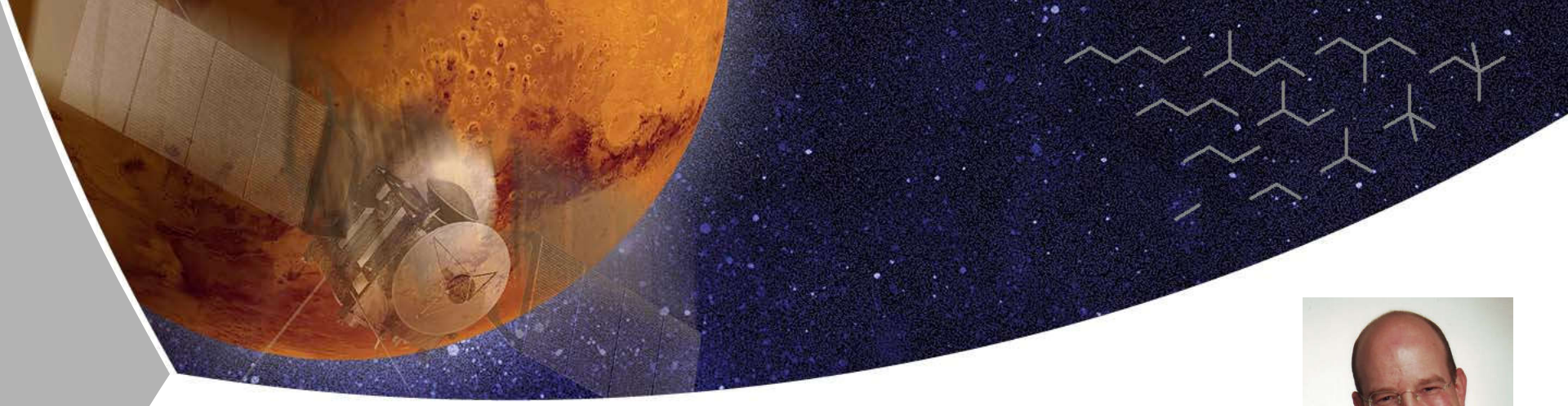


Markus Meringer

Algorithms for Chemical Space Enumeration and Applications in Astrobiology

Small Organic Molecules: Chemical Space, Reactions, Catalysis and Autocatalysis Workshop

Department of Statistics, University of Oxford, UK

March 24, 2020

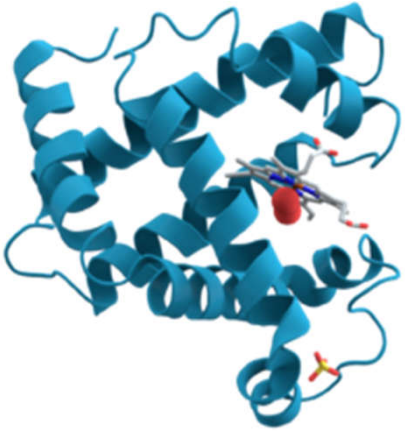
DLR Oberpfaffenhofen

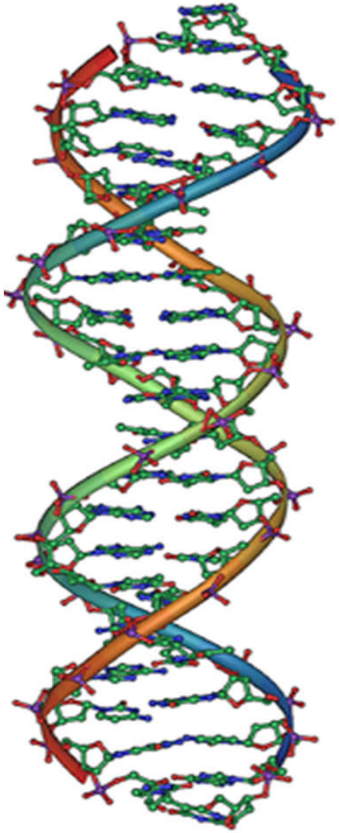
DLR Deutsches Zentrum für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

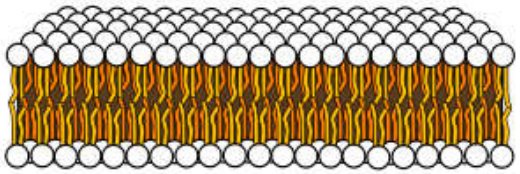
Outline

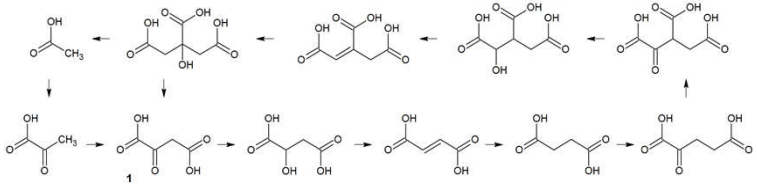
- Data structures and algorithms
  - representation of chemical structures in a computer
  - algorithms for enumerating chemical structures
- Applications in astrobiology
  - construction and
  - analysis of

  virtual libraries of
  - amino acids
  - nucleotide analogs
  - other biomolecules…









Representing chemical compounds: What precisely are we talking about?

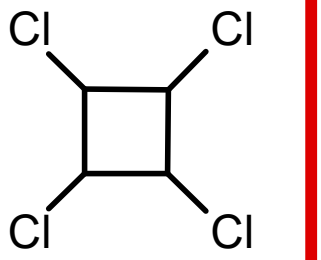
Different levels of abstraction

Composition

$C_4H_4Cl_4$

molecular formula

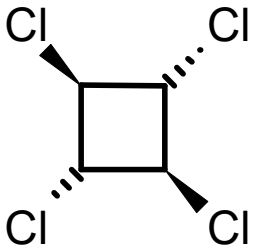
Constitution



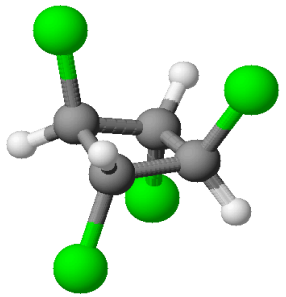
structural formula

Specialization

Configuration



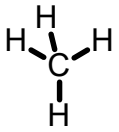
Conformation

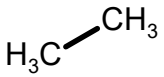


Generalization

From compositions to constitutions

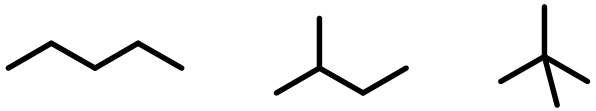
Example: Alkanes C_nH_{2n+2}

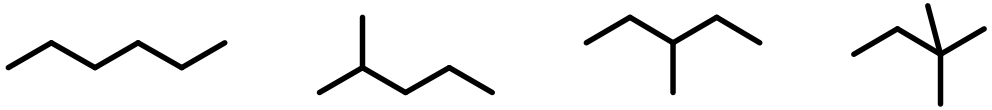
CH_4 

C_2H_6  or simply 

C_3H_8 

C_4H_{10} 

C_5H_{12} 

C_6H_{14} 

C_7H_{16} ... 9 *isomers* (try yourself – it‘s fun!)

Typically there are several, mostly very many structural formulas with the same molecular formula

Lists must be

- complete
- non-redundant

Exponential growth!

Applications: relating structure and properties

- From structure to physical, chemical, biological and pharmaceutical properties
  - structure-property relationships, esp. QSAR/QSPR
  - application of such relationships to predict properties of virtual structures (→ inverse QSAR)

- From physical and chemical properties (spectra) to structure

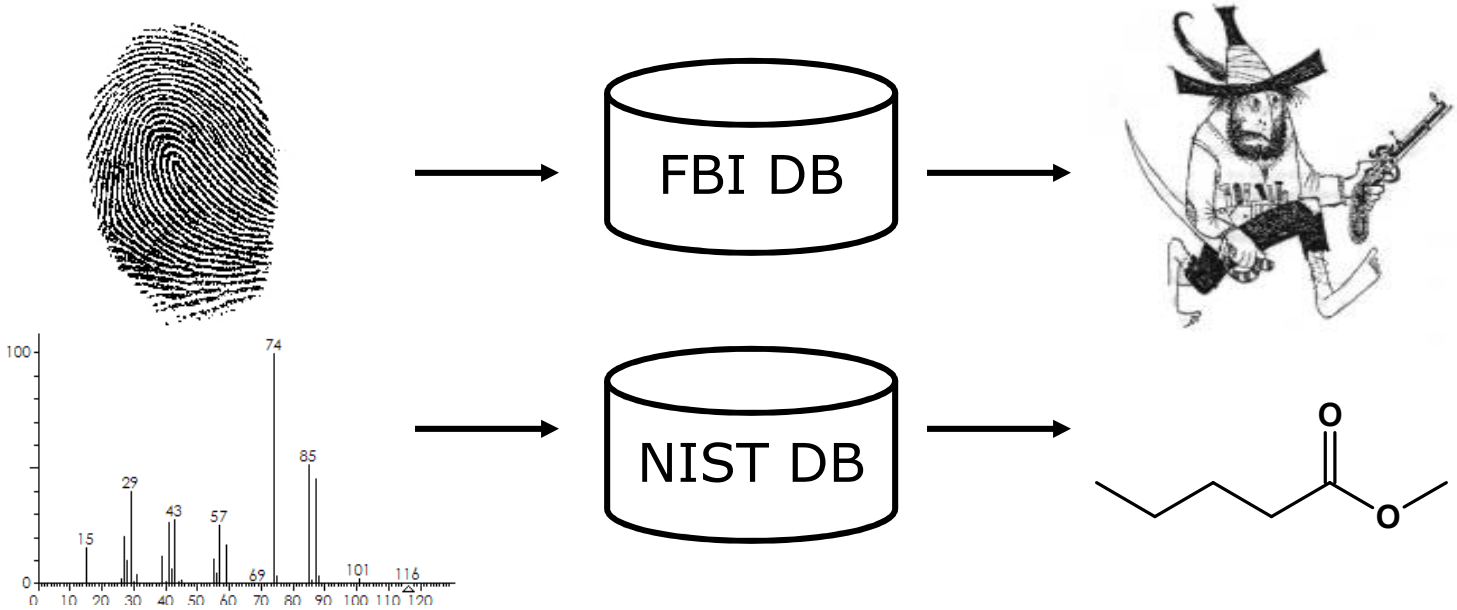
  computer-aided / automated

  molecular structure elucidation

  "CASE"

Structure elucidation by database searching

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

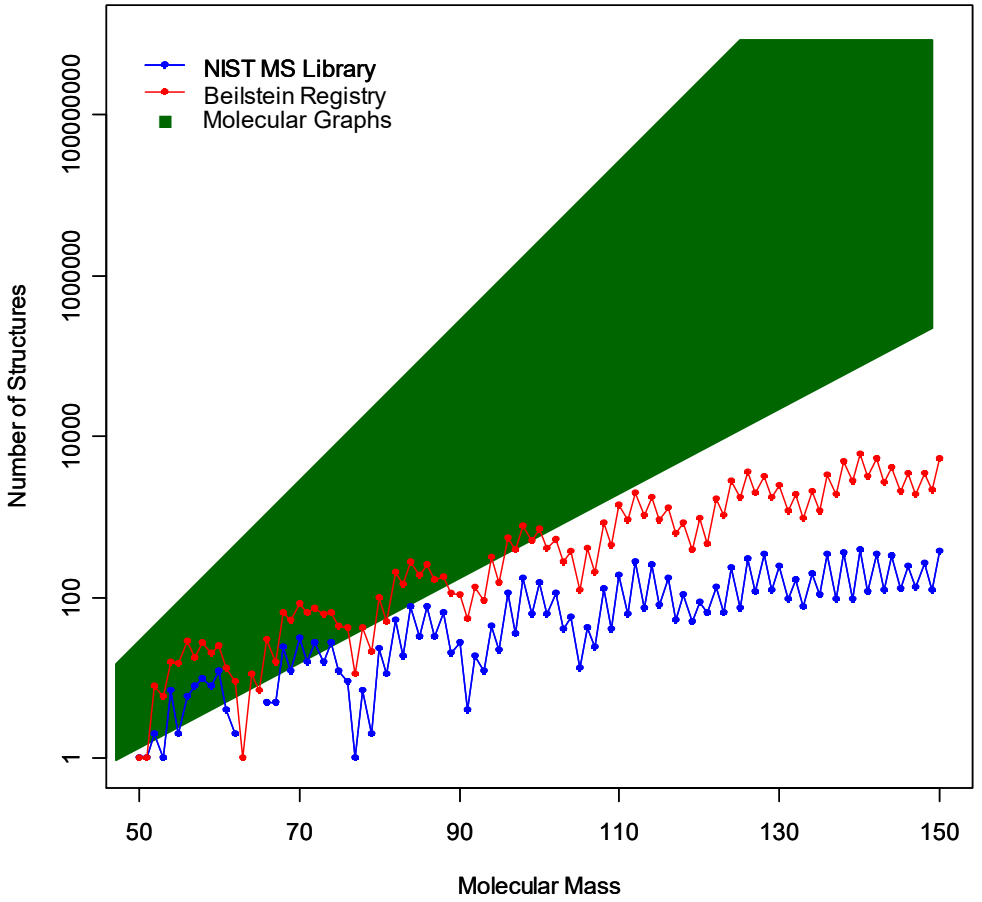


- Problem: only such data can be found that is stored in the database

Sizes of data bases

Structures:

- elements C, H, N, O
- at least 1 C-atom
- standard valencies
  C:4 H:1 N:3 O:2
- no charges
- no radicals
- only connected structures



Need for techniques to explore virtual chemical space in silico!

