Molecular Structure Elucidation: More accurate or more diverse models?

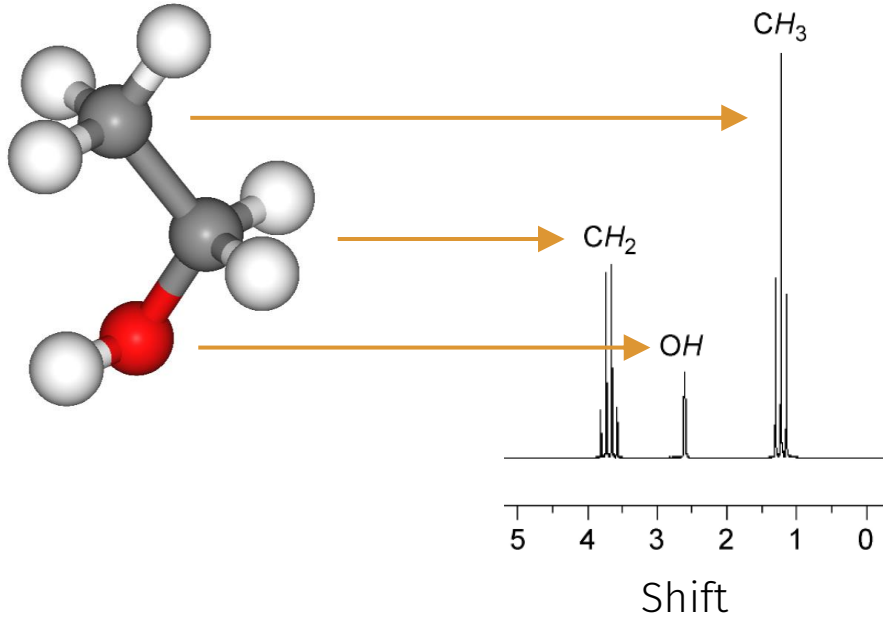
Guido Falk von Rudorff, University of Kassel


ferchault


nablachem.org/talks

Structure Elucidation

What we have: ^{1}H-NMR, ^{13}C-NMR, UV-VIS, stoichiometry, …



What we want: (Ranked) list of candidate **geometries**

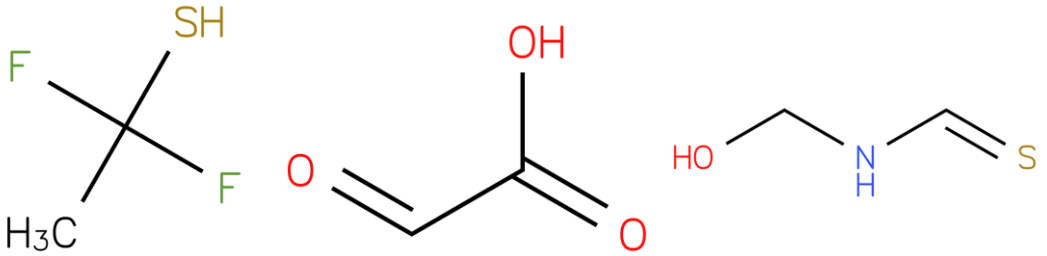
Scale

Take some atoms from {C, O, N, F, S}, H-saturated

349:

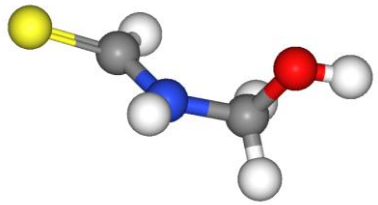
$C_2H_4F_2S$
$C_2H_2O_3$
C_2H_5NOS
...

9,917:



...

