Markus Meringer

50 Years of Chemical Space Exploration Through Computation

5 Years with Focus on Biomolecules

Gordon Research Conference

Origins of Life

Galveston, January 17-22, 2016

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

What precisely are we talking about?

Chemical Space is the space spanned by all possible stable chemical compounds – this is all combinations of atomic nuclei, in all possible topology isomers. [adapted from Wikipedia]

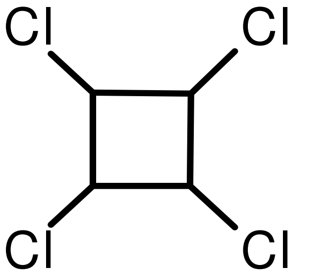
Different levels of abstraction for representing a molecule:

composition

$C_4H_4Cl_4$

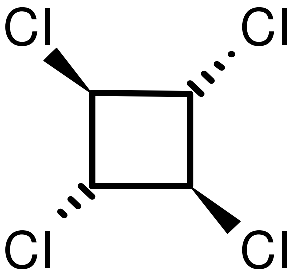
molecular formula

constitution

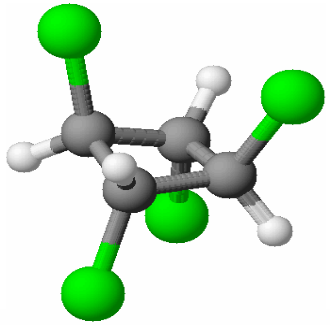


structural formula

configuration



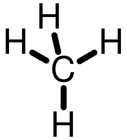
conformation



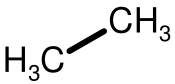
From compositions to constitutions

Example: Alkanes C_nH_{2n+2}

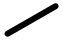
CH_4



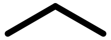
C_2H_6



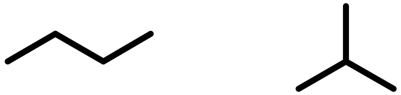
or simply



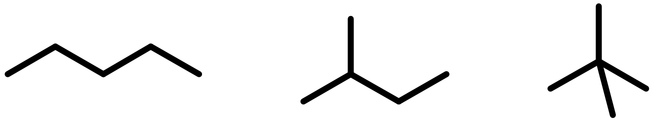
C_3H_8



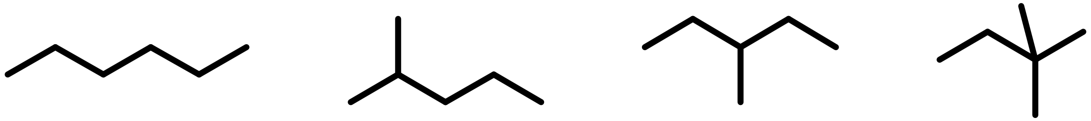
C_4H_{10}



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!

The DENDRAL project

- driven by exiobiologist J. Lederberg
- initiated 50 years ago (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