The DENDRAL project

- driven by exobiologist J. Lederberg
- initiated in the mid 1960's
- short for DENDRitic ALgorithm
- included an algorithm for generating acyclic structures
- partially funded by NASA
- aim: identifying unknown organic molecules by analyzing their mass spectra (MS) automatically
- perspective: processing of MS recorded on mars missions
- pioneer project in artificial intelligence, first expert system
- structure generators covering cyclic structures followed: StrGen, CONGEN, GENOA

R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum, J. Lederberg. Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project. McGraw-Hill Book Company, 1980.

DENDRAL approach to structure generation

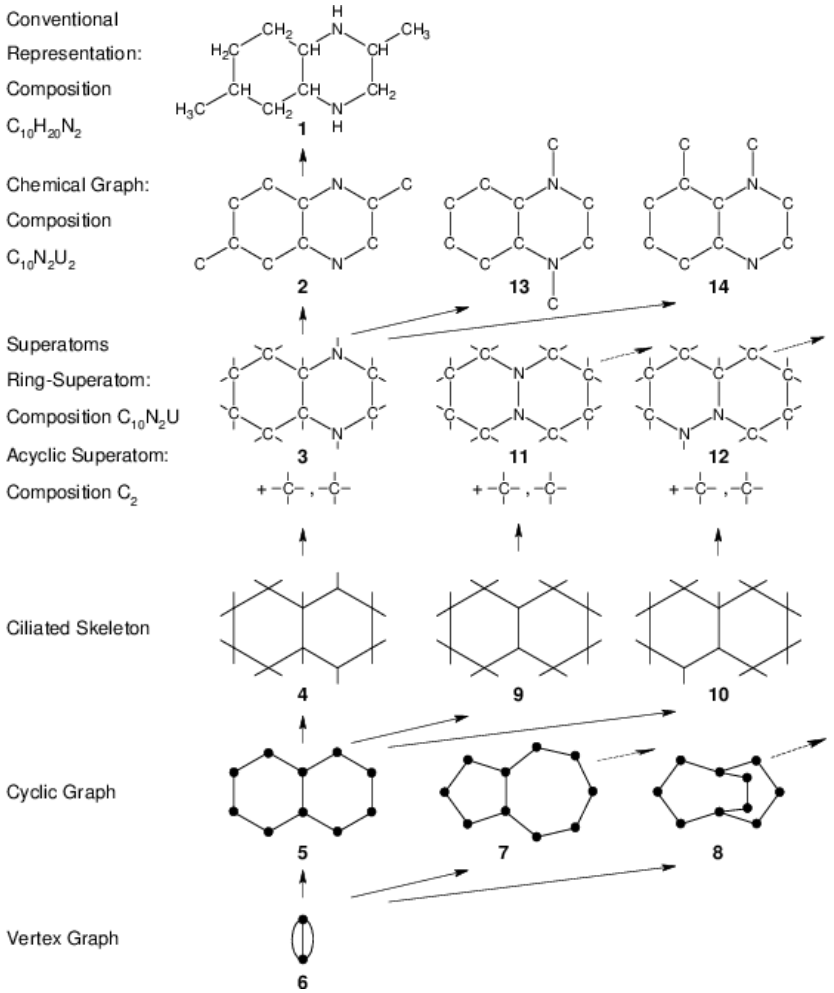
remove hydrogen

decompose into superatoms

strip element symbols

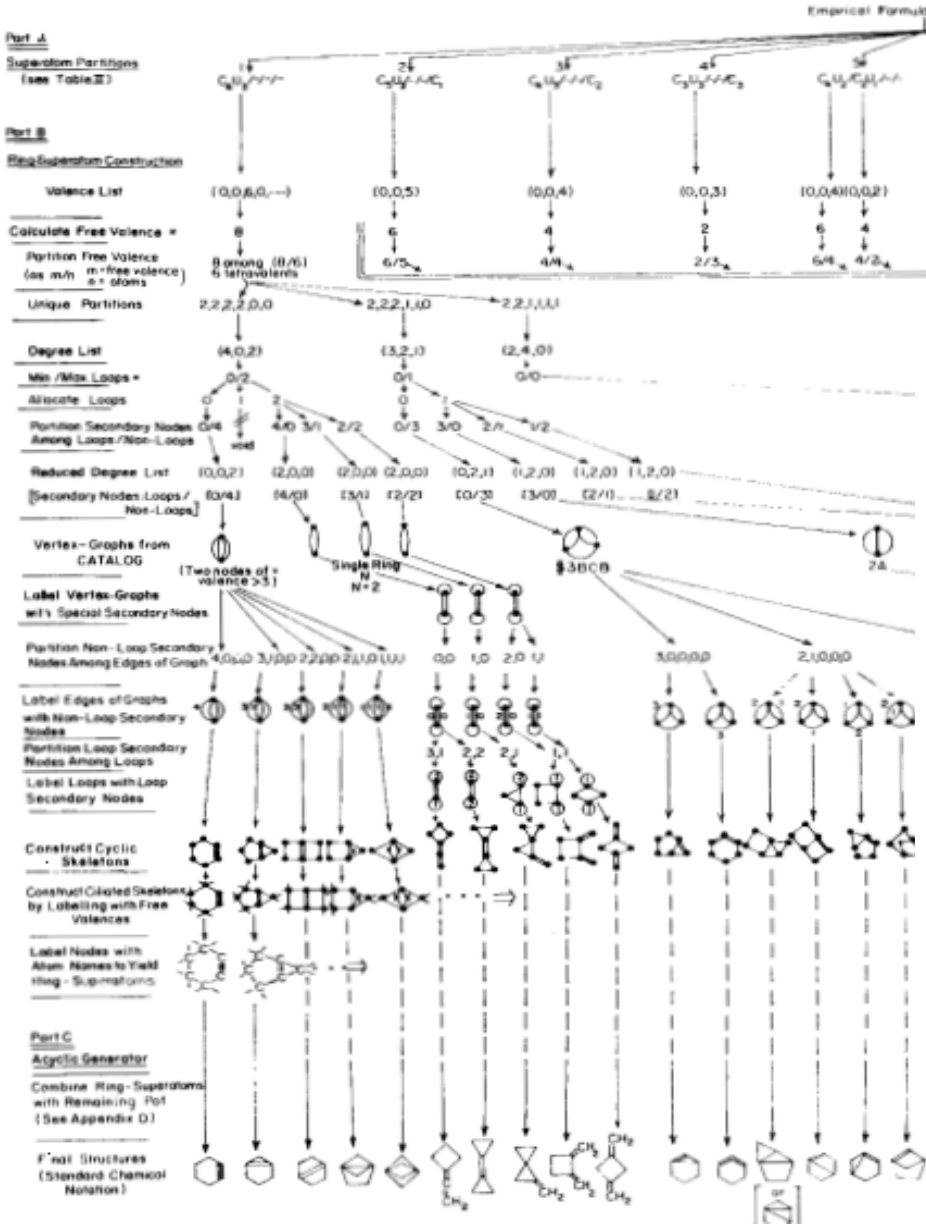
delete free valencies

replace chains of bivalent nodes by edges



LM Masinter, NS Sridharan, J Lederberg, DH Smith. Applications of Artificial Intelligence for Chemical Inference: XII. Exhaustive Generation of Cyclic and Acyclic Isomers. J. Am. Chem. Soc. 96(25) 7702-7717, 1974

Generating tree for C_6H_{10} isomers

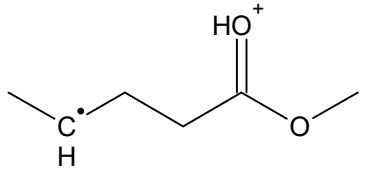


- masterpiece of computer programming
- especially in consideration of limited hardware resources, operation systems, programming languages available at this time
- however, this approach was very complicated
- particularly not suited to process structural constraints efficiently

Molecular graphs

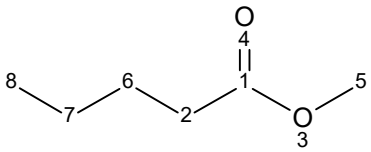
- Chemical compounds as molecular graphs

  vertices and edges (simple graph)
  + bond multiplicities (multigraph)
  + element & atomic state symbols



- Representation of molecular graphs in a computer: adjacency matrix

  - label atoms with numbers



  - write bond multiplicities into a matrix

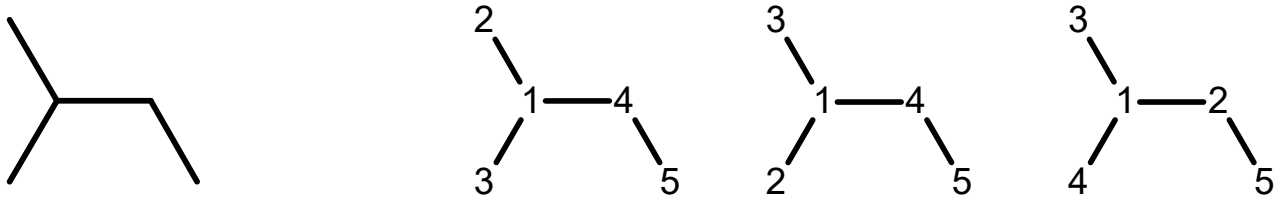
	1	2	3	4	5	6	7	8
1	0	1	1	2	0	0	0	0
2	1	0	0	0	0	1	0	0
3	1	0	0	0	1	0	0	0
4	2	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0
6	0	1	0	0	0	0	1	0
7	0	0	0	0	0	1	0	1
8	0	0	0	0	0	0	1	0

- Idea: fill adjacency matrix in all possible ways

Chemical compounds in nature and in silico

Chemical compounds

- in nature: atoms are not labeled
- in a computer: atoms have to be labeled



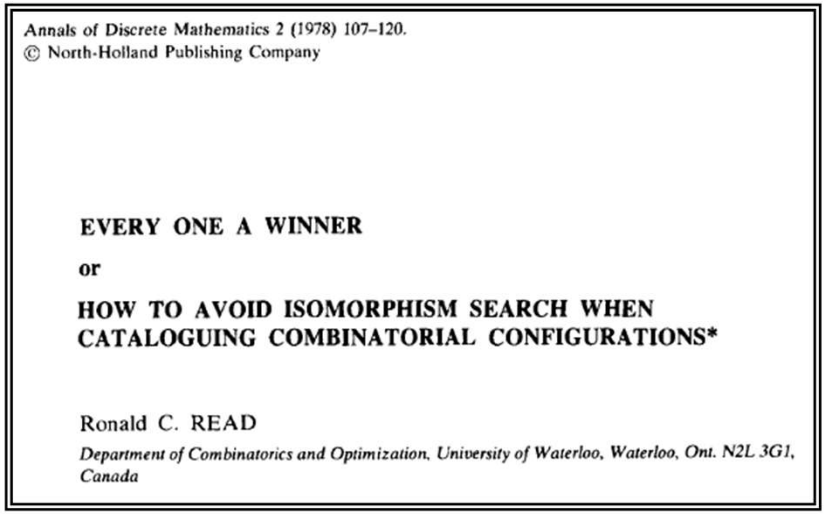
leads to problems

- up to n! different labeled (isomorphic) representations of an unlabeled structure
- deciding whether two labeled structures are isomorphic is computationally expensive
- “graph isomorphism problem”

Discrete mathematicians found solutions

Orderly generation

- principle found by Read in 1978
- reduced the number of isomorphism tests



Annals of Discrete Mathematics 2 (1978) 107–120.
© North-Holland Publishing Company