~52k:

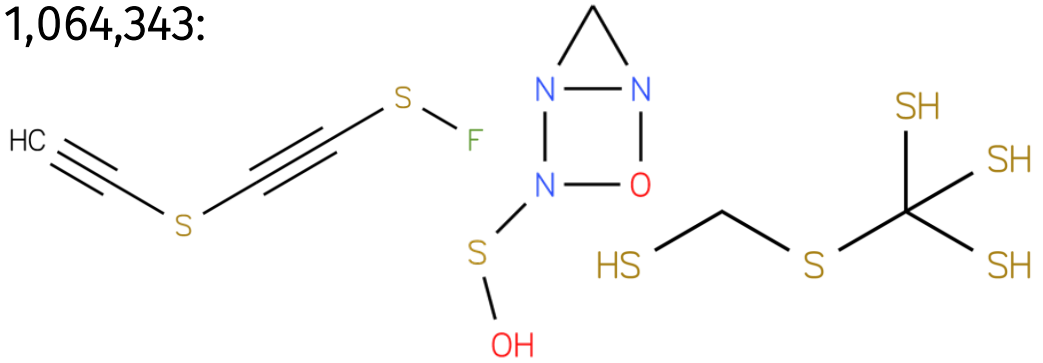


...

1,050:

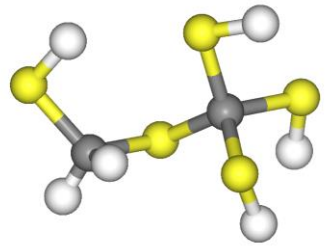
C_4HFS_2
$CH_3N_3O_2S$
$C_2H_6S_5$
...

1,064,343:



...

~23M:



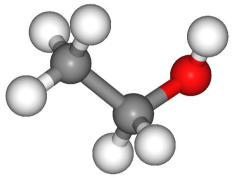
...

Ibuprofen: ~120B

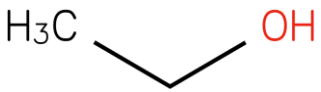
Identification Risks

Cost: Too many candidates for forward estimates[1-4]

Chemistry: Models often lack 3D information



vs.



hindrance and strain

Accuracy: Geometries nearly indistinguishable
Experimentally: ~0.2 ppm[5]

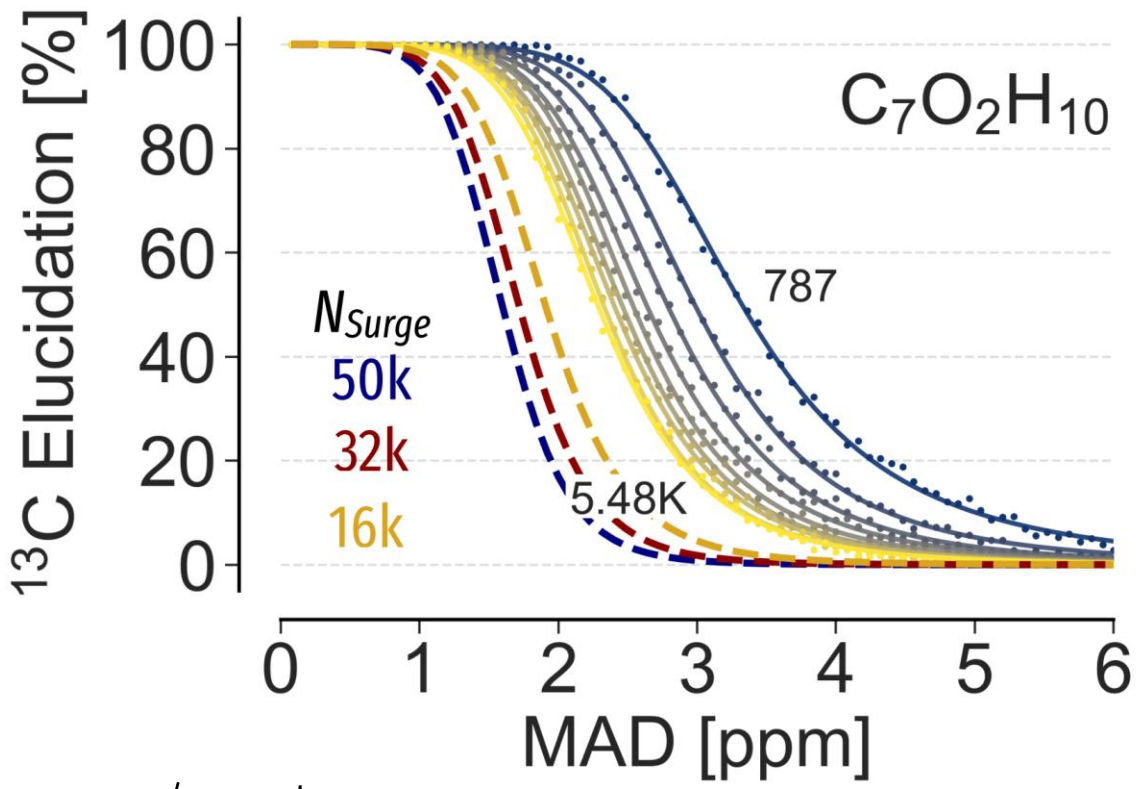
Estimation by adding noise to computed NMR shifts

