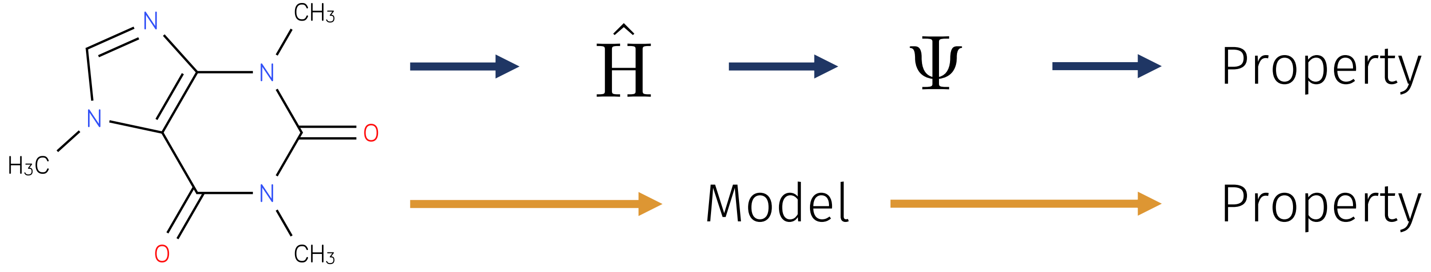
Random Sampling of Chemical Space: How to count without counting

Guido von Rudorff, University of Kassel

✉ vonrudorff@uni-kassel.de   🌐 nablachem.org/talks   ferchault   𝕏 @ferchault

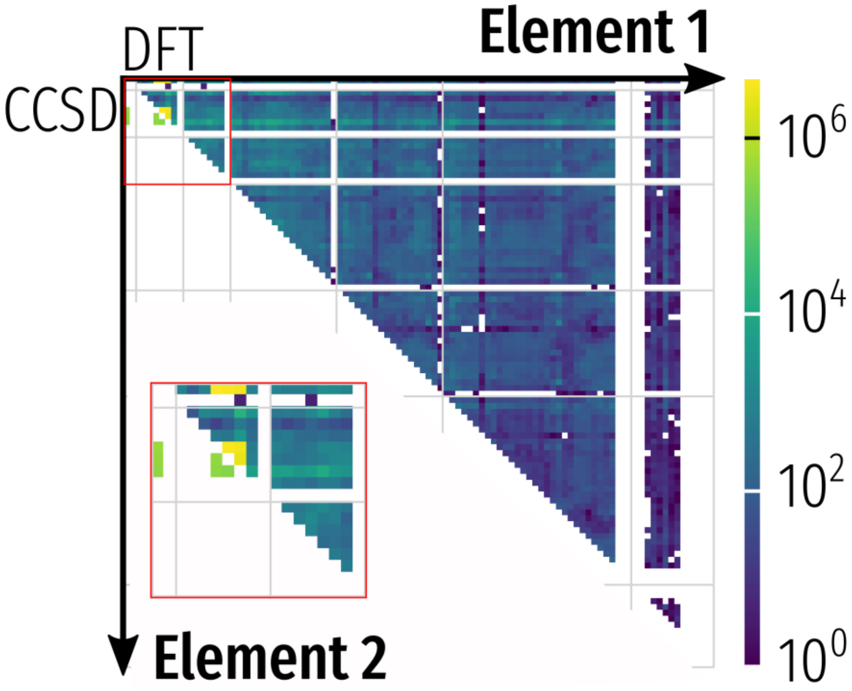
Learning solved?



$\hat{H}$ → Ψ → Property

Model → Property

Data Bias



Random Sampling

Allow for data-driven fundamental statements
“Most molecules do X”, “High X means low Y”