EVERY ONE A WINNER

or

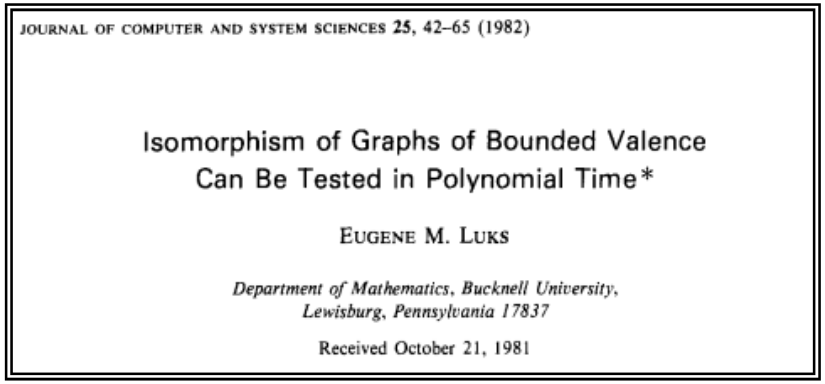
HOW TO AVOID ISOMORPHISM SEARCH WHEN CATALOGUING COMBINATORIAL CONFIGURATIONS*

Ronald C. READ

Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada

Fast isomorphism tests

- Luks found polynomial time algorithm in 1982
- note: molecular graphs have valences at most 4 (or maybe 6 for S)



JOURNAL OF COMPUTER AND SYSTEM SCIENCES 25, 42–65 (1982)

Isomorphism of Graphs of Bounded Valence Can Be Tested in Polynomial Time*

EUGENE M. LUKS

Department of Mathematics, Bucknell University, Lewisburg, Pennsylvania 17837

Received October 21, 1981

Order on edges of labeled graphs

Order on edges of graphs:

$e = (x,y)$, $e' = (x',y')$ with $x < y$, $x' < y'$

then $e < e'$, iff

$x < x'$ or ($x = x'$ and $y < y'$)

Examples:

$(1,2) < (2,3)$

$(1,2) < (1,3)$

Order on labeled graphs

Lexicographical order on graphs on n nodes

$\gamma = \{e_1,\ldots,e_t\}$ with $e_1 < \ldots < e_t$

$\gamma' = \{e'_1,\ldots,e'_{t'}\}$ with $e'_1 < \ldots < e'_{t'}$

then $\gamma < \gamma'$, iff

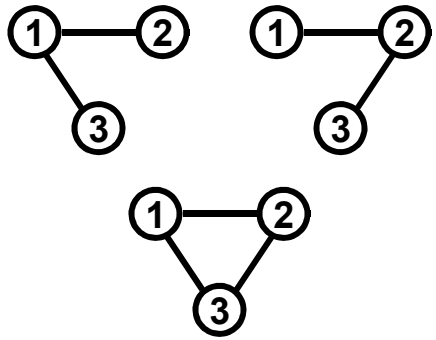
(there is an i with $e_i < e'_i$ and for all $j < i$: $e_j = e'_j$) or

($t < t'$ and for all $j \leq t$: $e_j = e'_j$)

Examples: graphs on 3 nodes 1, 2, 3

{(1,2),(1,3)} < {(1,2),(2,3)}

{(1,2),(1,3)} < {(1,2),(1,3),(2,3)}



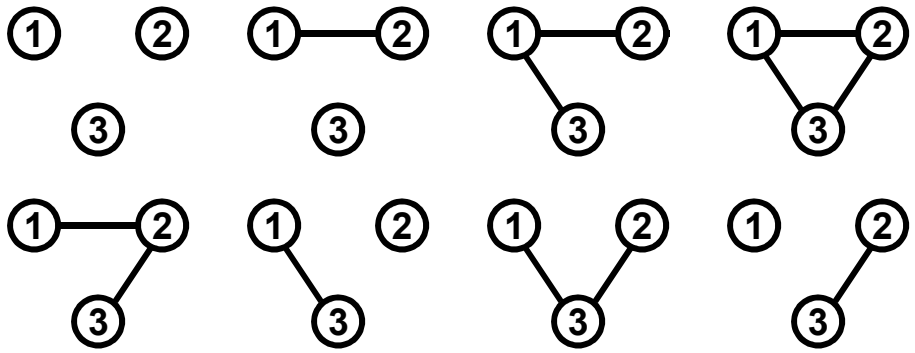
Generation of labeled graphs

Algorithm: Labeled Generation (γ)

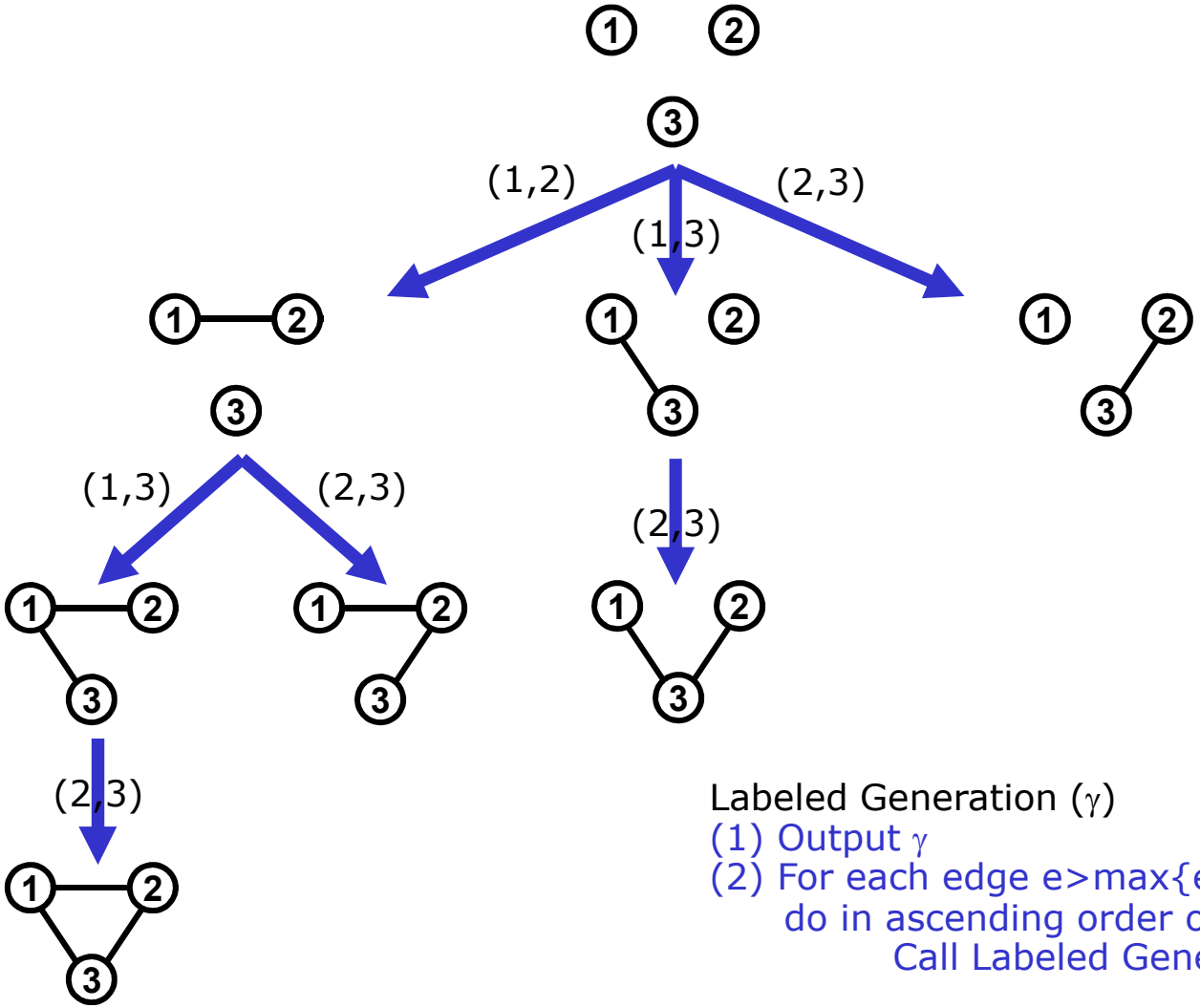
(1) Output γ

(2) For each edge $e > \max\{e' \in \gamma\}$
do in ascending order of e
Call Labeled Generation ($\gamma \cup \{e\}$)

Example: graphs on 3 nodes starting with the empty graph, Labeled Generation ({}) produces the output



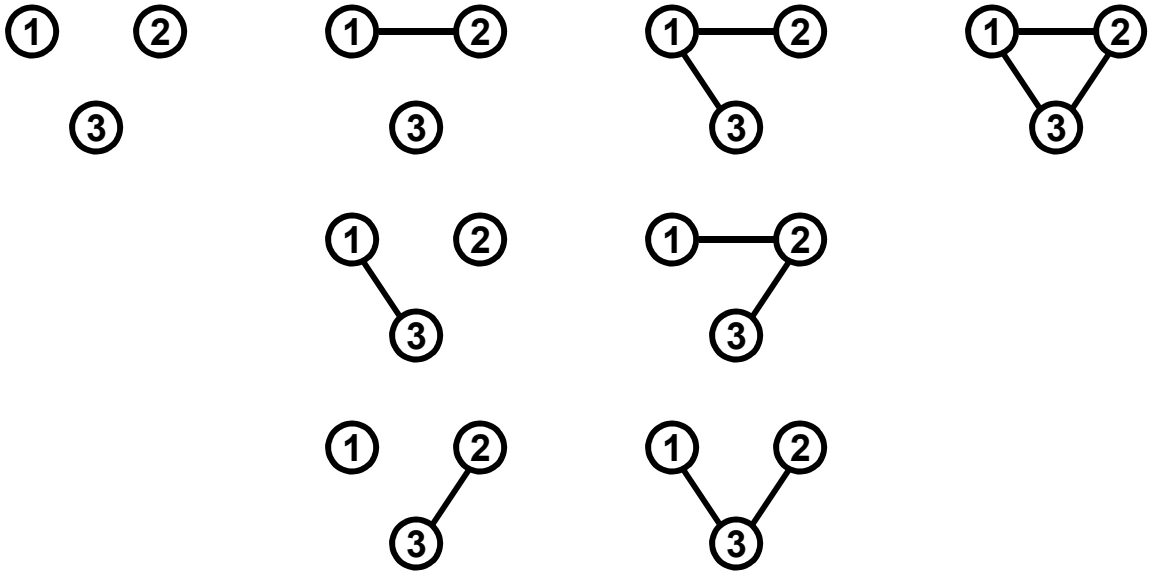
Example: labeled graphs on 3 nodes



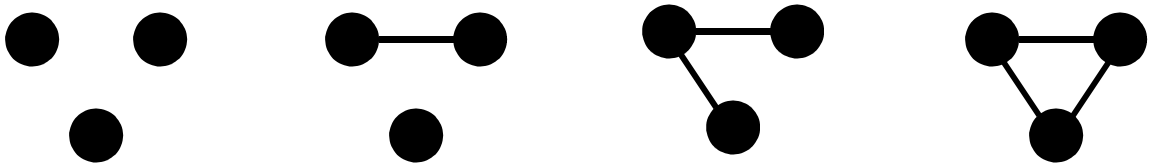
Labeled Generation (γ)
(1) Output γ
(2) For each edge e>max{e´$\in \gamma$}
do in ascending order of e
Call Labeled Generation ($\gamma \cup$ {e})

From labeled to unlabeled graphs

How to obtain from labeled graphs ...

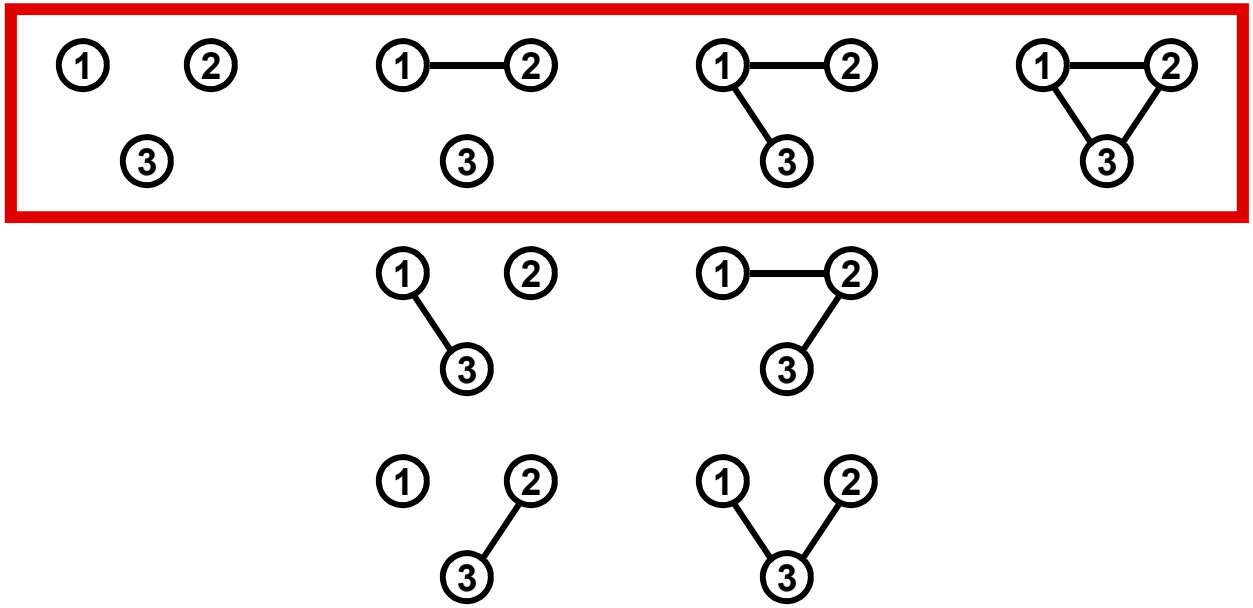


... unlabeled graphs ?