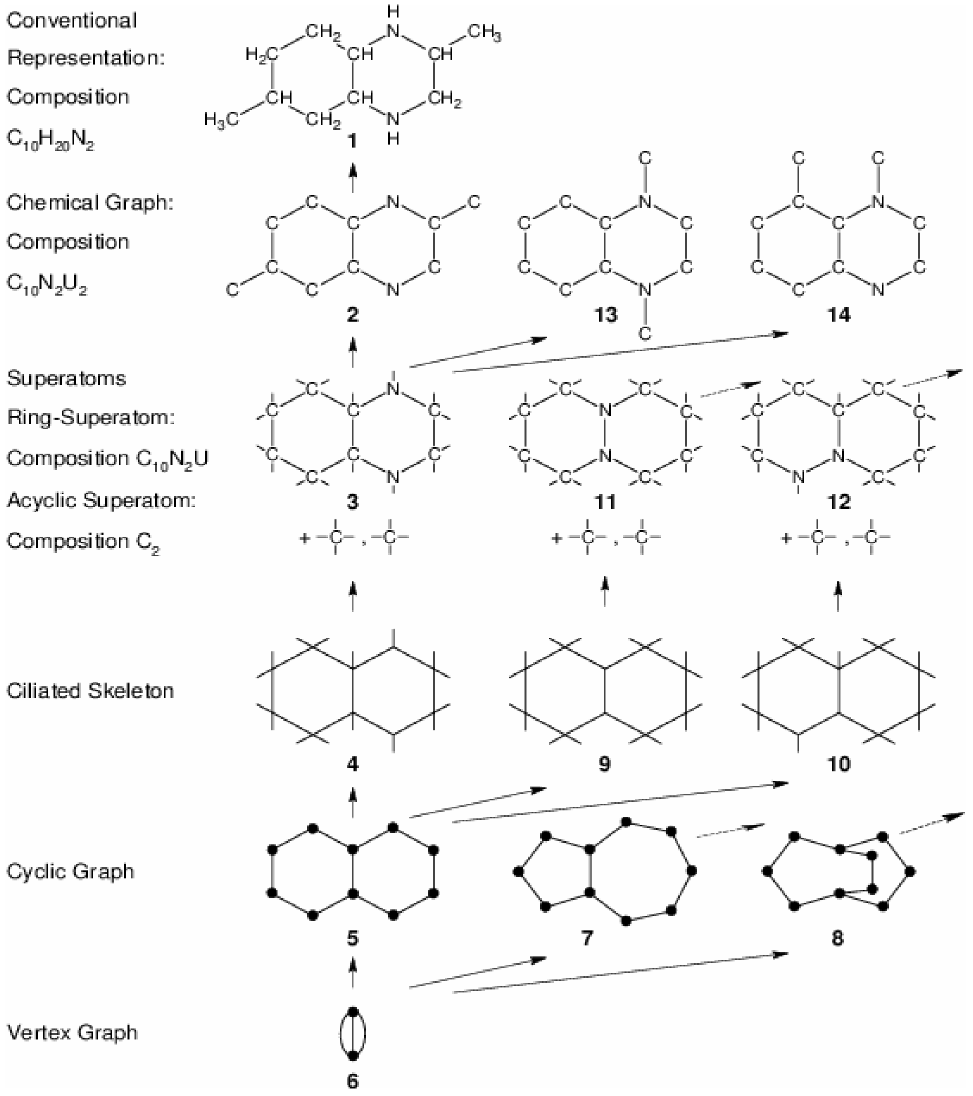
remove hydrogen

decompose into superatoms

strip element symbols

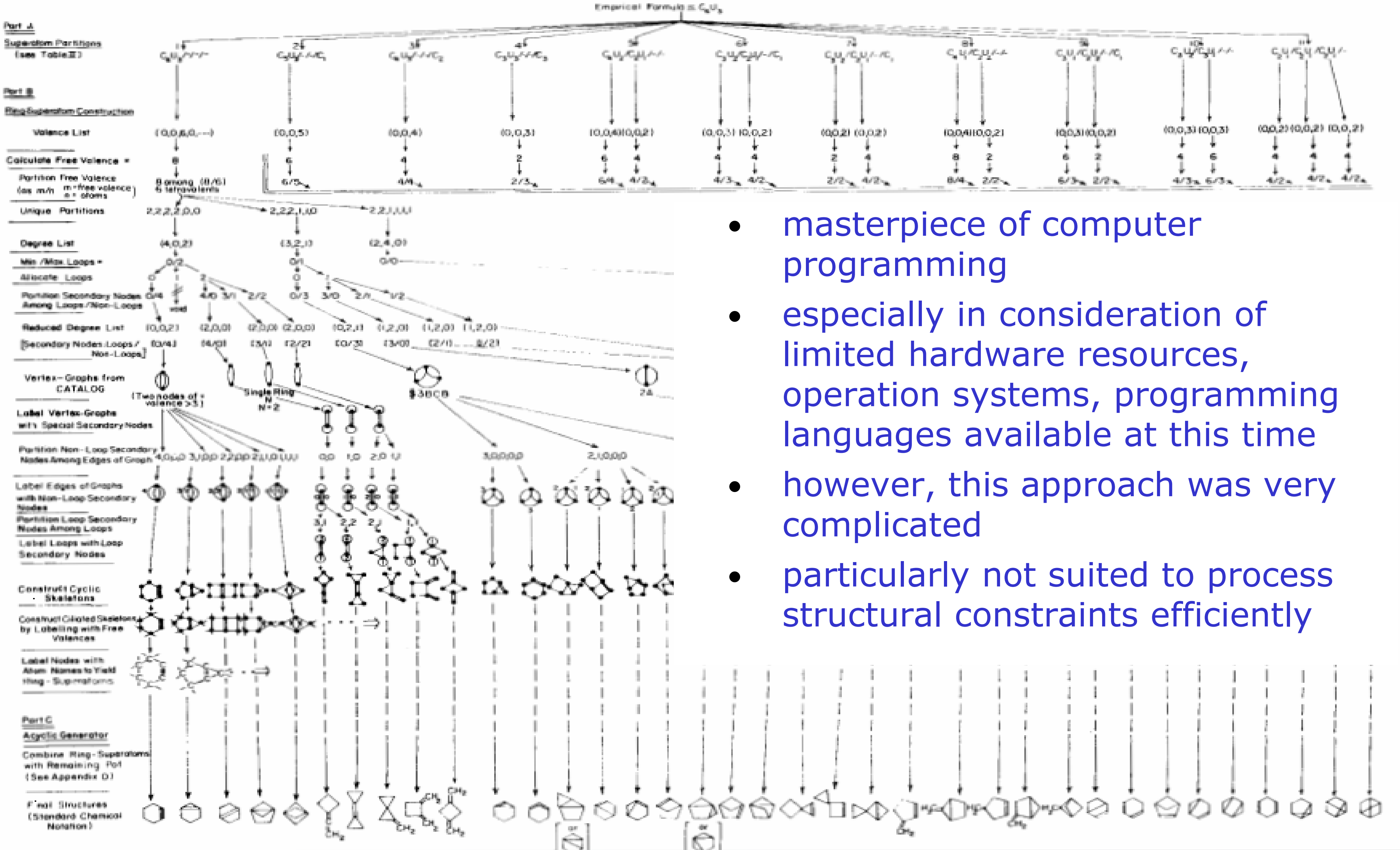
delete free valencies

replace chains of
bivalent nodes by edes



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

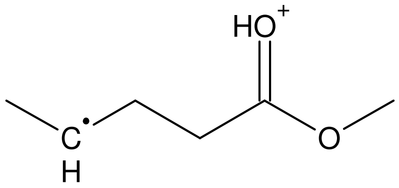

Empirical Formula = C6H10
Part A
Superatom Partitions (see Table II)
Part B
Ring Superatom Construction
Valence List
Calculate Free Valence *
Partition Free Valence (as m/n m = free valence n = atoms)
8 among (8/6) 6 tetravalents
Unique Partitions
Degree List
Min / Max Loops *
Allocate Loops
Partition Secondary Nodes Among Loops / Non-Loops
void
Reduced Degree List
[Secondary Nodes Loops / Non-Loops]
Vertex-Graphs from CATALOG
(Two nodes of valence > 3)
Single Ring N N = 2
Label Vertex-Graphs with Special Secondary Nodes
Partition Non-Loop Secondary Nodes Among Edges of Graph
Label Edges of Graphs with Non-Loop Secondary Nodes
Partition Loop Secondary Nodes Among Loops
Label Loops with Loop Secondary Nodes
Construct Cyclic Skeletons
Construct Ciliated Skeletons by Labelling with Free Valences
Label Nodes with Atom Names to Yield Ring-Superatoms
Part C
Acyclic Generator
Combine Ring-Superatoms with Remaining Part (See Appendix D)
Final Structures (Standard Chemical Notation)

- 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

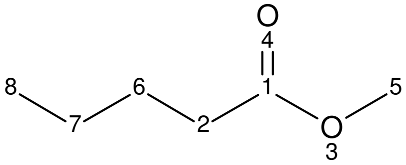
- Chemical compounds as molecular graphs

  vertices and edges (simple graph)
  + bonds 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

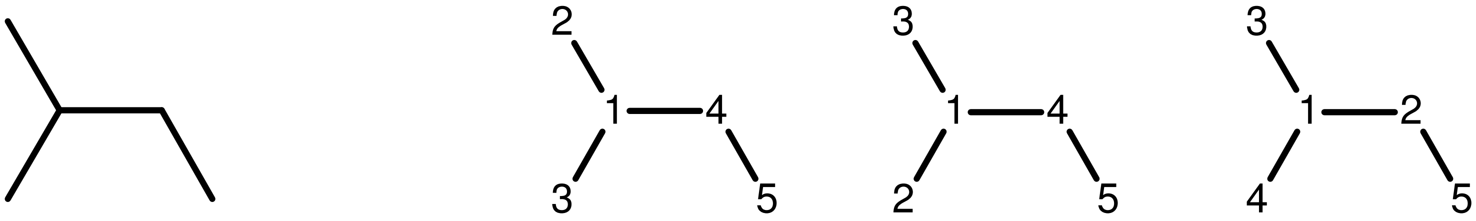
	1	2	3	4	5	6	7	8
1	0	1	1	2	0	0	0	0
2	1	0	0	0	0	1	0	0
3	1	0	0	0	1	0	0	0
4	2	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0
6	0	1	0	0	0	0	1	0
7	0	0	0	0	0	1	0	1
8	0	0	0	0	0	0	1	0

- Idea: fill adjacency matrix in all possible ways

Houston, we have a problem!

Chemical compounds

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



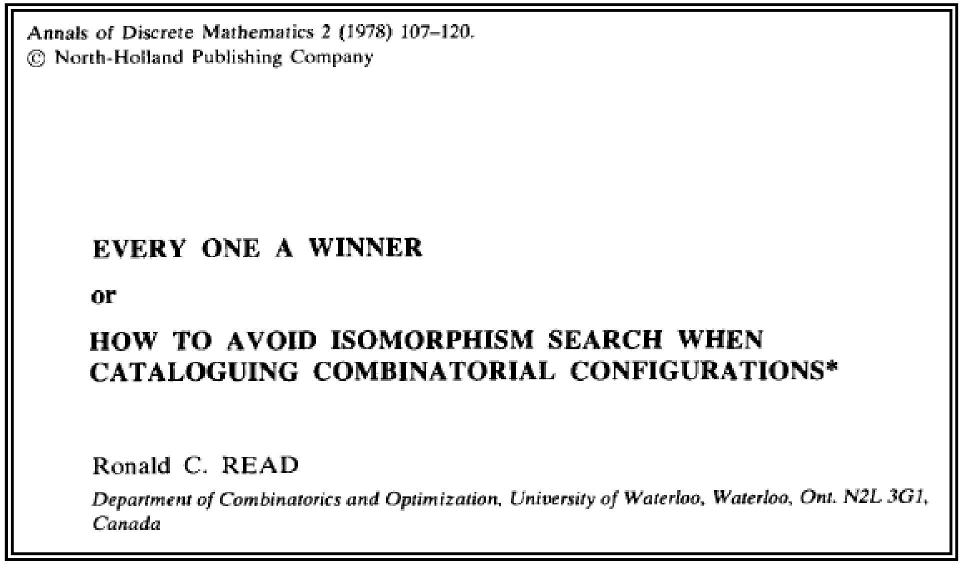
leads to problems

- 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

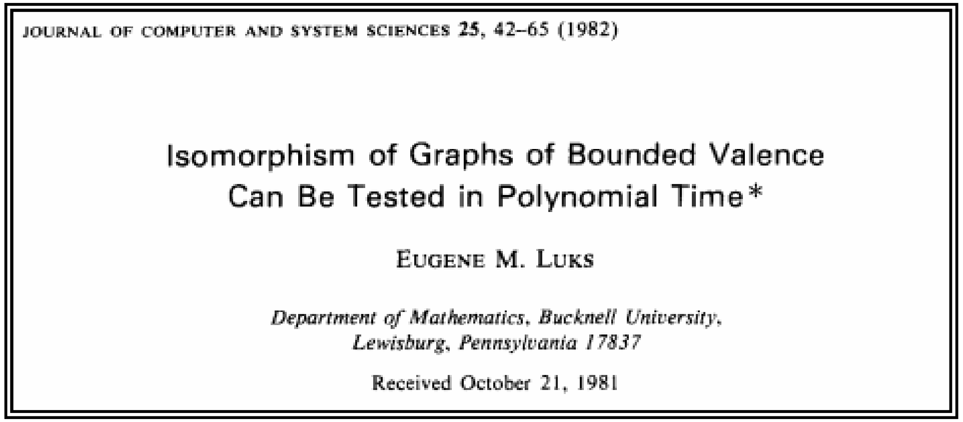
HOW TO AVOID ISOMORPHISM SEARCH WHEN CATALOGUING COMBINATORIAL CONFIGURATIONS*

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

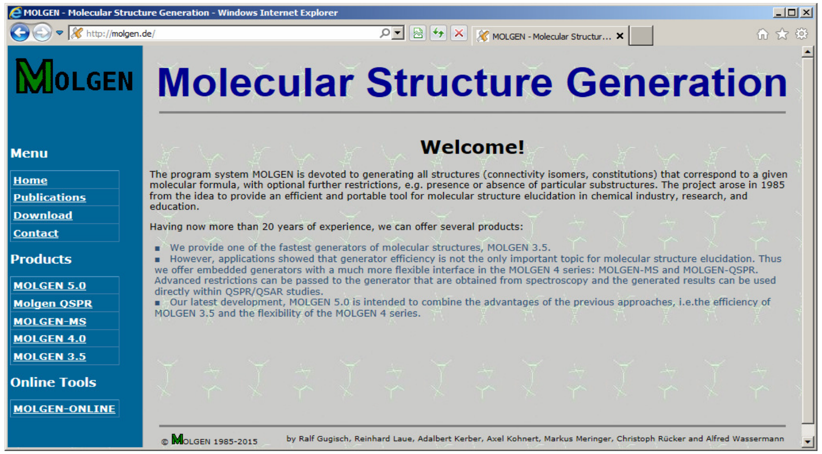
Received October 21, 1981

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)

based on “orderly generation”



Molecular Structure Generation

Welcome!

The program system MOLGEN is devoted to generating all structures (connectivity isomers, constitutions) that correspond to a given molecular formula, with optional further restrictions, e.g. presence or absence of particular substructures. The project arose in 1985 from the idea to provide an efficient and portable tool for molecular structure elucidation in chemical industry, research, and education.

Having now more than 20 years of experience, we can offer several products:

- We provide one of the fastest generators of molecular structures, MOLGEN 3.5.
- However, applications showed that generator efficiency is not the only important topic for molecular structure elucidation. Thus we offer embedded generators with a much more flexible interface in the MOLGEN 4 series: MOLGEN-MS and MOLGEN-QSPR. Advanced restrictions can be passed to the generator that are obtained from spectroscopy and the generated results can be used directly within QSPR/QSAR studies.
- Our latest development, MOLGEN 5.0 is intended to combine the advantages of the previous approaches, i.e.the efficiency of MOLGEN 3.5 and the flexibility of the MOLGEN 4 series.