[1] E Jonas, *NeurIPS*, 2019. [2] Z Huang et al, *Chem Sci*, 2021. [3] B Sridharan et al, *J Phys Chem Lett*, 2022.
[4] A. Howarth et al, *Chem Sci*, 2022. [5] F. Fardus-Reid et al, *Anal. Methods*, 2016.

How accurate do we need to be?

Database: ~5.5k spectra from QM9-NMR + noise

Possible graphs: ~50k

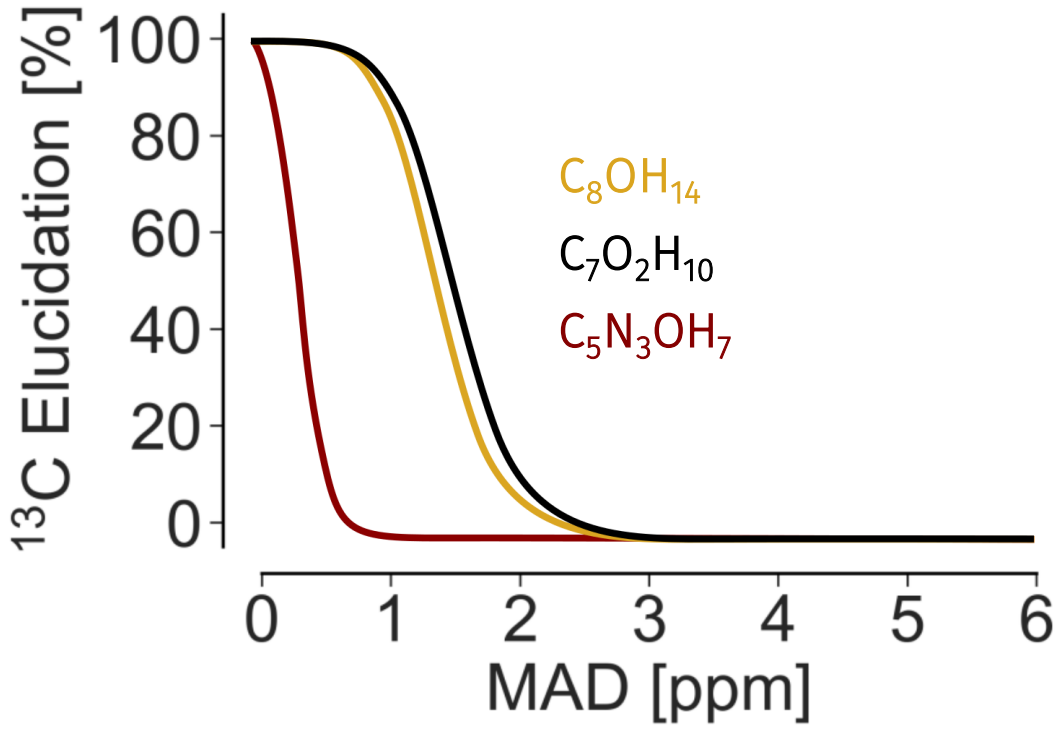


Optimistic: only one geometry/graph

D Lemm, **GFvR**, OA von Lilienfeld, *Digital Discovery*, 2023.