Canonical orbit representatives

Solution: Select from each orbit (column) the lexicographically minimal representative



Note: Testing minimality is a rather expensive procedure, up to n! permutations have to be checked

Testing minimality

γ is minimal, iff

for each permutation π of the symmetric group S_n:

$\gamma \leq \pi(\gamma)$

Example:

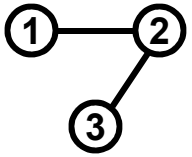
$\pi_3(\{(1,2),(2,3)\})$

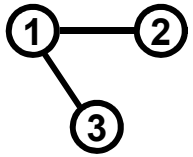
$= \{(2,1),(1,3)\}$

$= \{(1,2),(1,3)\}$

$< \{(1,2),(2,3)\}$

$\Rightarrow$ not minimal





x →	1	2	3
$\pi_1(x)$	1	2	3
$\pi_2(x)$	1	3	2
$\pi_3(x)$	2	1	3
$\pi_4(x)$	2	3	2
$\pi_5(x)$	3	1	2
$\pi_6(x)$	3	2	1

Note: Using algebraic and group-theoretic methods, costs for testing minimality can be reduced considerably

J Braun, R Gugisch, A Kerber, R Laue, M Meringer, C Rücker:
MOLGEN-CID - A canonizer for molecules and graphs accessible through the Internet. *J Chem Inf Comput Sci* 44: 542-548, 2004

Generation of unlabeled graphs

Algorithm: Labeled Generation (γ)

(1) Output γ

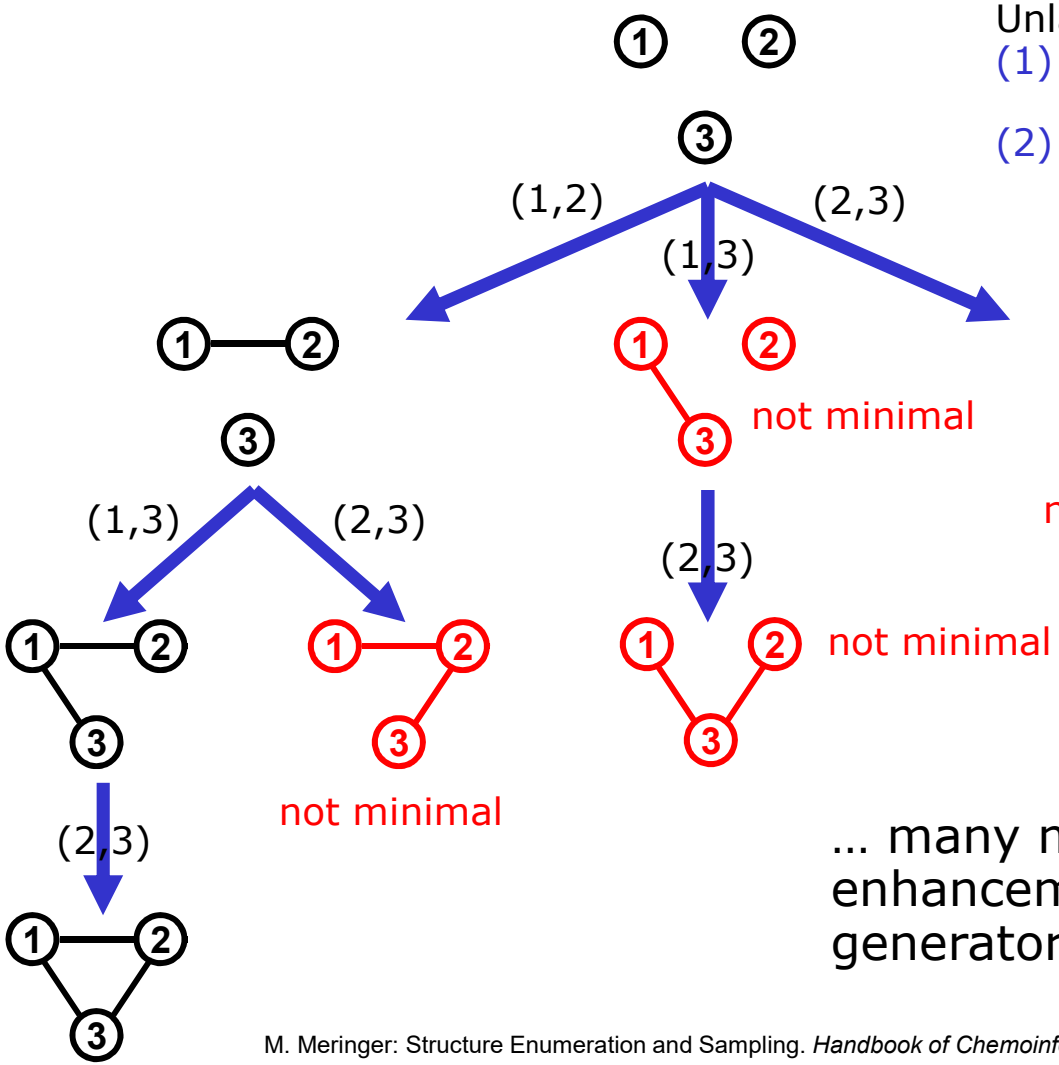
(2) For each edge e>max{e´ $\in \gamma$}
do in ascending order of e
Call Labeled Generation ($\gamma \cup$ {e})

Algorithm: Unlabeled Generation (γ)

(1) If γ is minimal in its orbit then
Output γ

(2) For each edge e>max{e´ $\in \gamma$}
do in ascending order of e
Call Unlabeled Generation ($\gamma \cup$ {e})

Example: unlabeled graphs on 3 nodes



Unlabeled Generation (γ)
(1) If γ is minimal in its orbit then
Output γ
(2) For each edge $e > \max\{e' \in \gamma\}$
do in ascending order of e
Call Unlabeled Generation
($\gamma \cup \{e\}$)

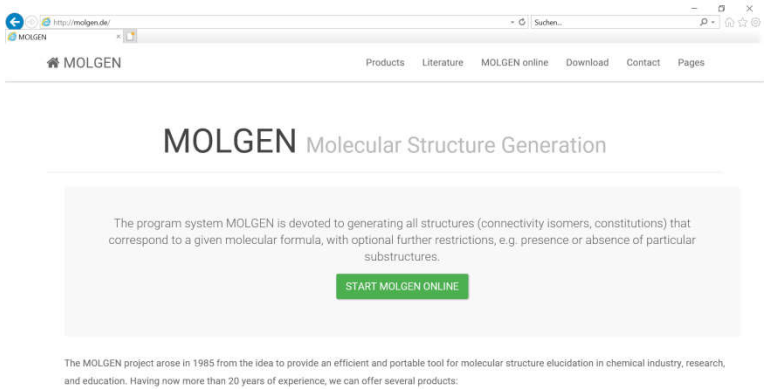
… many more refinements and enhancements lead to an efficient generator of molecular graphs …

M. Meringer: Structure Enumeration and Sampling. *Handbook of Chemoinformatics Algorithms*. CRC/Chapman&Hall, 233-267, 2010.

A new generation of structure generators

www.molgen.de

- MOLGEN 3.5 (1997, Win 95)
- MOLGEN 4.0 (1998, UNIX)
- MOLGEN 5.0 (2007, Win, Linux)
- others, e.g. Assemble, OMG

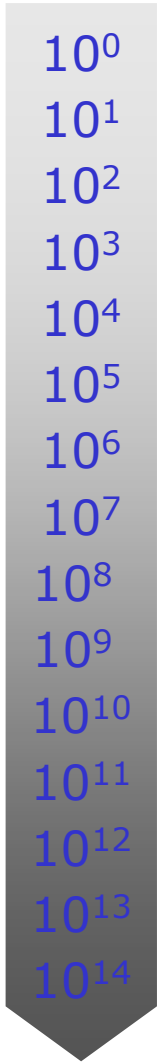


Computational example with constraints

Restrictions	no. of isomers	CPU-time
Chemical formula $C_6H_8O_6$ only	2,558,517	838 s
no triple bonds	2,434,123	703 s
hydrogen distribution $1CH_2$,$2CH_1$,3C,4OH	79,831	25 s
no substructure -O-O-	35,058	97 s
hybridization 1Csp3-2H,2Csp3-1H,3Csp2-OH,1Osp2-OH	990	8 s
minimal size of rings =5	348	5 s
contains at least one CO_3 branch	15	11 s

T. Grüner, A. Kerber, R. Laue, M. Meringer: MOLGEN 4.0. MATCH Communications in Mathematical and in Computer Chemistry 37, 205-208, 1998.

Orders of magnitude of structural spaces and data bases

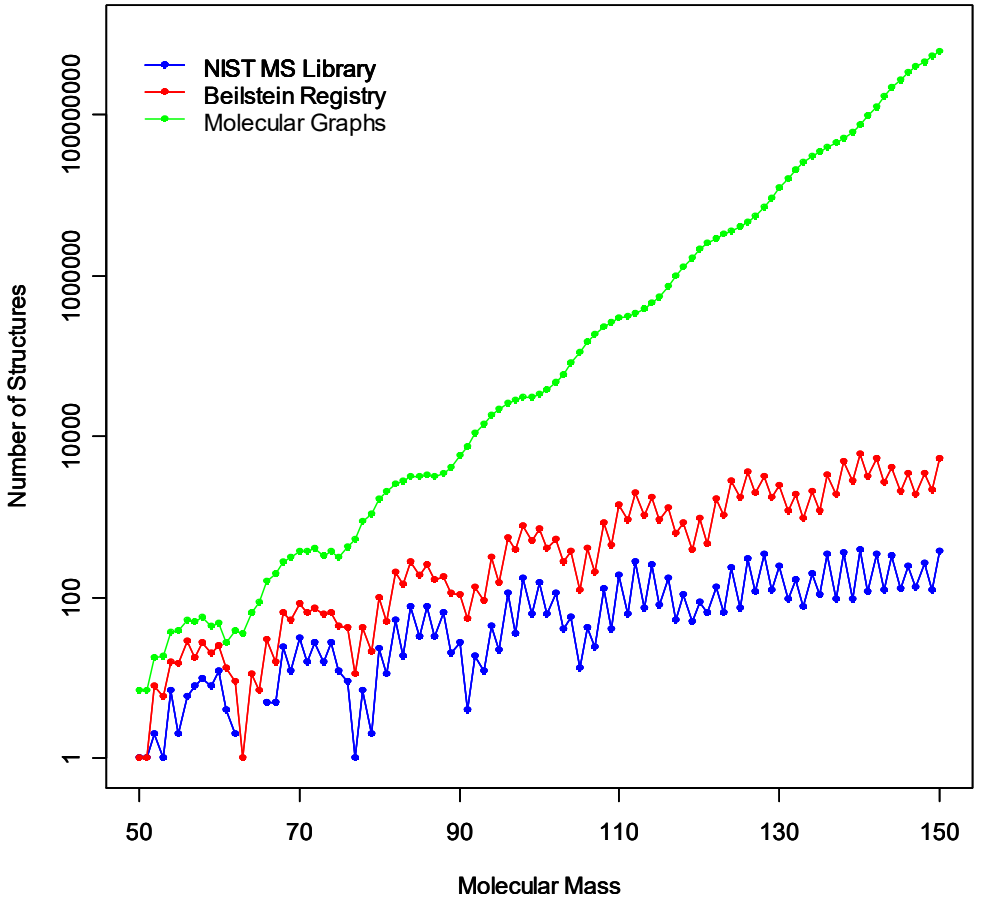


	10^0	paper and pencil (e.g. small alkanes)
	10^1	
	10^2	object lessons (e.g. 217 isomers of C_6H_6)
	10^3	automated structure elucidation via MS
	10^4	
	10^5	
NMR Shift DB (4.1e5)	10^6	
NIST MS DB (2.2e6)	10^7	automated structure elucidation via NMR
PubChem (1.8e8)	10^8	
GDB-13 (9.8e8)	10^9	molecular graphs (C,H,N,O, ≤150Da: 3.7e9)
	10^{10}	molecular graphs (C,H,O, ≤180Da: 6.7e10)
	10^{11}	
	10^{12}	
GDB-17 (1.7e12)	10^{13}	constitutional isomers of TRP (1.9e13)
	10^{14}	quartic graphs on 23 points (4.3e14)

Sizes of data bases and numbers of molecular graphs

Structures:

- elements C, H, N, O
- at least 1 C-atom
- standard valencies
  C:4 H:1 N:3 O:2
- no charges
- no radicals
- no stereoisomers
- only connected structures



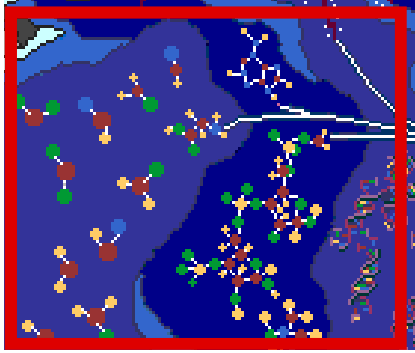
A. Kerber, R. Laue, M. Meringer, C. Rücker: Molecules in Silico: Potential versus Known Organic Compounds. MATCH 54 (2), 301-312, 2005.