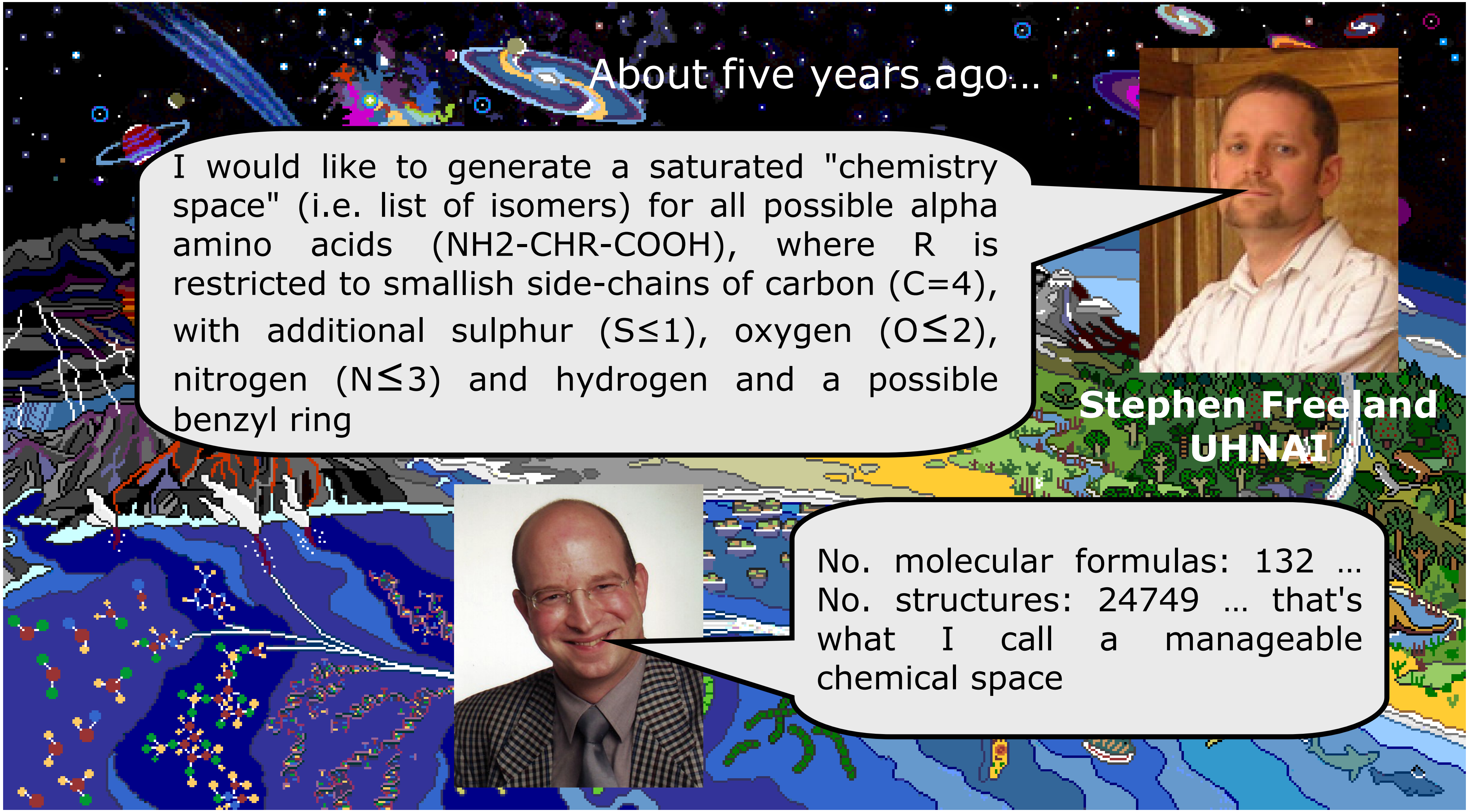
© MOLGEN 1985-2015 by Ralf Gugisch, Reinhard Laue, Adalbert Kerber, Axel Kohnert, Markus Meringer, Christoph Rücker and Alfred Wassermann

Computational examples:

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.

Crossing disciplinary boundaries



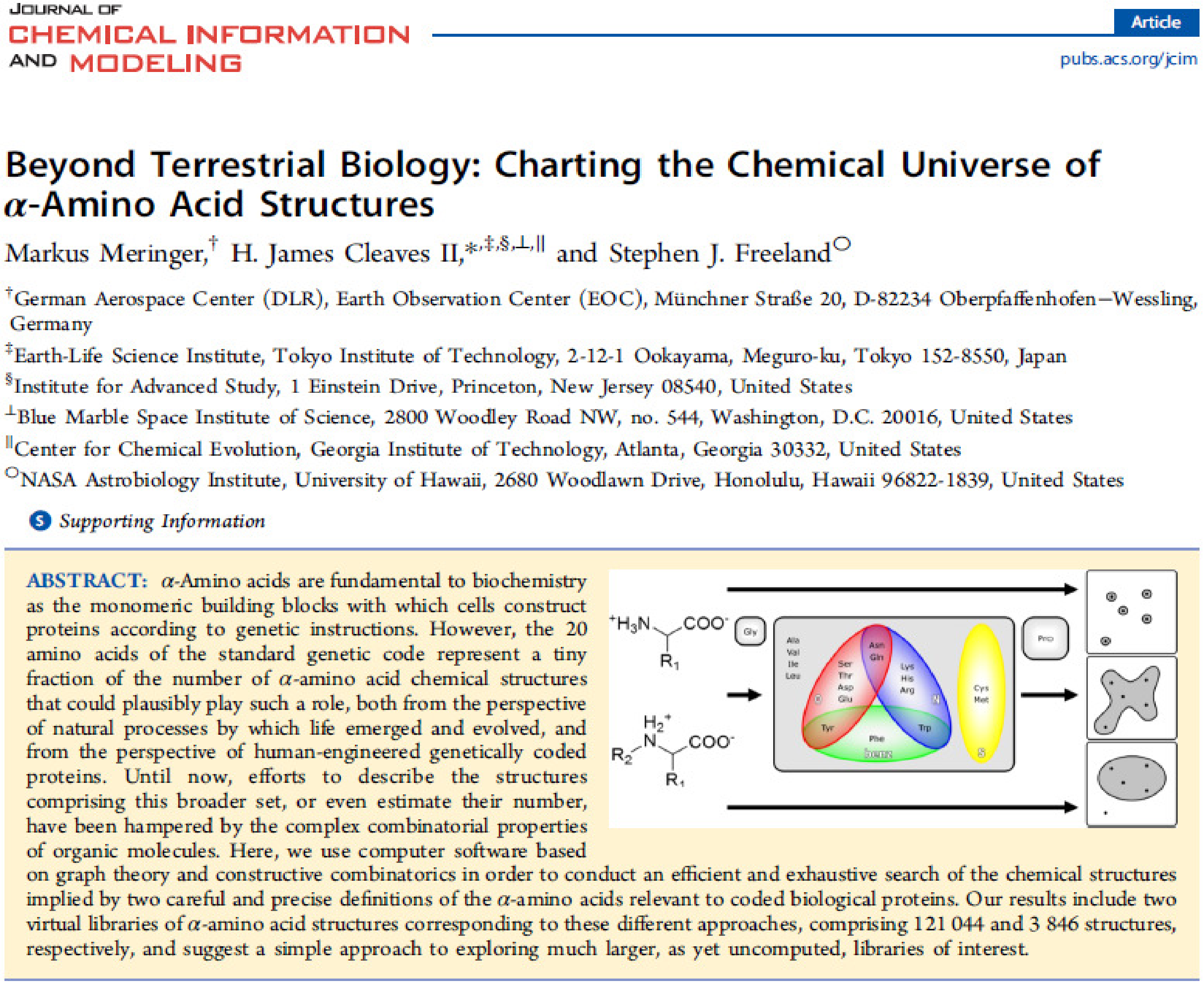
About five years ago...

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



JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

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

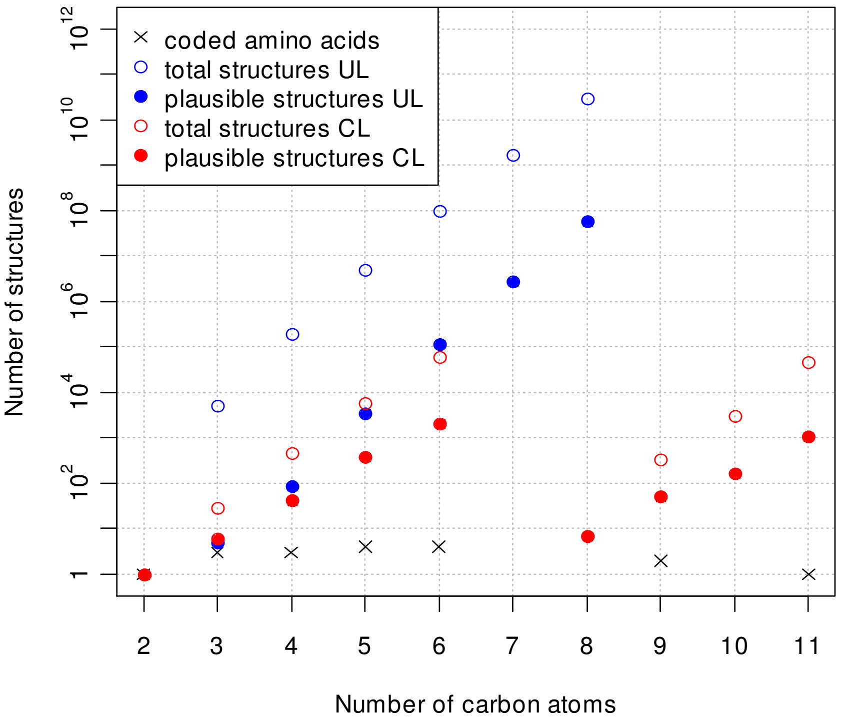
[⊥] Blue Marble Space Institute of Science, 2800 Woodley Road NW, no. 544, Washington, D.C. 20016, United States

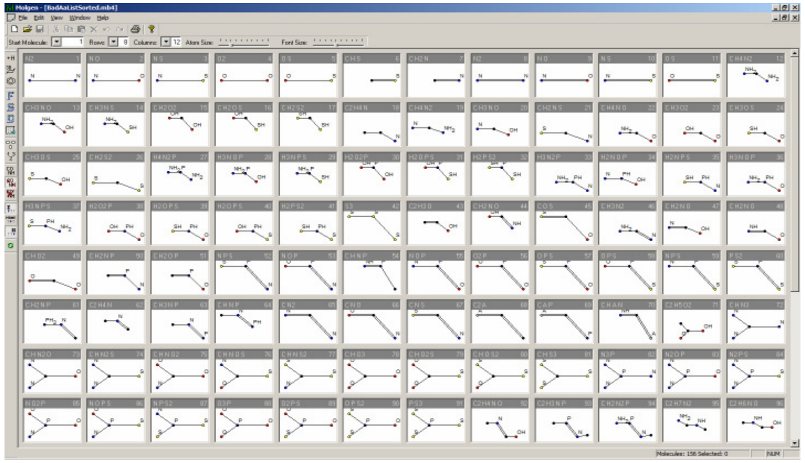
[∥] Center for Chemical Evolution, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

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

S *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.