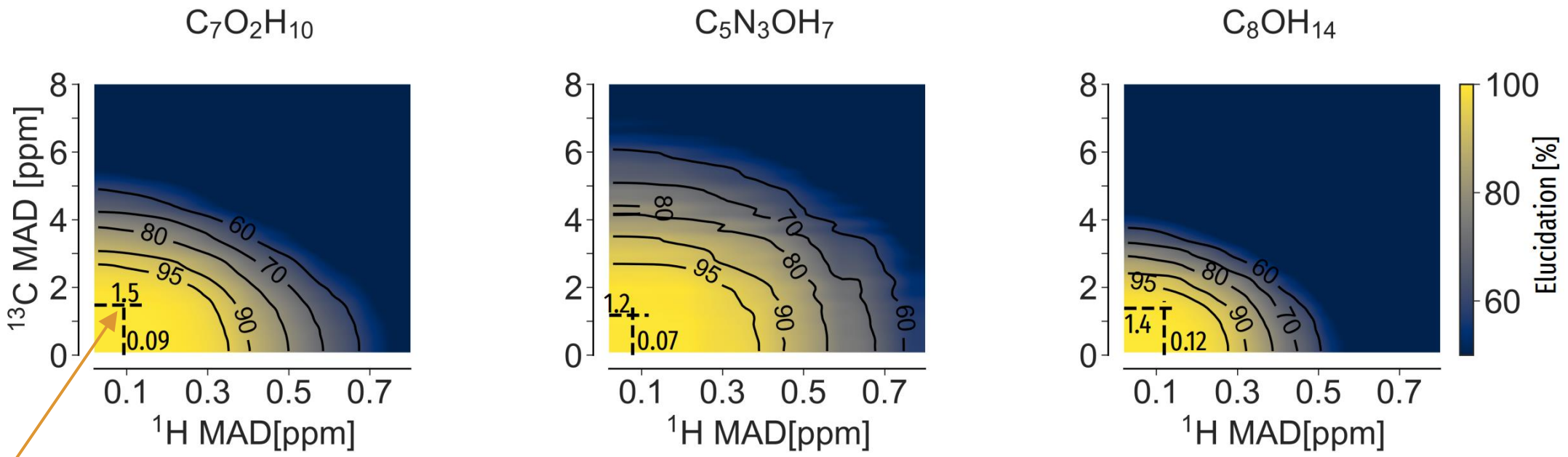
Impact of Stoichiometry

Depends on
- number of graphs
- self-similarity



D Lemm, **GFvR**, OA von Lilienfeld, *Digital Discovery*, 2023.

Combining ^{13}C and ^{1}H

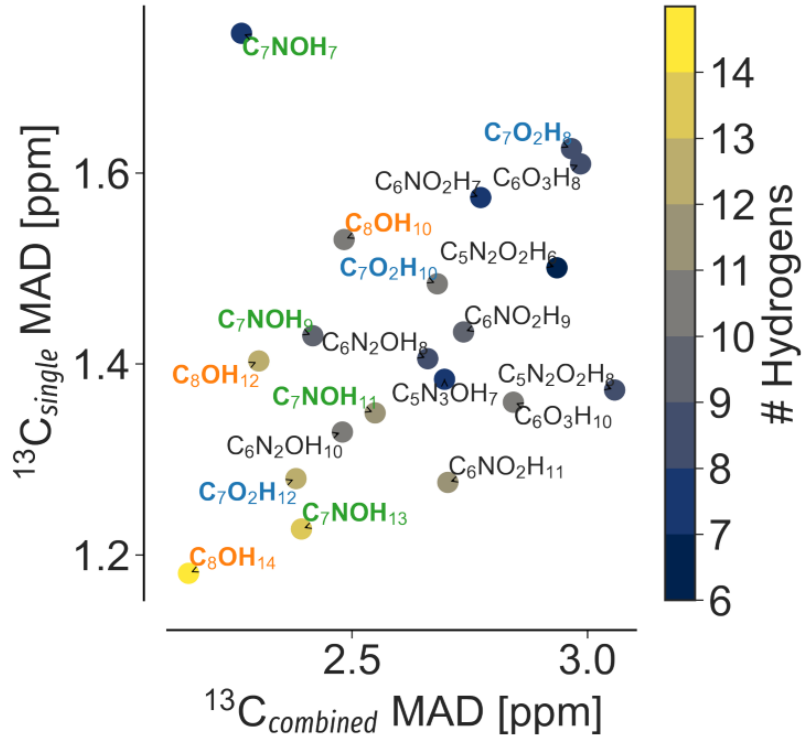


95% for only one spectrum

D Lemm, **GFvR**, OA von Lilienfeld, *Digital Discovery*, 2023.