How structure generation was rediscovered for astrobiology

About nine years ago...

I would like to generate a saturated "chemistry space" (i.e. list of isomers) for all possible alpha amino acids (NH2-CHR-COOH), where R is restricted to smallish side-chains of carbon (C=4), with additional sulphur (S≤1), oxygen (O≤2), nitrogen (N≤3) and hydrogen and a possible benzyl ring

Stephen Freeland
UHNAI



No. molecular formulas: 132 …
No. structures: 24749 … that's what I call a manageable chemical space

Amino acid libraries resulting from the studies at UHNAI

Jim Cleaves

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Beyond Terrestrial Biology: Charting the Chemical Universe of α-Amino Acid Structures

Markus Meringer,[†] H. James Cleaves II,[*,‡,§,⊥,∥] and Stephen J. Freeland[○]

[†]German Aerospace Center (DLR), Earth Observation Center (EOC), Münchner Straße 20, D-82234 Oberpfaffenhofen–Wessling, Germany

[‡]Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan

[§]Institute for Advanced Study, 1 Einstein Drive, Princeton, New Jersey 08540, United States

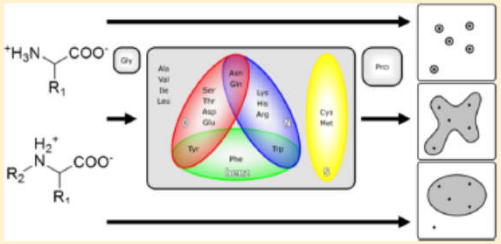
[⊥]Blue Marble Space Institute of Science, 2800 Woodley Road NW, no. 544, Washington, D.C. 20016, United States

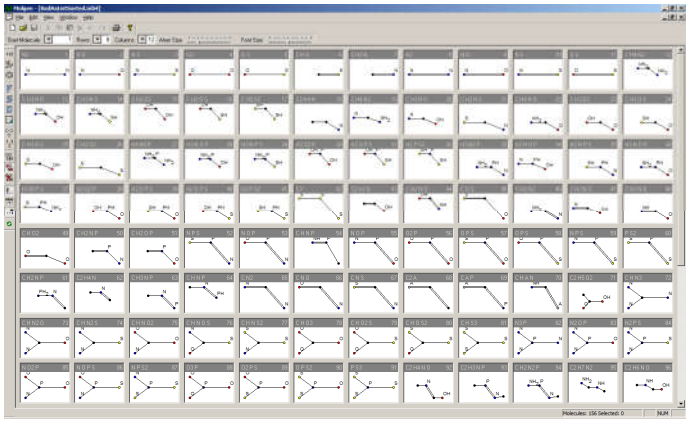
[∥]Center for Chemical Evolution, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

[○]NASA Astrobiology Institute, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822-1839, United States

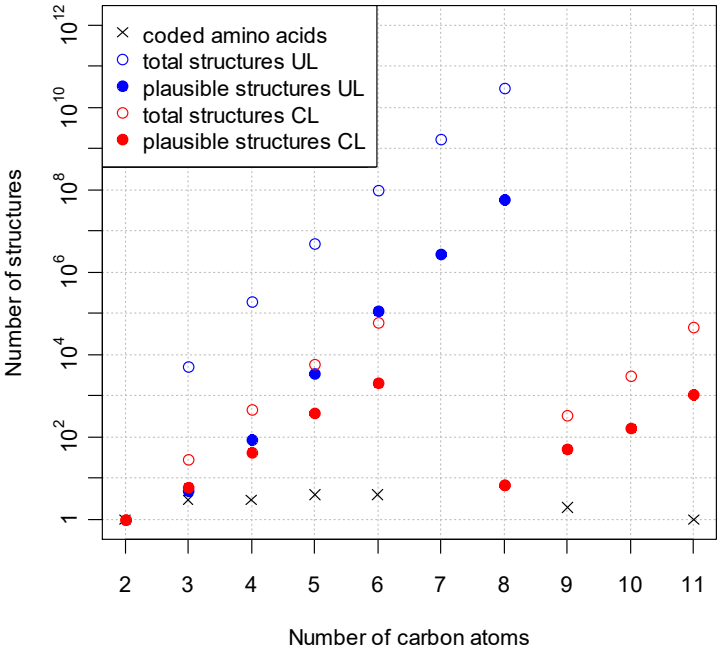
Supporting Information

ABSTRACT: α-Amino acids are fundamental to biochemistry as the monomeric building blocks with which cells construct proteins according to genetic instructions. However, the 20 amino acids of the standard genetic code represent a tiny fraction of the number of α-amino acid chemical structures that could plausibly play such a role, both from the perspective of natural processes by which life emerged and evolved, and from the perspective of human-engineered genetically coded proteins. Until now, efforts to describe the structures comprising this broader set, or even estimate their number, have been hampered by the complex combinatorial properties of organic molecules. Here, we use computer software based on graph theory and constructive combinatorics in order to conduct an efficient and exhaustive search of the chemical structures implied by two careful and precise definitions of the α-amino acids relevant to coded biological proteins. Our results include two virtual libraries of α-amino acid structures corresponding to these different approaches, comprising 121 044 and 3 846 structures, respectively, and suggest a simple approach to exploring much larger, as yet uncomputed, libraries of interest.





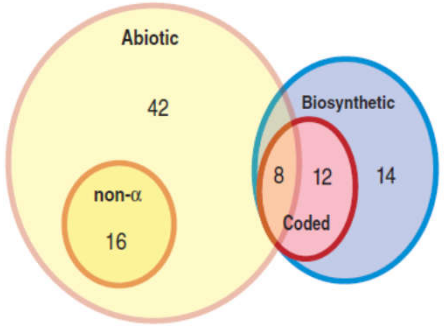
156-membered badlist

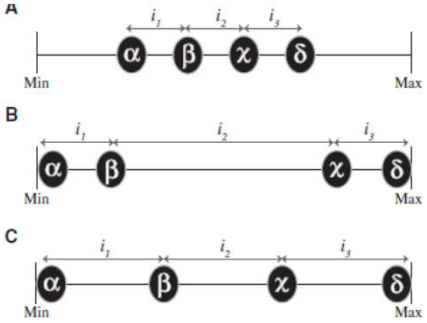


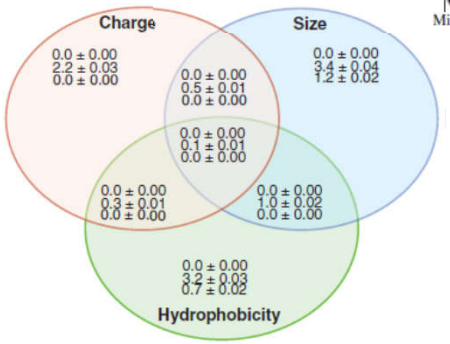
Application:
Verify a model on selection of the amino acid alphabet

- Model established previously on a small set of known amino acids
  - abiotic
  - coded
  - biosynthetic
- The 20 biologically encoded amino acids cover chemical space optimally in terms of
  - range and
  - evenness

  with respect to 3 properties
  - charge,
  - size and
  - hydrophobicity



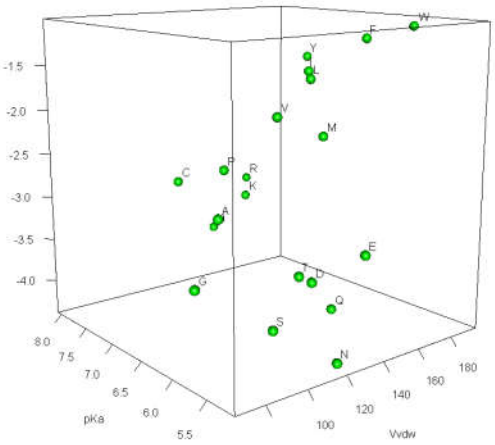




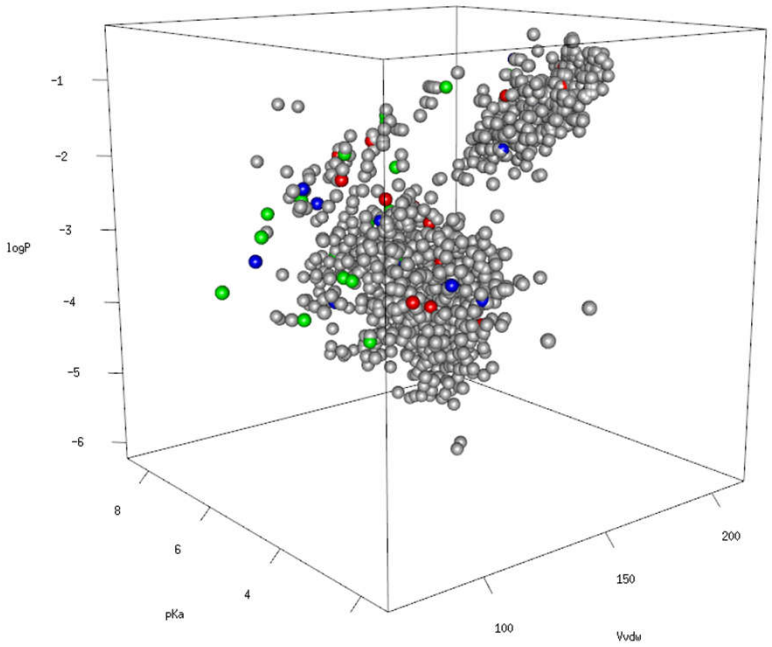
Philip GK, Freeland SJ: Did evolution select a nonrandom "alphabet" of amino acids? Astrobiology 11(3), 235 (2011)

Property calculation for the generated library

- Calculation of physico-chemical properties
  - hydrophobicity represented by logP (MOLGEN-QSPR)
  - size represented by Van der Waals volume V_{vdw} (MOLGEN-QSPR)
  - charge represented by pK_a (JChem)



… gives a 3D mapping of our amino acid chemical space



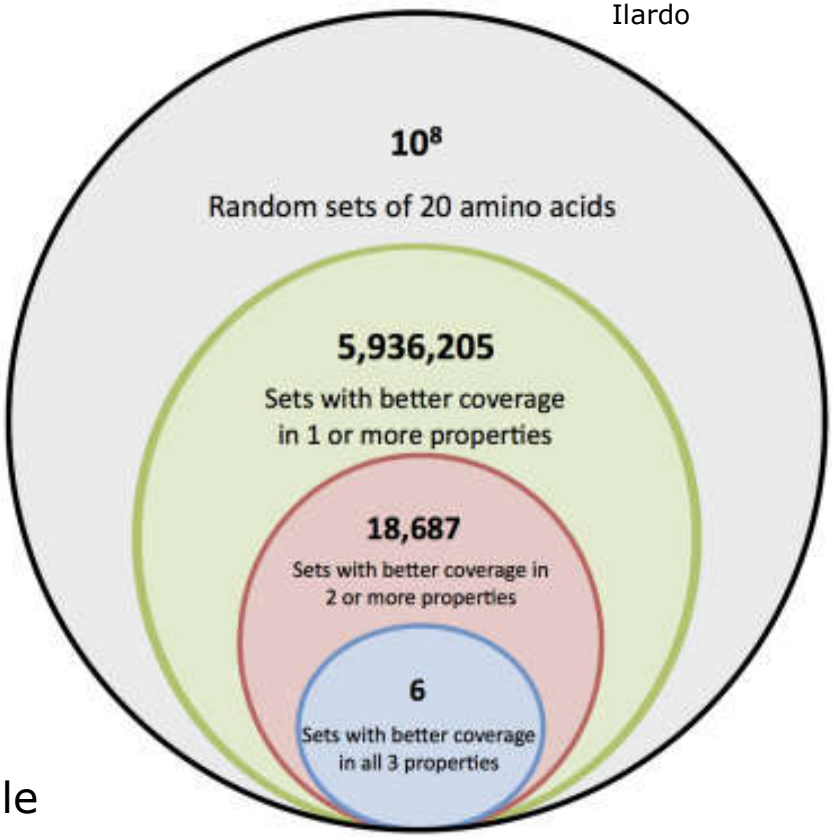
20 biologically encoded amino acids colored green