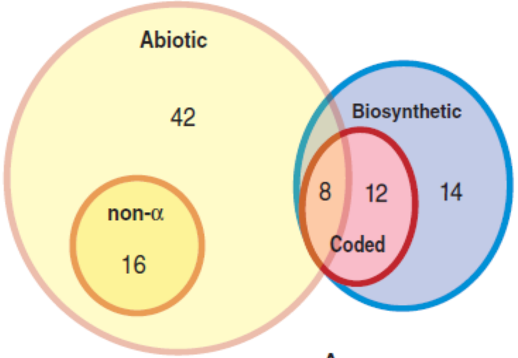
156-membered badlist

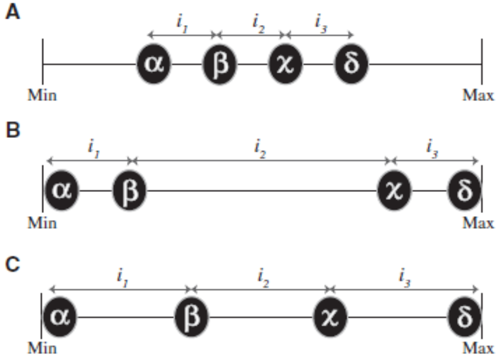
Poster 28

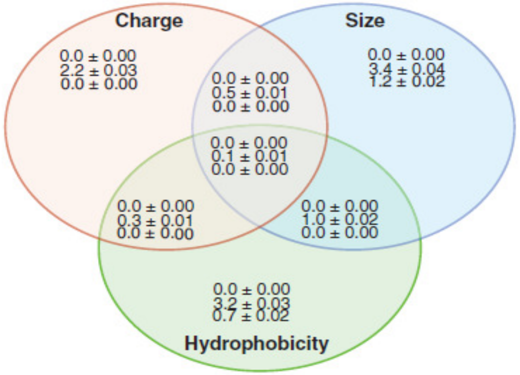
Application:
Verify a model on selection of the amino acid alphabet

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

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





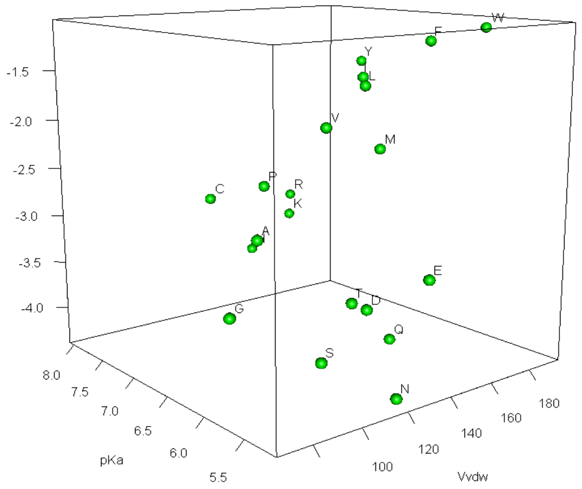


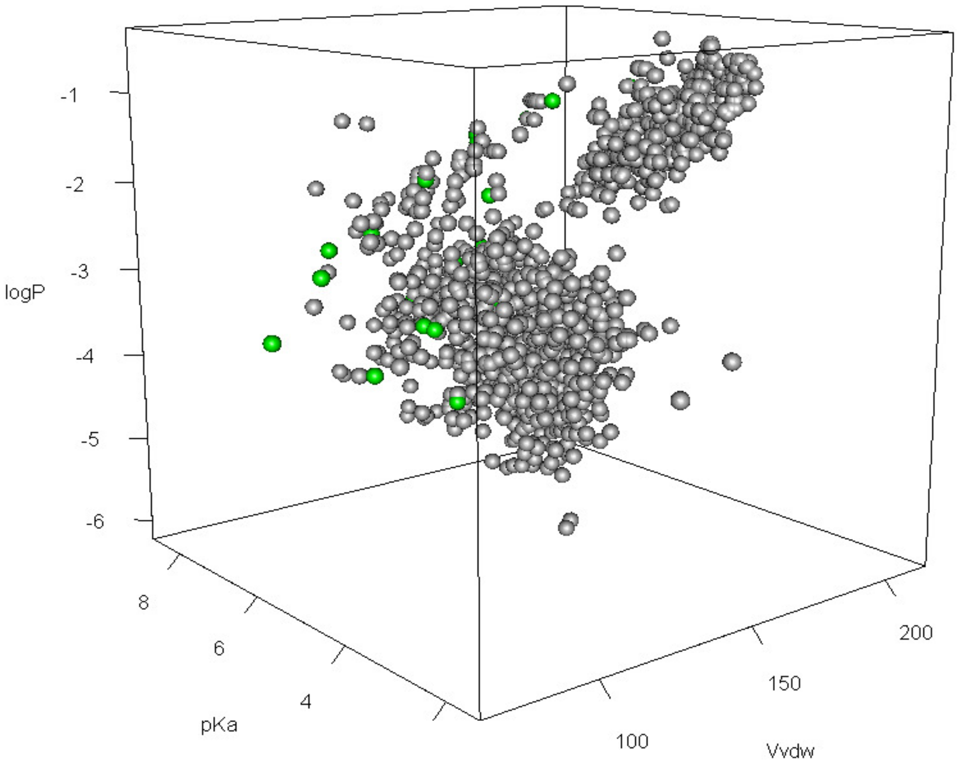
Philip GK, Freeland SJ: Did evolution select a nonrandom "alphabet" of amino acids? Astrobiology 11(3), 235 (2011)

… research continued at ELSI …

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

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





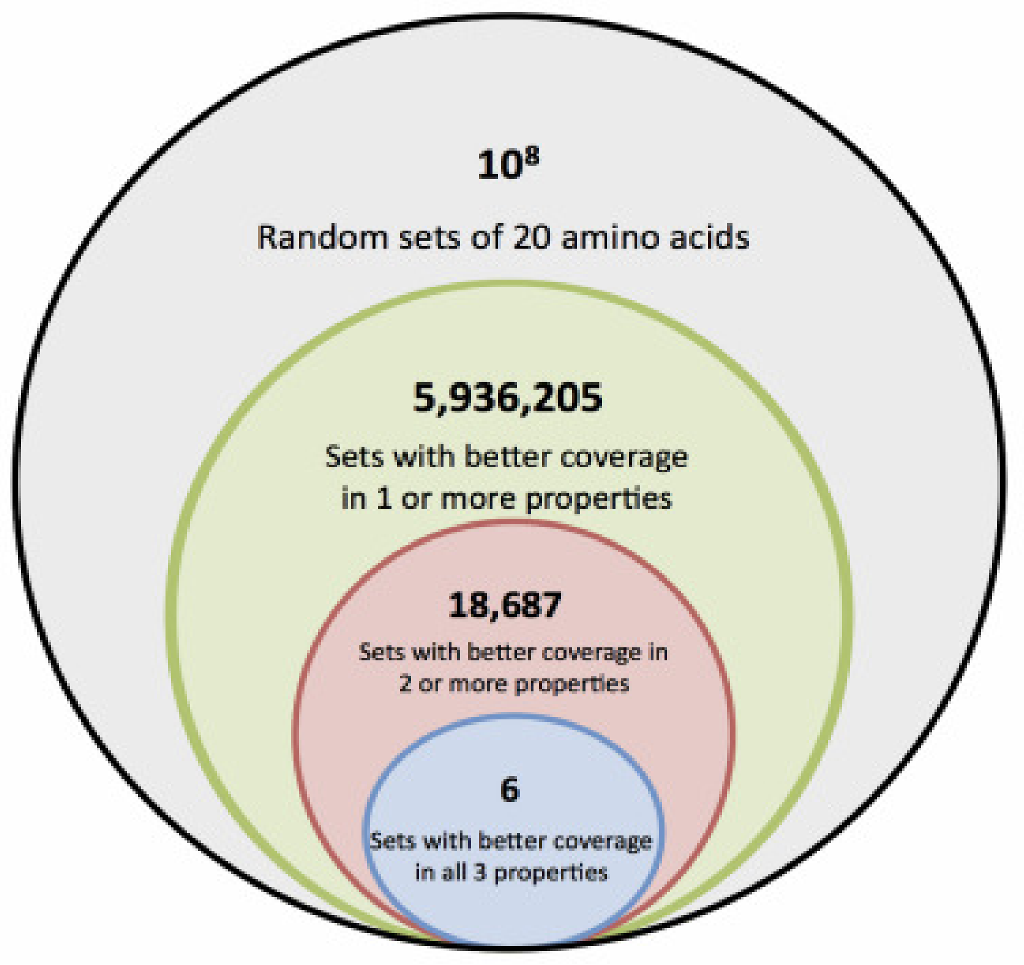
20 biologically encoded amino acids colored green