Impact of Stoichiometry

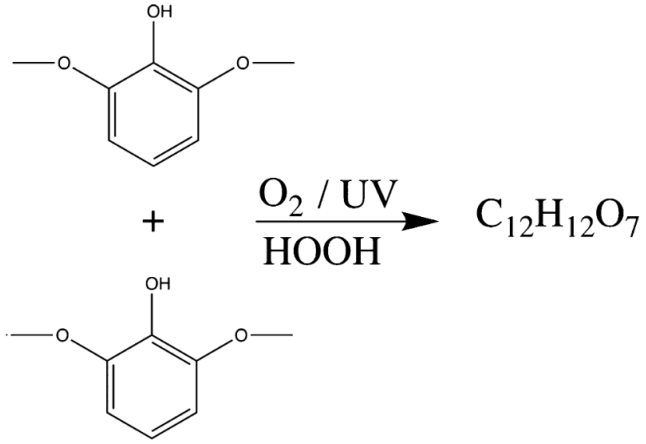
- Twice the permissible error: 1/100th of data points needed
- More hydrogens / fewer rings: harder problem despite more information

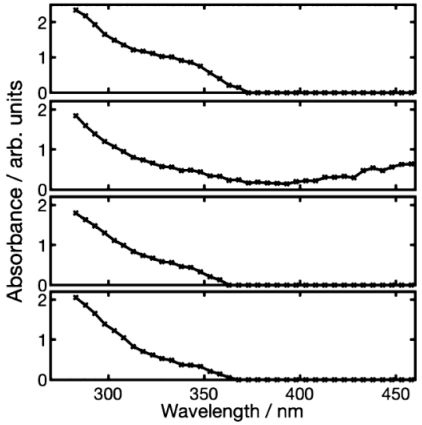


D Lemm, **GFvR**, OA von Lilienfeld, *Digital Discovery*, 2023.

UV-VIS case

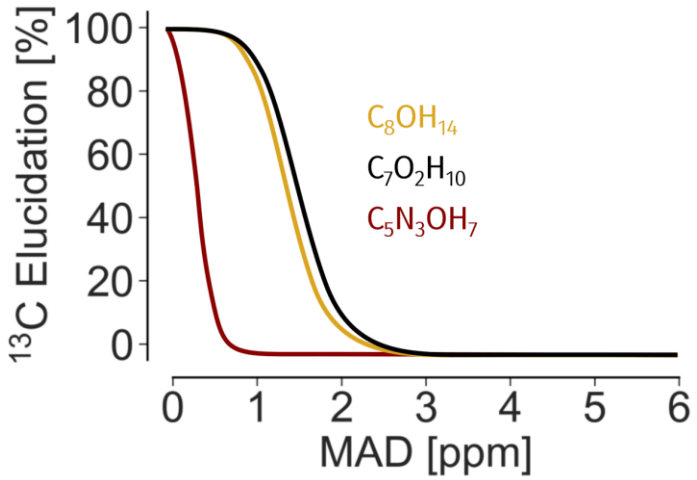
total molecules with two C_6 rings	263,917,411
and which have OH groups	263,917,411
and which have no oxygen chain longer than 2	161,160,394
and which have an oxygen connecting the carbon rings	115,715,458
and which have one aromatic ring	134,944
and which are stable	64,121
and which match spectrum 1	36,518

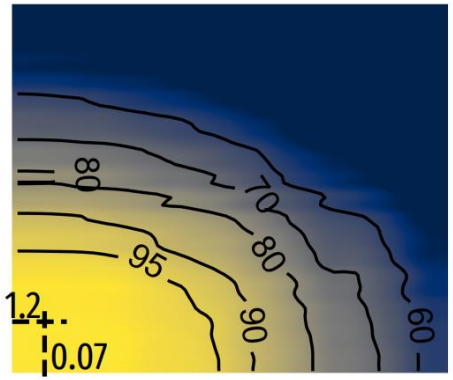




E. Tapavicza, GFvR, D. O. De Haan, M. Contin, C. George, M. Riva, O. A. von Lilienfeld, *Environ. Sci. Technol.* 2021.

Summary





Tradeoff | Multiple properties might be more valuable than accuracy

Data | Small-data regime potentially sufficient

NMR | Accuracy and gain quantified

NMR | ^{13}C alone insufficient even for small molecules

ferchault

nablachem.org/talks
vonrudorff@uni-kassel.de

Care about automatic differentiation?
We're hiring a postdoc