Statistical Analysis

Melissa Ilardo

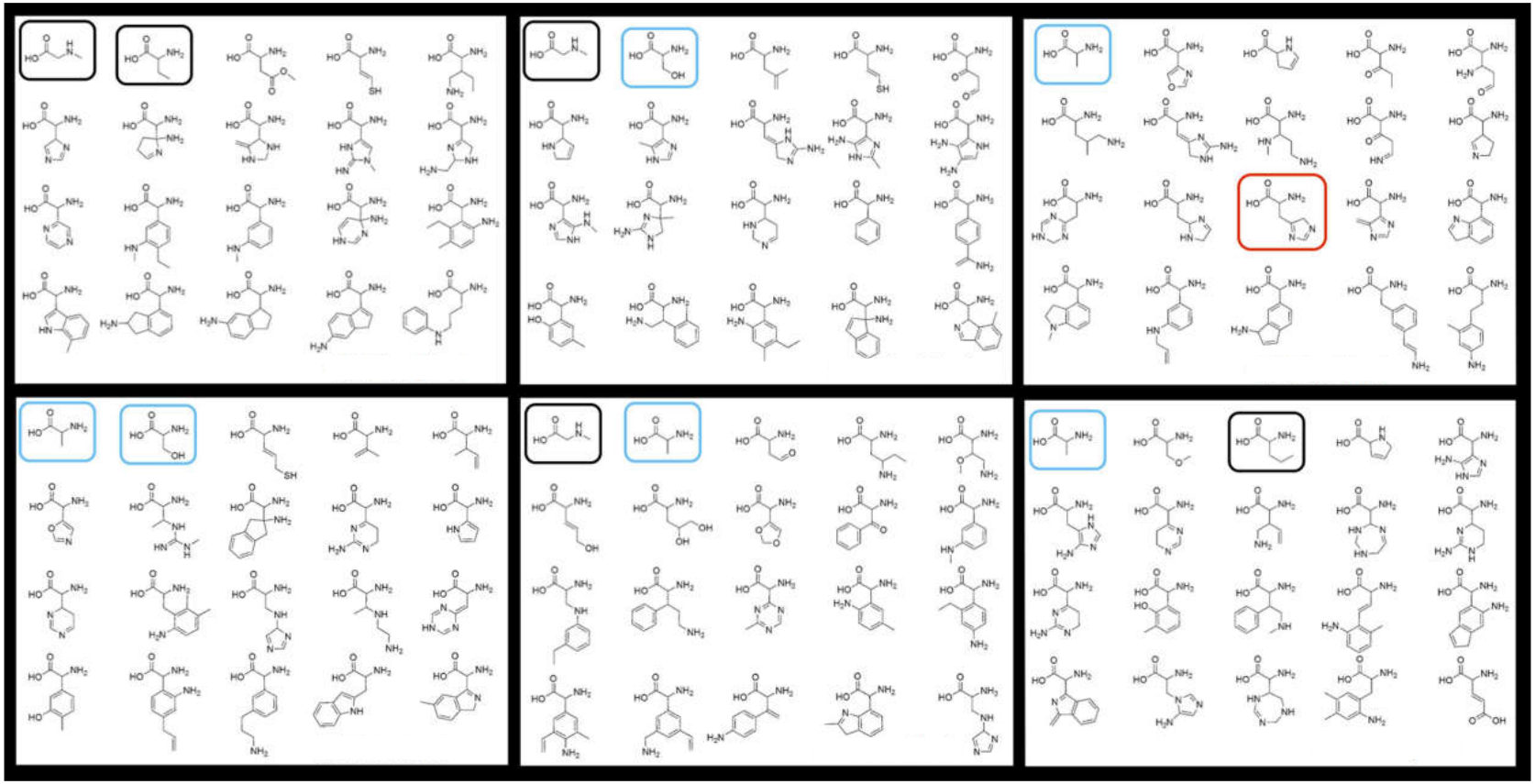
Bakhtiyor Rasulev

- Adaptive analysis gives insight to the adaptive properties of the amino acid alphabet
- Method:
  - sample 10^8 random sets of 20 amino acids from a virtual library of 1913
  - compute *coverage* of chemical space in terms of
    - range and evenness in
    - three dimensions (logP, V_{vdw}, pK_a)
- Results:
  - better sets do exist,
  - but they are rare,
  - and energetically less favorable



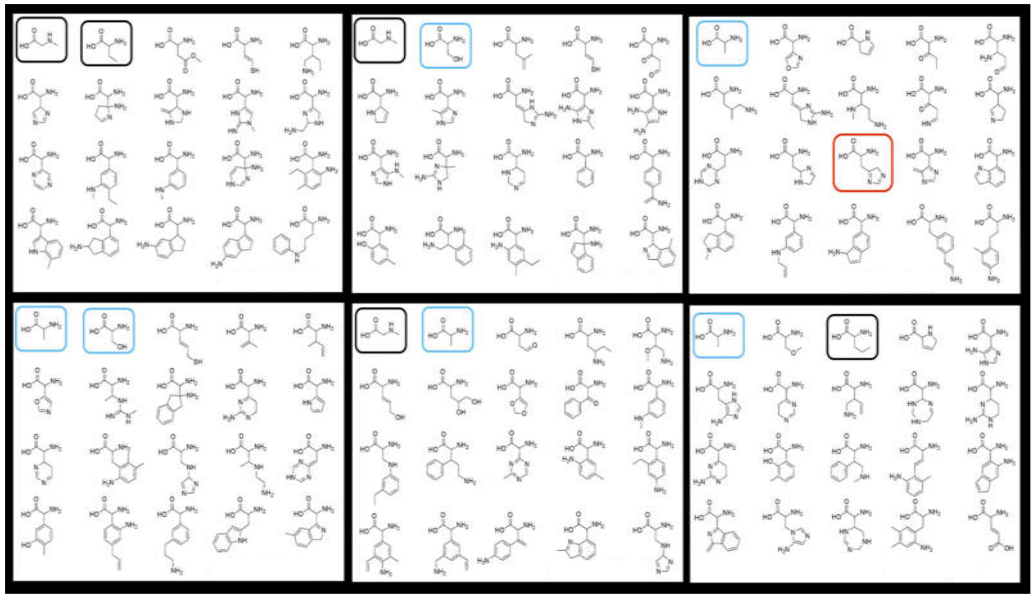
Ilardo et al. Extraordinarily Adaptive Properties of the Genetically Encoded Amino Acids. Sci Rep 5, 9414 (2015)

6 sets with better coverage



black: meteoritic red: encoded blue: both

Simple statistics by basic combinatorics



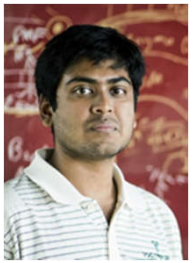
black: meteoritic

red: encoded

blue: both

- 5 of the 6 better sets (~83%) include at least one encoded AA
- the probability that a random set of 20 includes at least one encoded amino acid is only 19%

Latest results: even the subsets of the genetically encoded amino acids show adaptive properties!

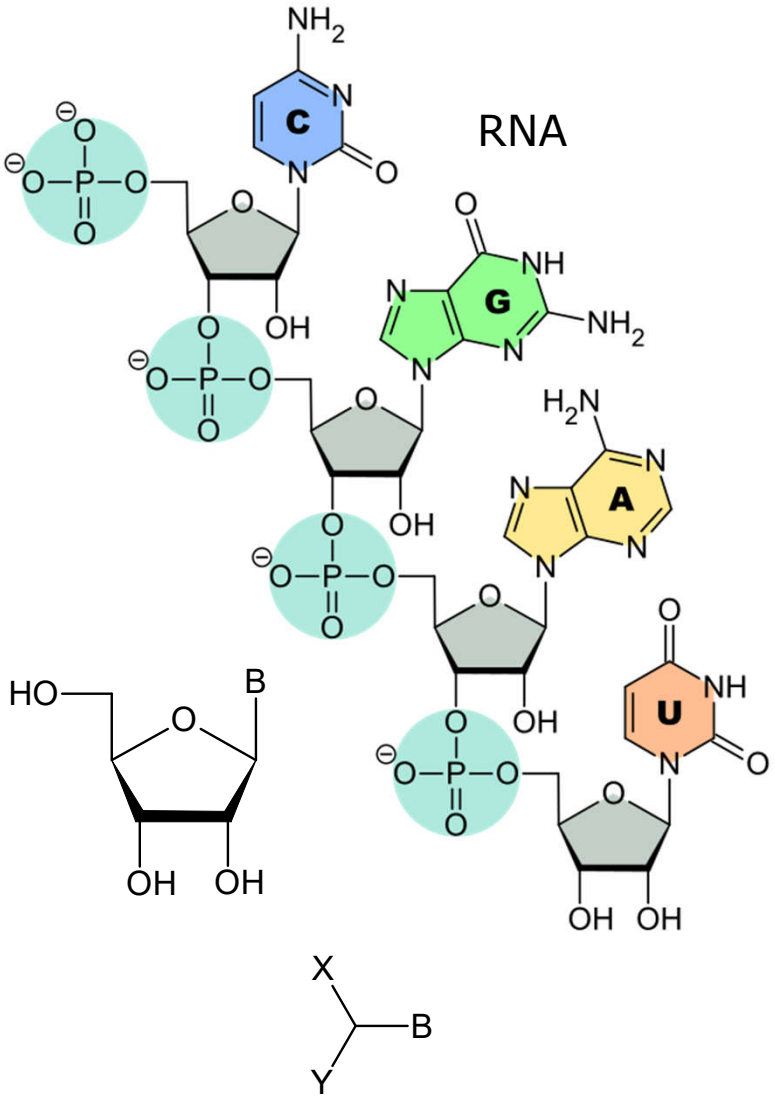


Rudrarup Bose

Ilardo M et al. Adaptive Properties of the Genetically Encoded Amino Acid Alphabet Are Inherited from Its Subsets. Sci. Rep. 9, 12468 (2019)

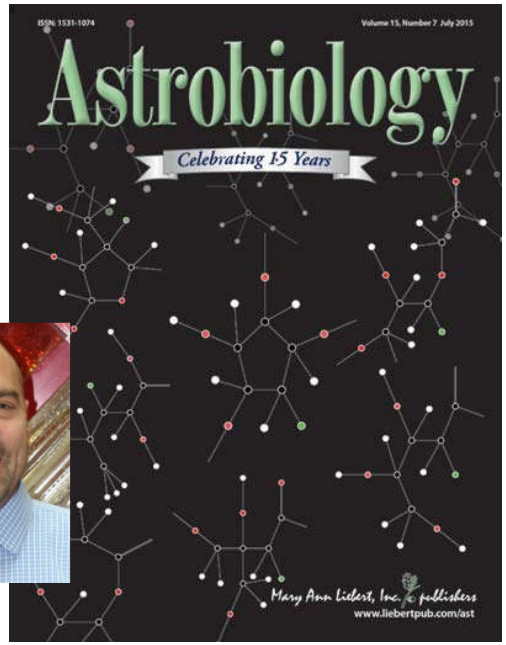
Nucleotides

- Monomeric building blocks of
  - DNA
  - RNA
- Structure
  - linker: phosphate group
  - core: sugar (ribose)
  - base: C, G, A, T or U
- Idea
  - generate isomers of ribose
  - and more general analogues of the core structure
  - analyze the resulting nucleoside libraries