Statistical analysis

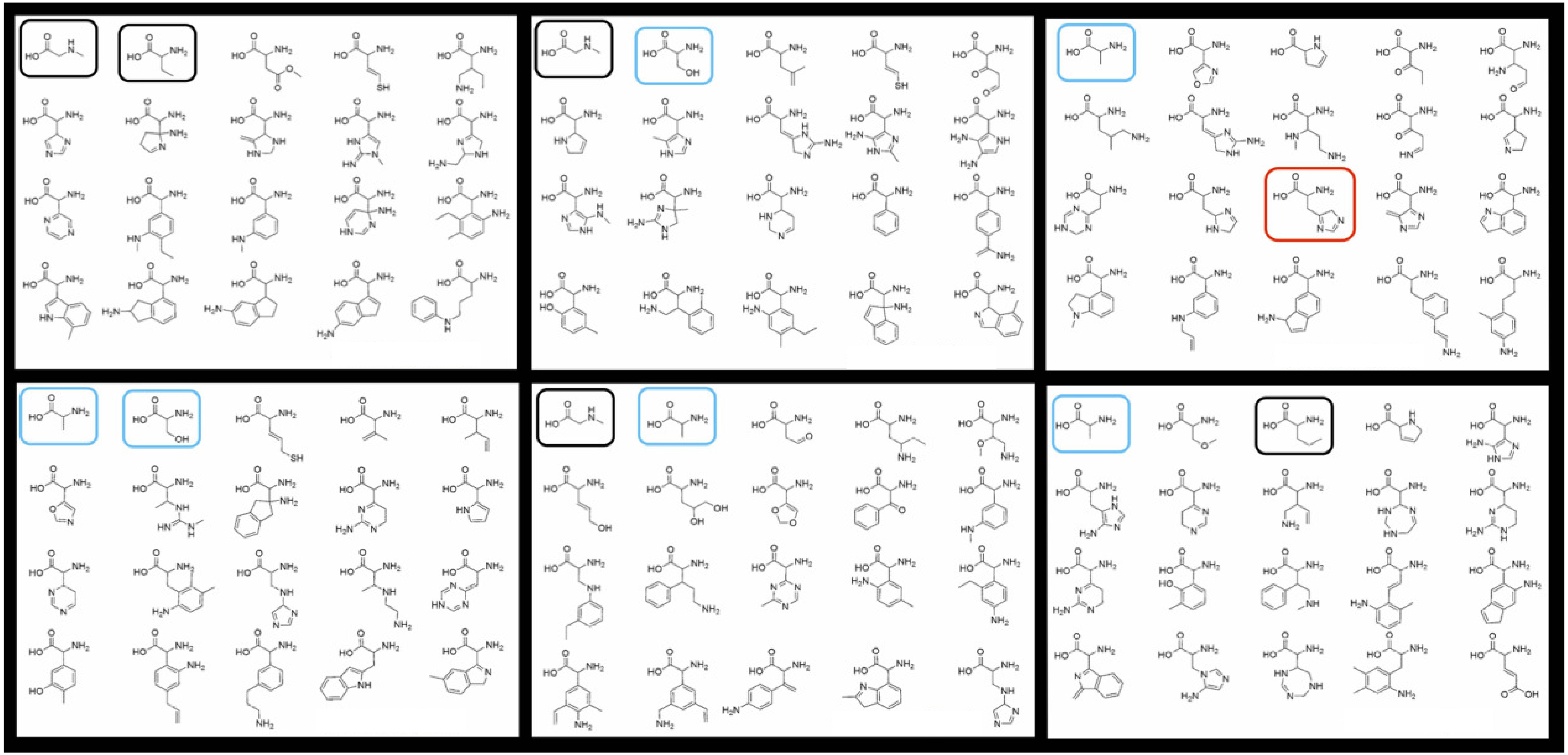
- Adaptive analysis gives insight to the adaptive properties of the amino acid alphabet

- Method:
  - sampling 10^8 random sets of 20 amino acids
  - comparing *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

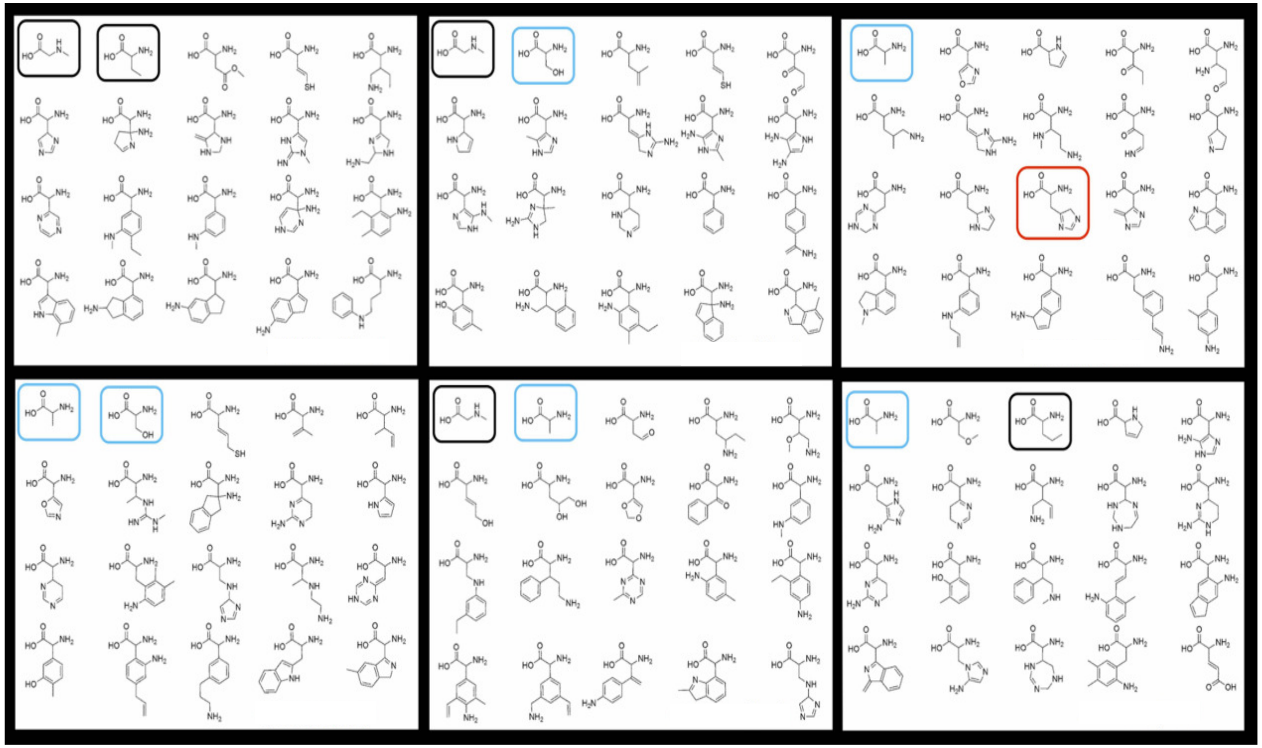

10^8
Random sets of 20 amino acids
5,936,205
Sets with better coverage in 1 or more properties
18,687
Sets with better coverage in 2 or more properties
6
Sets with better coverage in all 3 properties

6 sets with better coverage



black: meteoritic red: encoded blue: both

Simple statistics by basic combinatorics



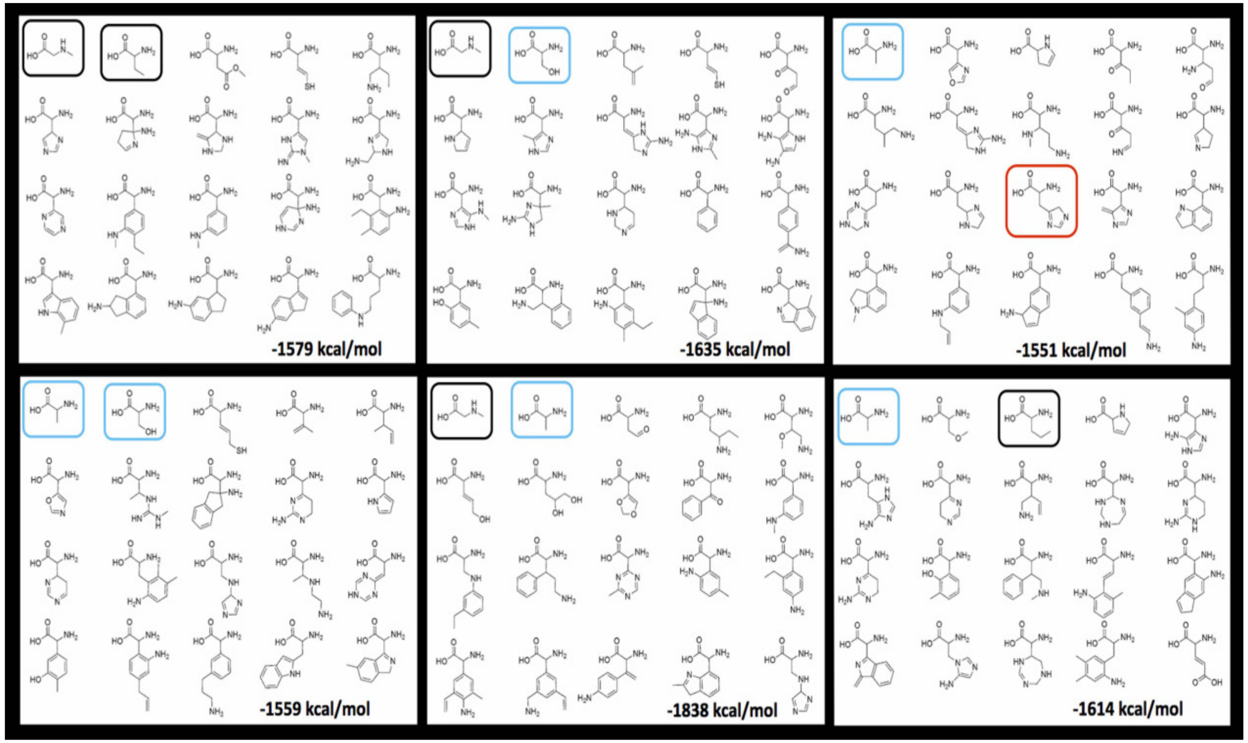
black: meteoritic

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%
- similar situation for meteoritic amino acids