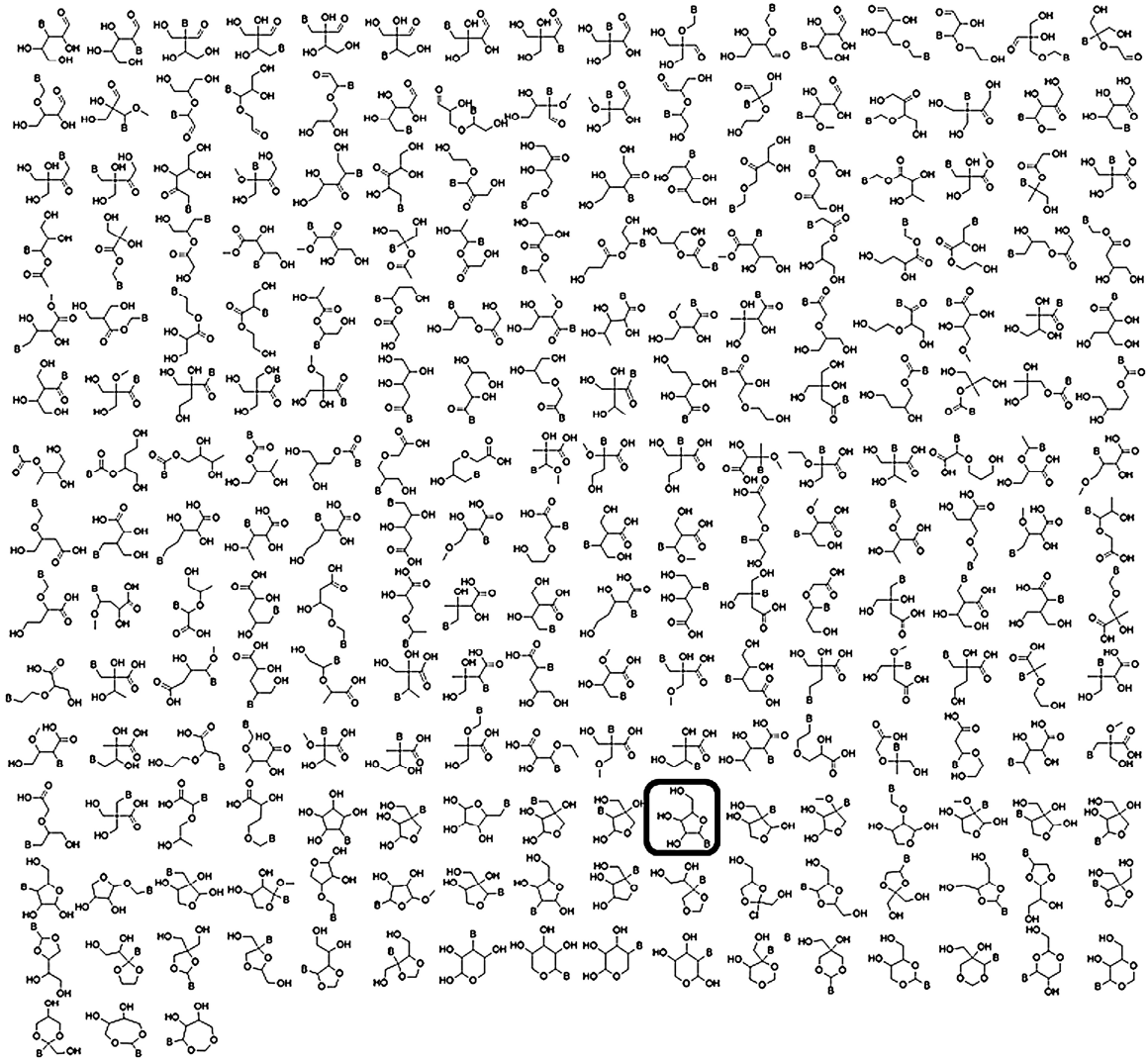


“The 227 faces of RNA”

Isomers of ribose



Jay Goodwin



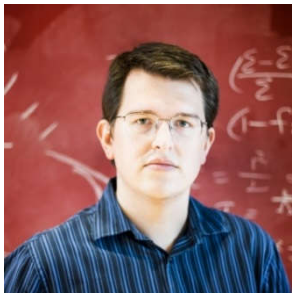
Conclusion:

ribonucleosides may have competed with a multitude of alternative structures

Cleaves HJ, Meringer M, Goodwin J. 227 Views of RNA: Is RNA Unique in Its Chemical Isomer Space? Astrobiology 15(7), 538 (2015)

Chemical space of general nucleosides

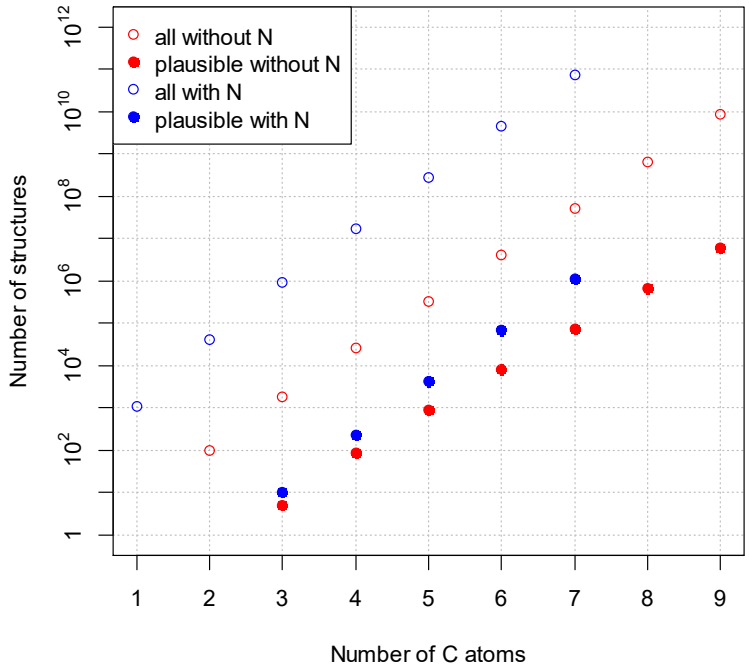
Chris Butch



MOLGEN input

- Formulas
  - C2-7H5-15O[h=0]0-2O[h=1]2-4Cl -sum O=2-4
  - C1-6H5-15N[h=0]0-2N[h=1]0-2N[h=2]0-2O[h=0]0-4O[h=1]0-4Cl -sum N[h=1]+N[h=2]+O[h=1]=2-6 -sum N=1-2 -sum O=0-4
- Rings
  - ringsize 5-10
- Bonds
  - maxbond 2
- Badlist
  - BadHetCl: 2 items
  - BadAaNucList: 181 items
  - BadRingList: 13 items
  - BadAromaticsList: 14 items

Sizes of libraries



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING
Article
Cite This: J. Chem. Inf. Model. 2019, 59, 4266−4277
pubs.acs.org/jcim

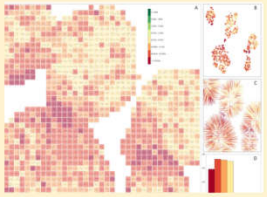
One Among Millions: The Chemical Space of Nucleic Acid-Like Molecules

Henderson James Cleaves, II,*,†,‡,§,# Christopher Butch,†,§,∥,# Pieter Buys Burger,∥ Jay Goodwin,∥ and Markus Meringer⊥,#

†Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-IE-1 Ookayama, Meguro-ku, Tokyo 152-8551, Japan
‡Institute for Advanced Study, Princeton, New Jersey 08540, United States
§Blue Marble Space Institute for Science, 1515 Gallatin St. NW, Washington, DC 20011, United States
∥Department of Chemistry, Emory University, 1515 Dickey Dr., Atlanta, Georgia 30322, United States
⊥German Aerospace Center (DLR), Earth Observation Center (EOC), Münchner Straße 20, 82234 Oberpfaffenhofen-Wessling, Germany

Supporting Information

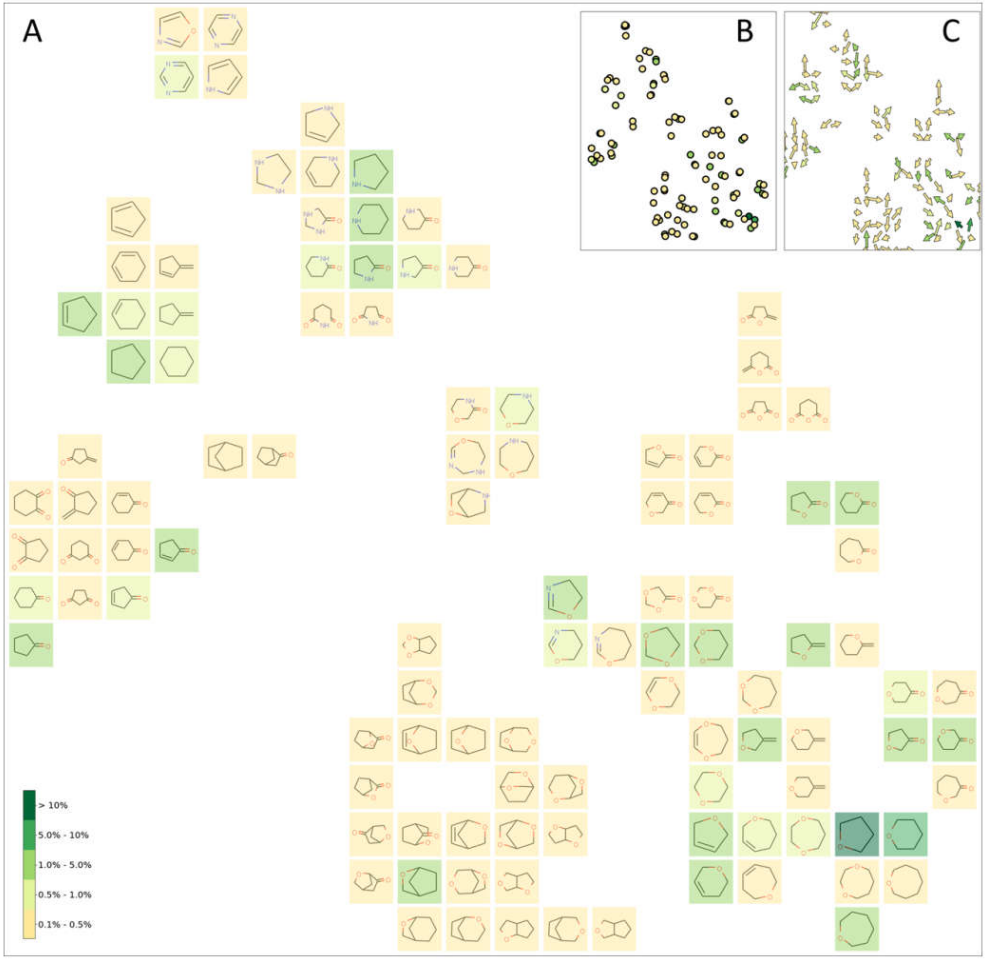
ABSTRACT: Biology encodes hereditary information in DNA and RNA, which are finely tuned to their biological functions and modes of biological production. The central role of nucleic acids in biological information flow makes them key targets of pharmaceutical research. Indeed, other nucleic acid-like polymers can play similar roles to natural nucleic acids both *in vivo* and *in vitro*; yet despite remarkable advances over the last few decades, much remains unknown regarding which structures are compatible with molecular information storage. Chemical space describes the structures and properties of molecules that could exist within a given molecular formula or other classification system. Using structure generation methods, we explore nucleic acid analogues within the formula ranges $BC_{3-7}H_{5-15}O_{2-4}$ and $BC_{3-6}H_{5-15}N_{1-2}O_{0-4}$, where B is a recognition element (e.g., a nucleobase). Other restrictions included two obligatory points of attachment for inclusion into a linear polymer and substructures predicting chemical stability. These sets contain 86,007 (CHO) and 75,309 (CHNO) compositionally isomeric structures, representing 706,568 CHO and 454,422 CHNO stereoisomers, that diversely and densely occupy this space. These libraries point toward there being large spaces of unexplored chemistry relevant to pharmacology and biochemistry and efforts to understand the origins of life.



Analysis includes

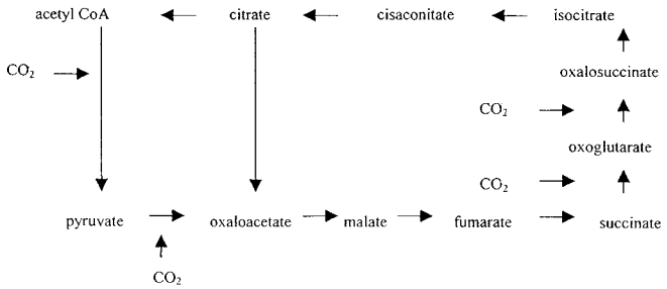
- geometric descriptors
- shape similarity
- synthetic accessibility
- drug-likeness

Potential anti-viral drugs!



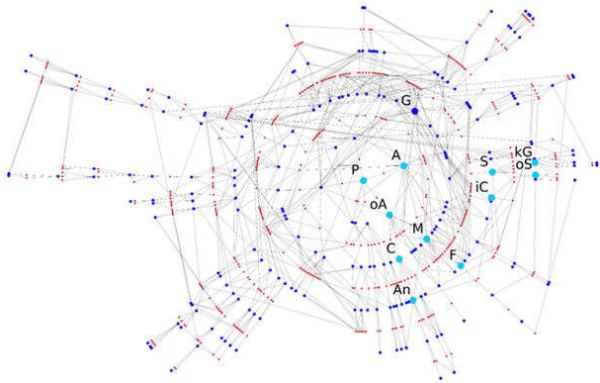
2D-similarity based heatmap of cyclic scaffolds

The rTCA chemical space



Several approaches

- Database search (Morowitz et al, 2000)
  - formulas $C_xH_yO_z$, $1 \le x \le 6$, $1 \le y < 99$, $1 \le z < 99$
    $x/y \le 1$, $y/z \le 2$ for $1 \le x \le 3$,
    $x/y \le 1$, $y/z \le 1.5$ for $4 \le x \le 6$
  - prescribed C=O, forbidden C-O-C, O-O, no cyclic compounds, no triple bonds
  - retrieved 153 hits in Beilstein, including the 11 members of rTCA



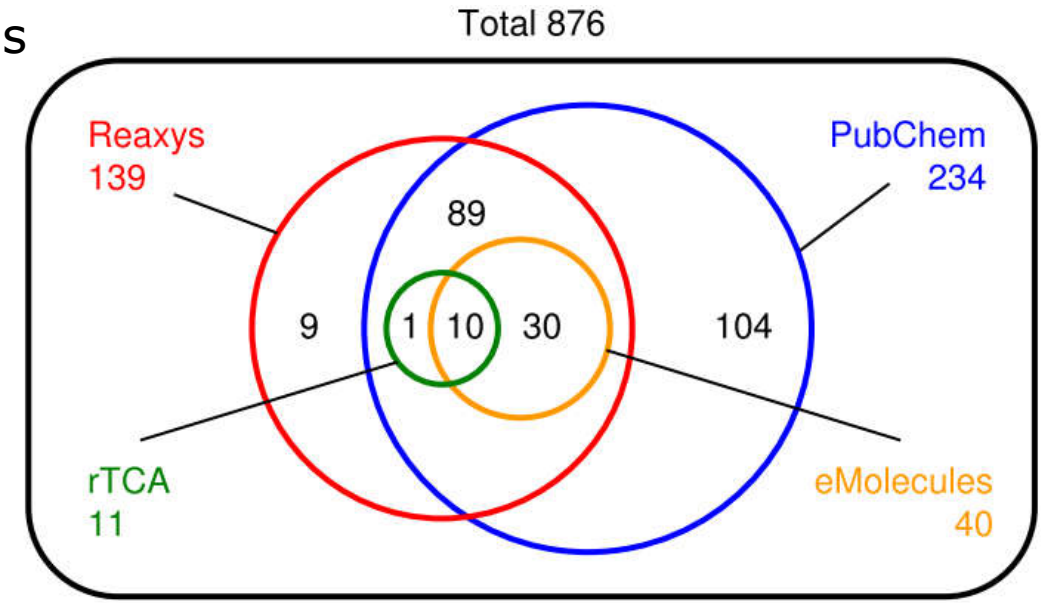
- Reaction-based structure generation (Zubarev et al, 2015)
  - 7 reaction types
  - recursively applied until all 11 rTCA compounds were generated (reaction network)
  - delivered a total of 175 structures (actually 221)

Morowitz HJ, Kostelnik JG, Yang J, Cody GD: The origin of intermediary metabolism. PNAS 97(14), 7704 (2000)
Zubarev DY, Rappoport DR, Aspuru-Guzikk, A: Uncertainty of Prebiotic Scenarios: The Case of Non-Enzymatic Reverse Tricarboxylic Acid Cycle. Scientific Reports 5, 8009 (2015)

Exhaustive enumeration of the rTCA chemical space

Third approach:

- Formula-based structure generation
  - Morowitz rules can almost directly be used as input for MOLGEN
  - additional constraints to exclude hydrates and enols
  - generated 876 structures
  - overlap with Morowitz set: 119
  - overlap with Zubarev set: 70
  - overlap with current databases …



Perspective:

- Search autocatalytic cycles in generated set(s)

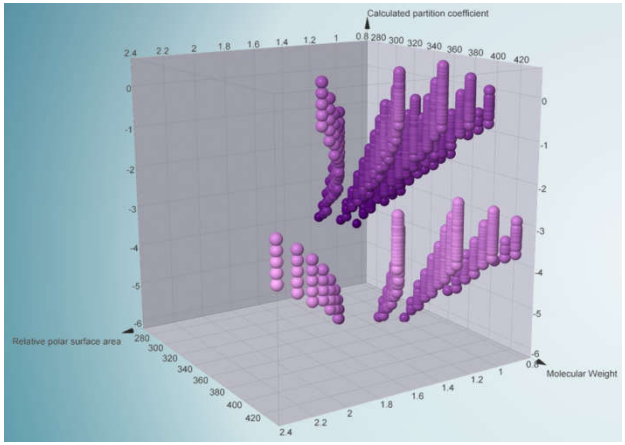
Meringer, M., Cleaves, H.J. Computational exploration of the chemical structure space of possible reverse tricarboxylic acid cycle constituents. Sci Rep 7, 17540 (2017)

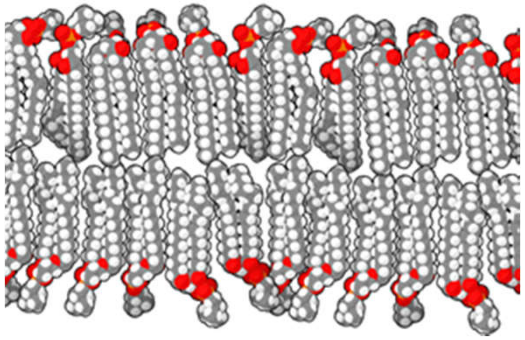
… more chemical space work in progress …

Ric Gillams

Selene Forget

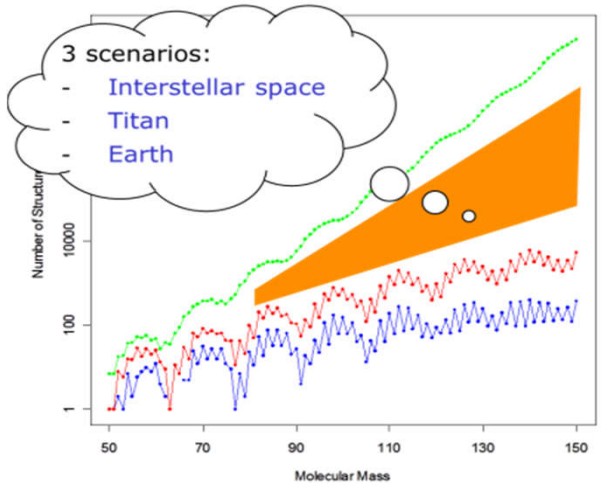
- libraries of lipids and their potential to form bilayers

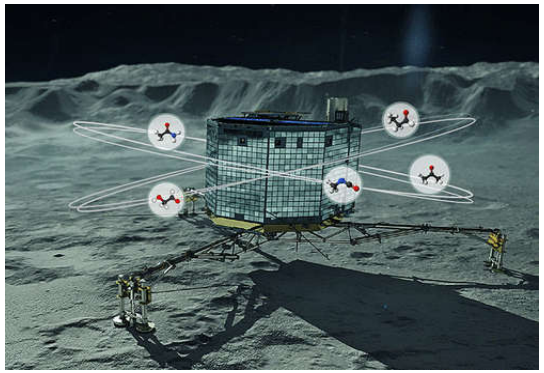




- general small molecule space and exploration missions

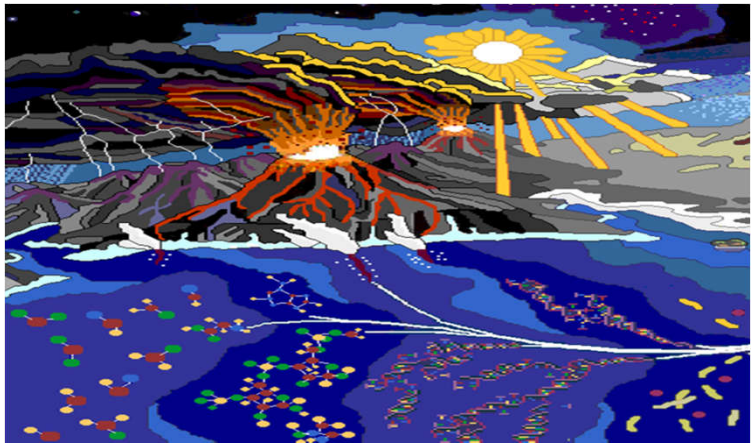
Chaitanya Giri



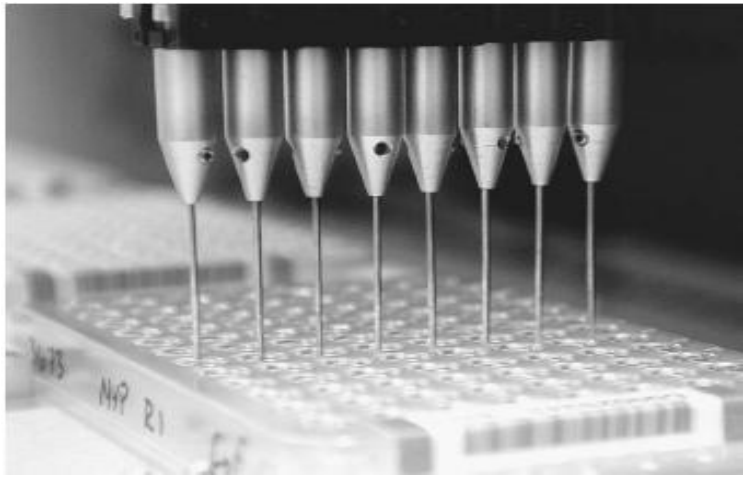


Meringer M, Giri C, Cleaves HJ. Fitting Cometary Sampling and Composition Mass Spectral Results Using Non-negative Least Squares: Reducing Detection Ambiguity for In Situ Solar System Organic Compound Measurements. ACS Earth Space Chem 2, 1256-1261 (2018)

Can this help to solve the mystery about Earth's gigantic prebiotic combinatorial chemistry experiment that led to the origin of life?



"This is one of science's great unsolved problems that is bound to get much more attention in the near future and turn much more computational as everything else in the biosciences." - Jotun Hein



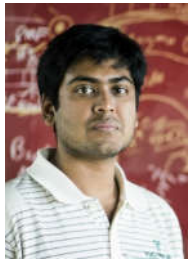
Acknowledgements

…to the contributors…

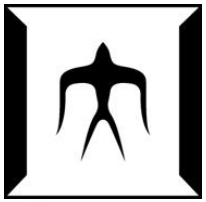


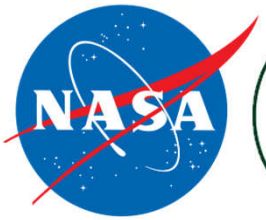


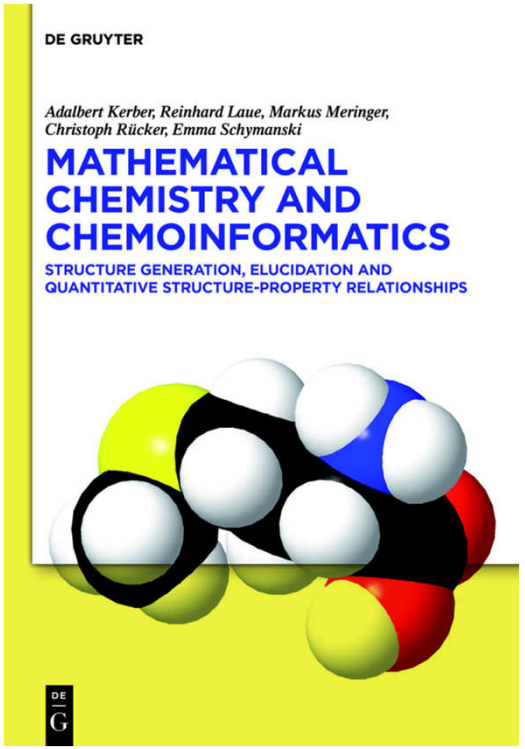


…the institutions…



… to the MOLGEN team
former Mathematics II
University of Bayreuth
www.molgen.de

THANKS FOR YOUR ATTENTION!