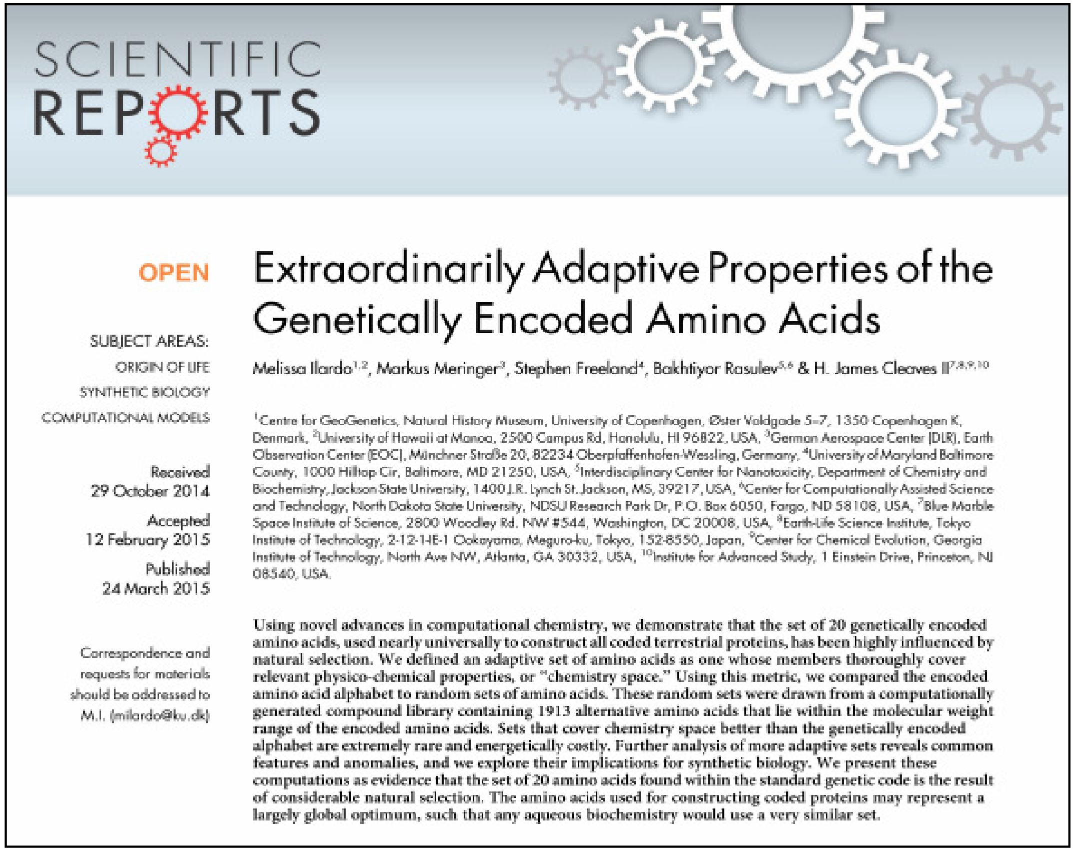
Heats of formation $\Delta H_f°$



- sums of ΔHf° for the encoded set is -2306 kcal/mol
- clearly below the sums for the sets with better coverage
- this additional criterion
  - improves the original model
  - to make the encoded set unique again

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

Results published last year



SCIENTIFIC REPORTS

OPEN

Extraordinarily Adaptive Properties of the Genetically Encoded Amino Acids

SUBJECT AREAS:
ORIGIN OF LIFE
SYNTHETIC BIOLOGY
COMPUTATIONAL MODELS

Received
29 October 2014

Accepted
12 February 2015

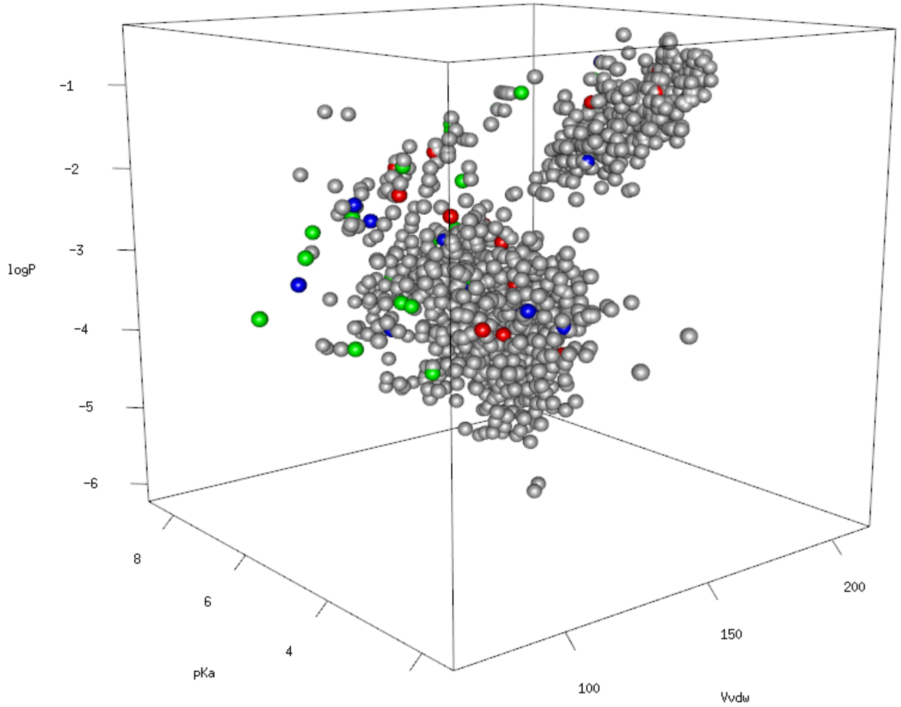
Published
24 March 2015

Melissa Ilardo[1,2], Markus Meringer[3], Stephen Freeland[4], Bakhtiyor Rasulev[5,6] & H. James Cleaves II[7,8,9,10]

[1]Centre for GeoGenetics, Natural History Museum, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark, [2]University of Hawaii at Manoa, 2500 Campus Rd, Honolulu, HI 96822, USA, [3]German Aerospace Center (DLR), Earth Observation Center (EOC), Münchner Straße 20, 82234 Oberpfaffenhofen-Wessling, Germany, [4]University of Maryland Baltimore County, 1000 Hilltop Cir, Baltimore, MD 21250, USA, [5]Interdisciplinary Center for Nanotoxicity, Department of Chemistry and Biochemistry, Jackson State University, 1400 J.R. Lynch St. Jackson, MS, 39217, USA, [6]Center for Computationally Assisted Science and Technology, North Dakota State University, NDSU Research Park Dr, P.O. Box 6050, Fargo, ND 58108, USA, [7]Blue Marble Space Institute of Science, 2800 Woodley Rd. NW #544, Washington, DC 20008, USA, [8]Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-1-IE-1 Ookayama, Meguro-ku, Tokyo, 152-8550, Japan, [9]Center for Chemical Evolution, Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332, USA, [10]Institute for Advanced Study, 1 Einstein Drive, Princeton, NJ 08540, USA.

Correspondence and requests for materials should be addressed to M.I. (milardo@ku.dk)

Using novel advances in computational chemistry, we demonstrate that the set of 20 genetically encoded amino acids, used nearly universally to construct all coded terrestrial proteins, has been highly influenced by natural selection. We defined an adaptive set of amino acids as one whose members thoroughly cover relevant physico-chemical properties, or "chemistry space." Using this metric, we compared the encoded amino acid alphabet to random sets of amino acids. These random sets were drawn from a computationally generated compound library containing 1913 alternative amino acids that lie within the molecular weight range of the encoded amino acids. Sets that cover chemistry space better than the genetically encoded alphabet are extremely rare and energetically costly. Further analysis of more adaptive sets reveals common features and anomalies, and we explore their implications for synthetic biology. We present these computations as evidence that the set of 20 amino acids found within the standard genetic code is the result of considerable natural selection. The amino acids used for constructing coded proteins may represent a largely global optimum, such that any aqueous biochemistry would use a very similar set.



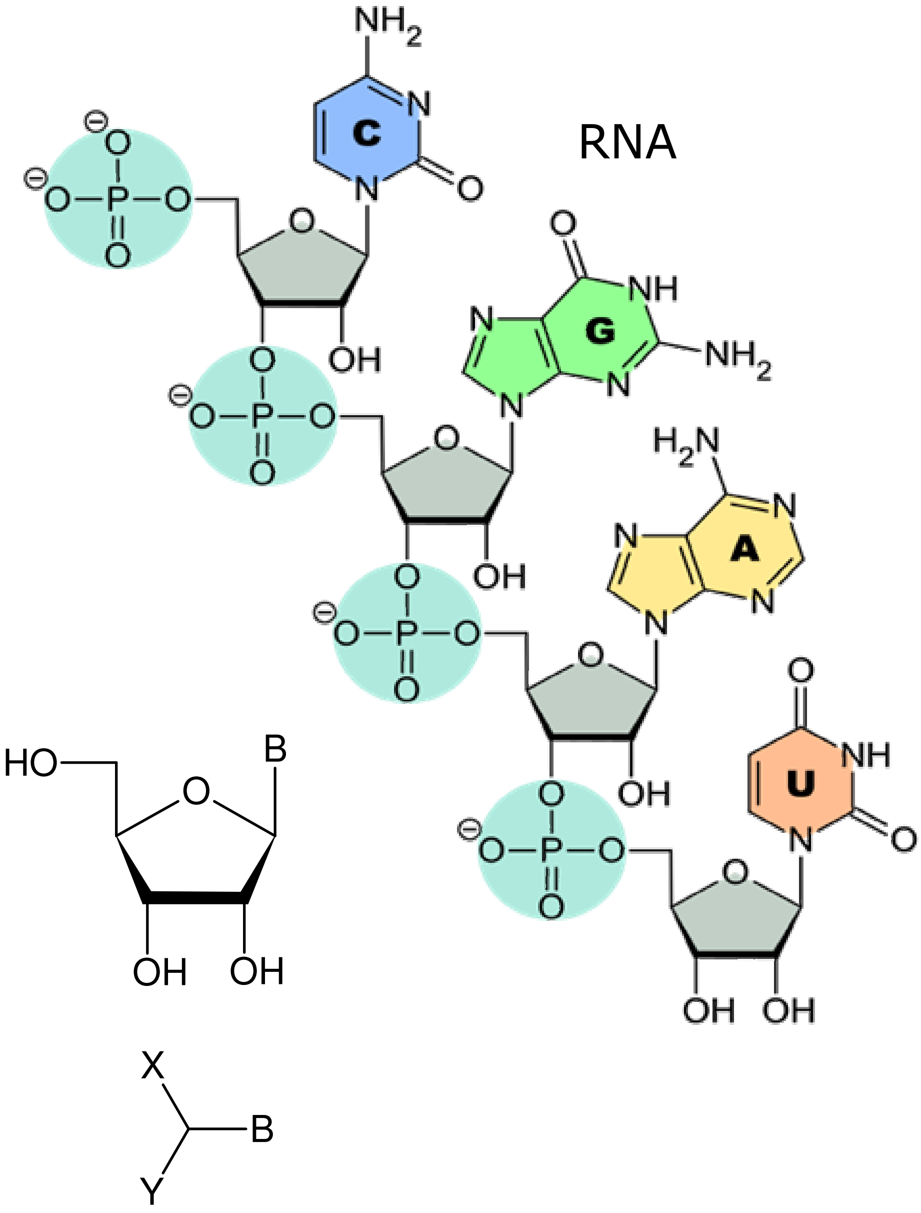
two sets with better coverage colored blue and red

Interactive graphics:

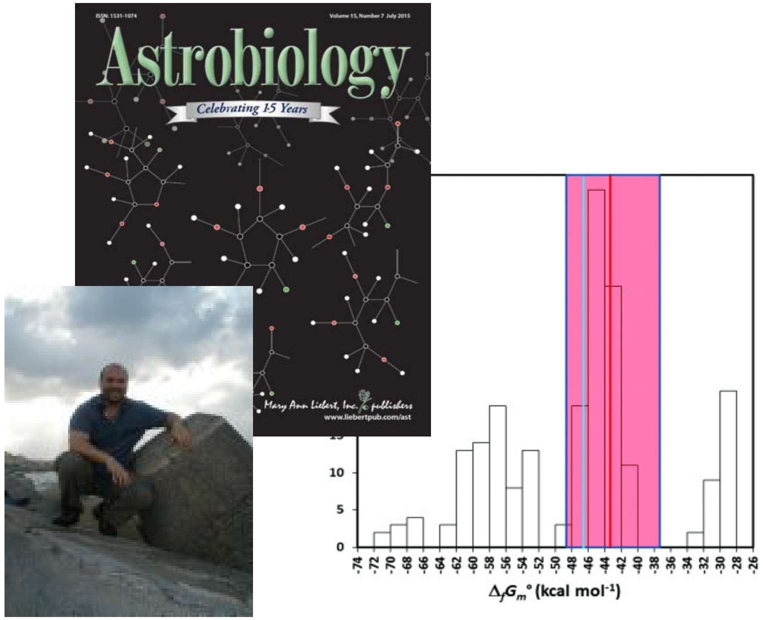
www.molgen.de/graphics/AdaptPropCodedAA/Fig2a/index.html
www.molgen.de/graphics/AdaptPropCodedAA/Fig2b/index.html

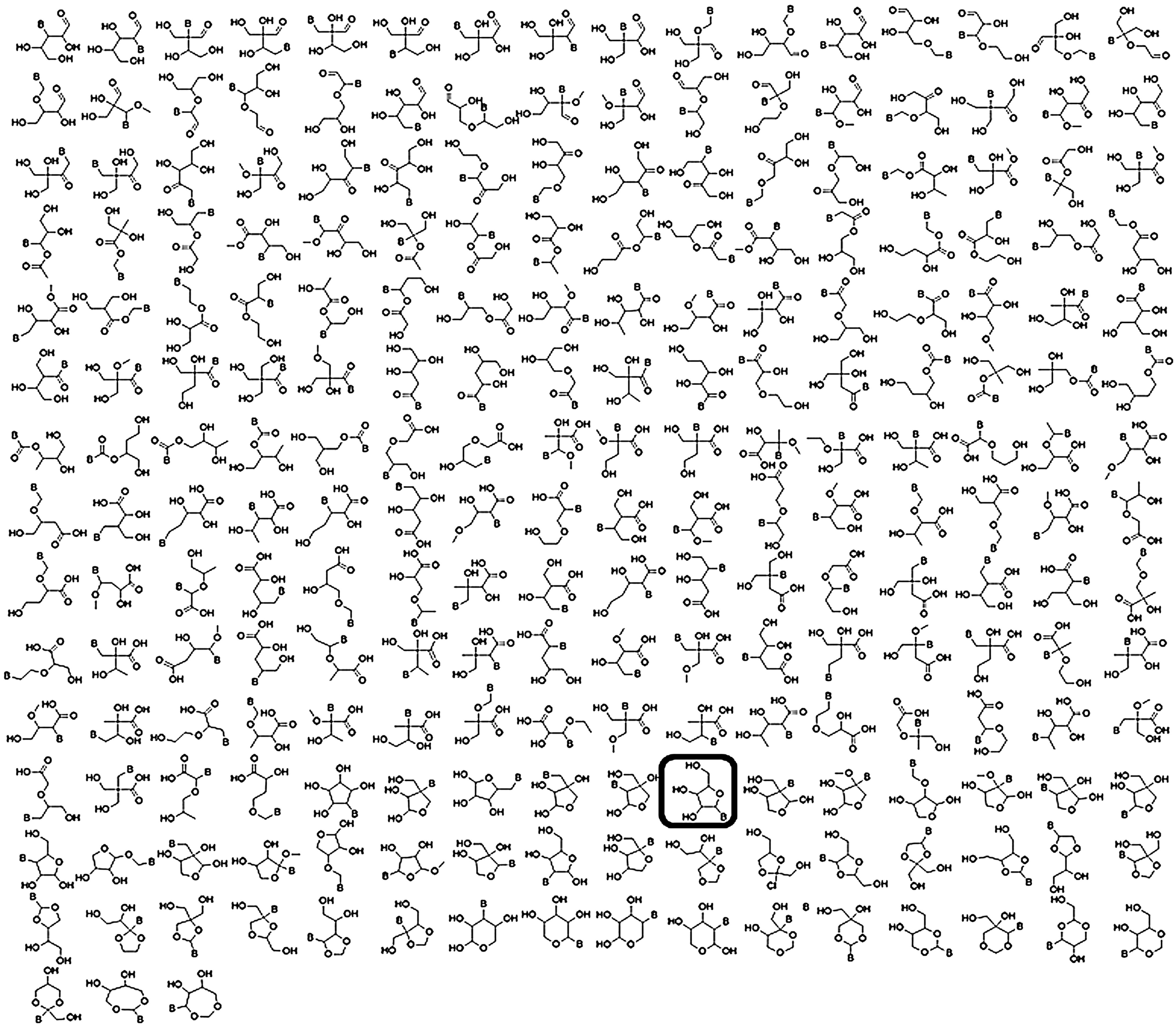
Nucleotides

- 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



First results: Isomers of ribose





Conclusion:

ribonucleosides may have competed with a multitude of alternative structures

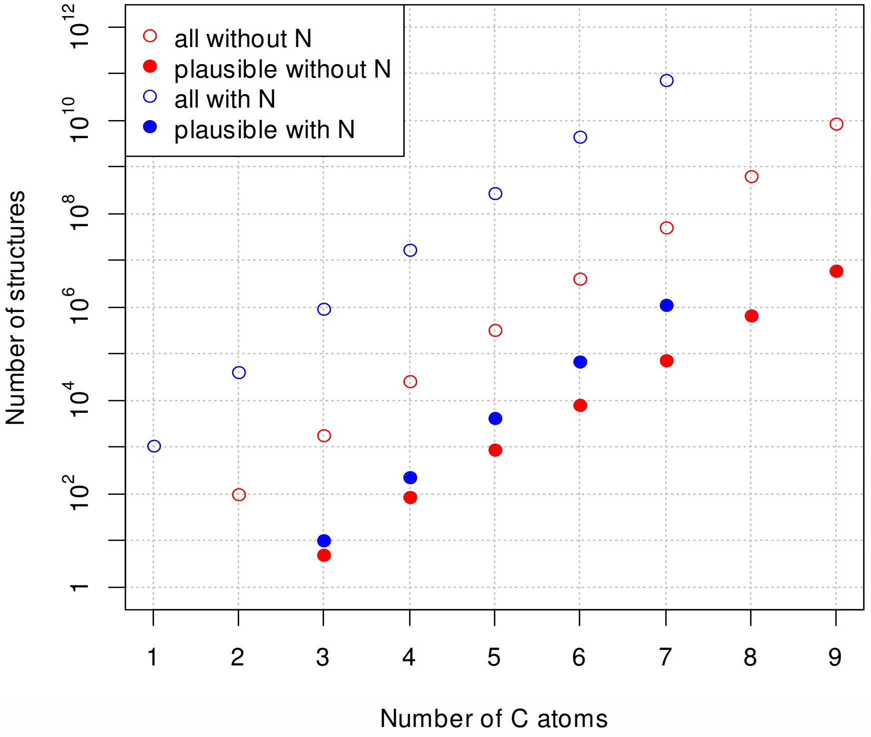
Cleaves HJ, Meringer M, Goodwin J. 227 Views of RNA: Is RNA Unique in Its Chemical Isomer Space? Astrobiology 15(7), 538 (2015)

Outlook: explore chemical space of general nucleosides

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

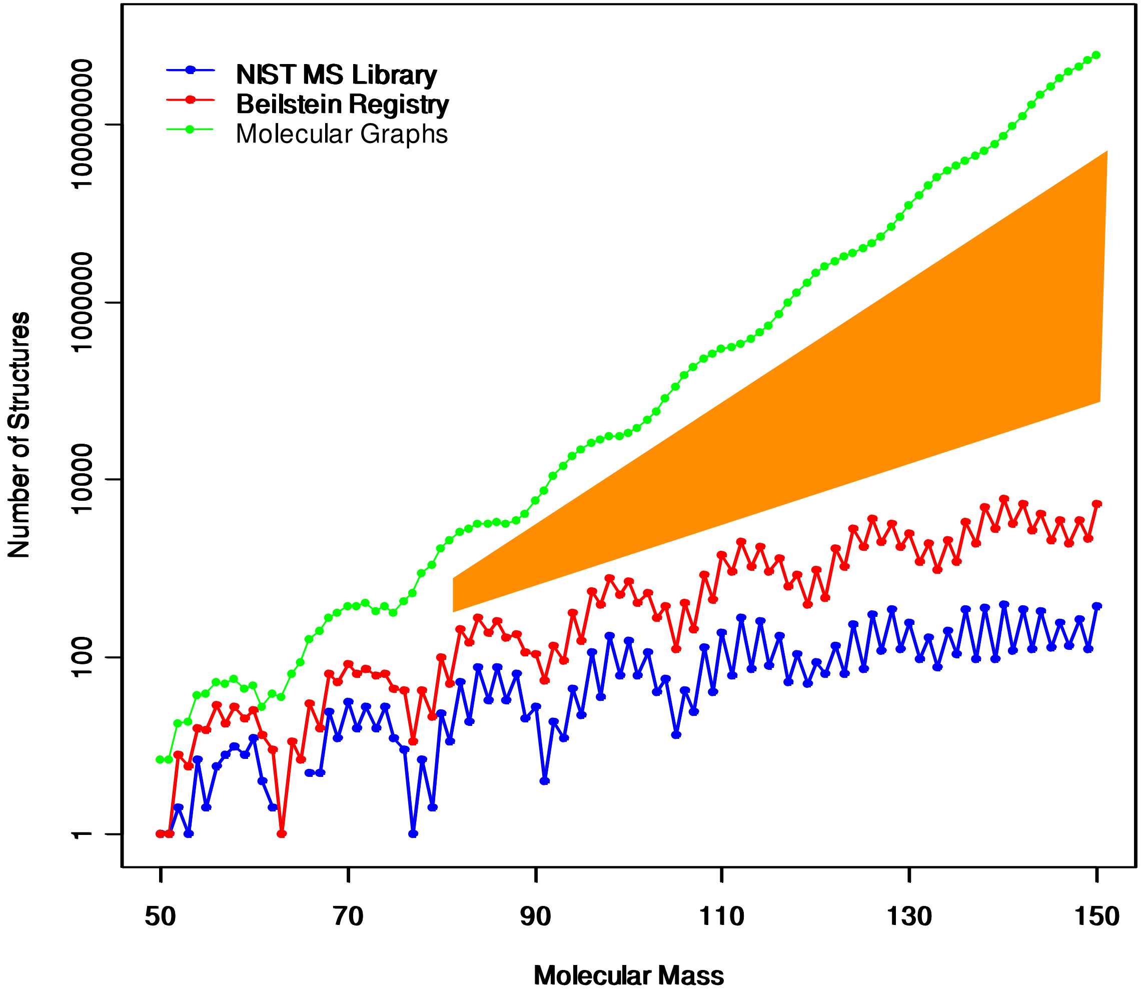
Sizes of libraries



Outlook: general small molecule compound spaces

Structures:

- elements C, H, N, O , S
- at least 1 C-atom
- standard valencies
- no charges
- no radicals
- no stereoisomers
- only connected structures
- chemically plausible structures

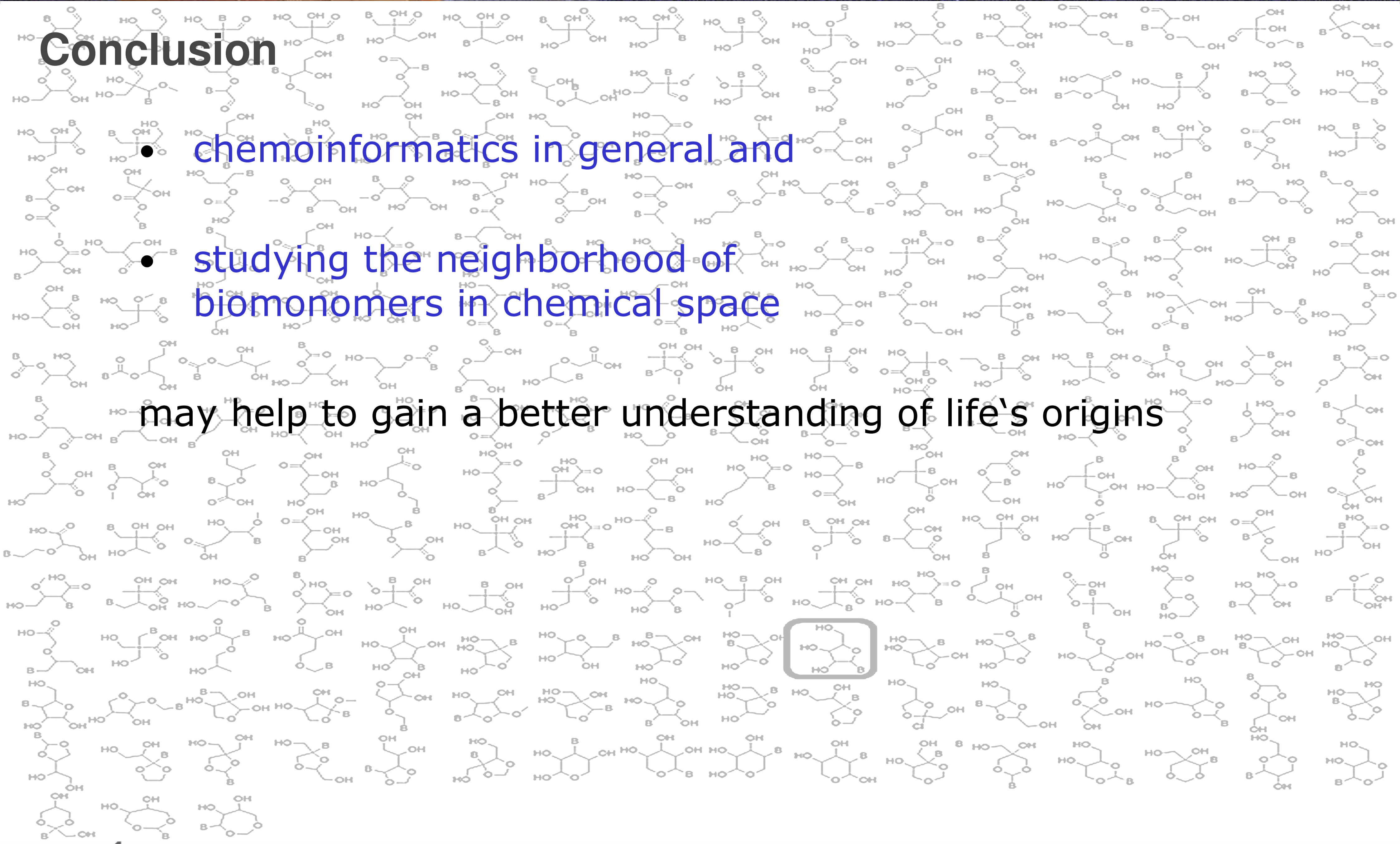


A. Kerber, R. Laue, M. Meringer, C. Rücker: Molecules in Silico: Potential versus Known Organic Compounds. MATCH 54 (2), 301-312, 2005.

Conclusion

- chemoinformatics in general and
- studying the neighborhood of biomonomers in chemical space

may help to gain a better understanding of life‘s origins

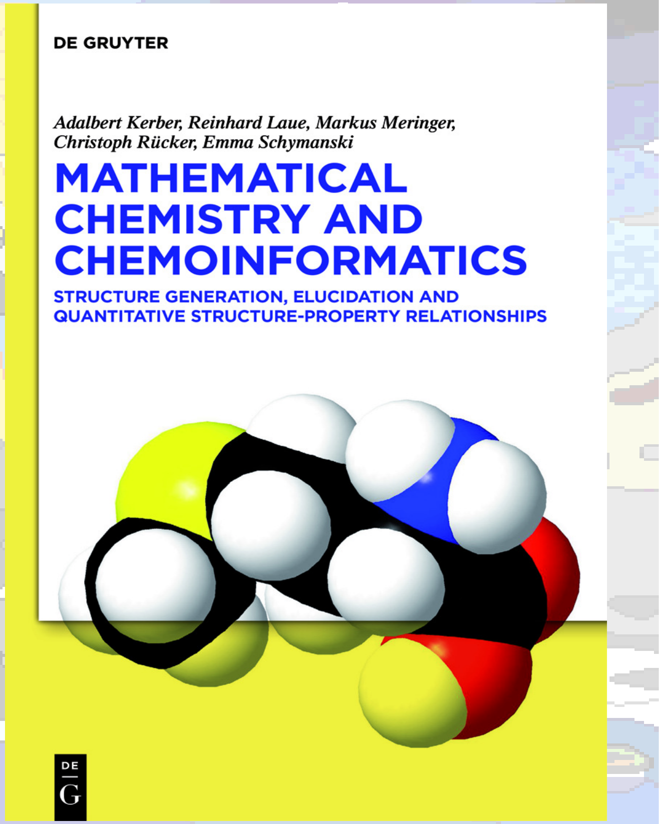


Acknowledgements

Stephen Freeland
Jim Cleaves
Melissa Ilardo
Bakhtiyor Rasulev
Jay Goodwin

NASA Astrobiology Institute
University of Hawaii

Earth Life Science Institute
Tokyo Institute of Technology



DE GRUYTER

Adalbert Kerber, Reinhard Laue, Markus Meringer, Christoph Rücker, Emma Schymanski

MATHEMATICAL CHEMISTRY AND CHEMOINFORMATICS

STRUCTURE GENERATION, ELUCIDATION AND QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIPS

DE G

MOLGEN Team
former Mathematics II
University of Bayreuth

THANKS FOR YOUR ATTENTION!