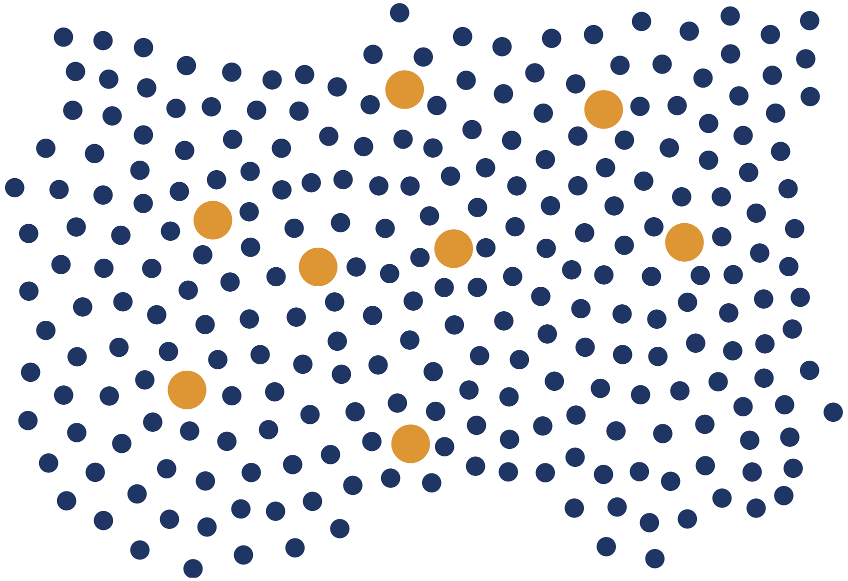
Transferability
More reliable understanding of trends

Lower data bias
More realistic generalisation error

More data efficiency
Maximally spanning coverage

Formal statements
Often require uniform sampling

Measure coverage
Generative methods

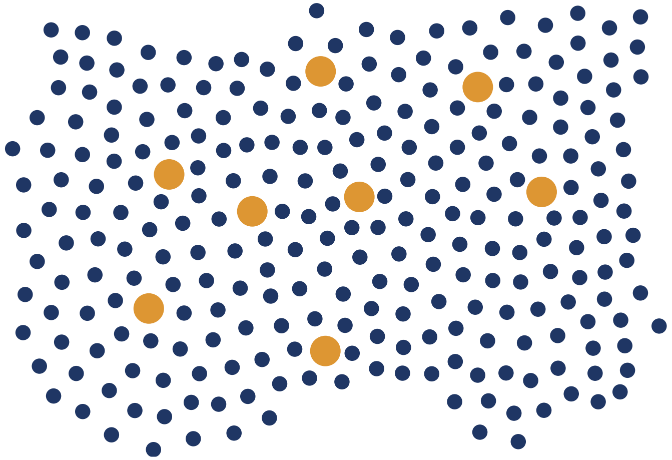


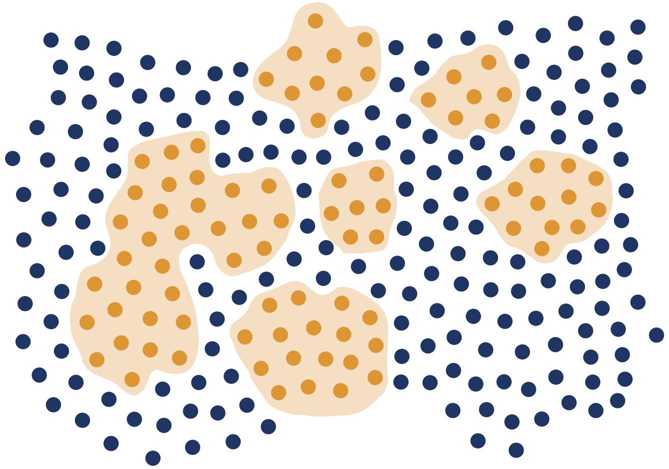
Problems

- Total number unknown
- Distribution unknown

Properties of domains?

Speed does not matter:
even enumeration is impossible.





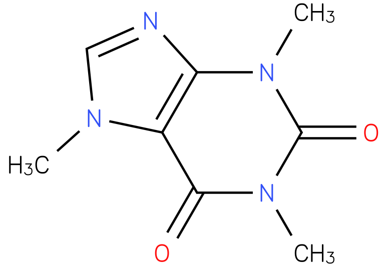
Random Sampling

Goal

Sample all molecules (with given constraints) with known probabilities.

Sampling

- Choose weighted random sum formula — $C_8H_{10}N_4O_2$
- Choose weighted random degree sequence — (8*4, 10*1, 4*3, 2*2)
- Choose weighted random molecular graph



Requirements

- Find all sum formulas and degree sequences — Solved, Seconds
- Sample loop-free multigraphs with given degree sequences uniformly — Solved, Seconds[1]
- Find weights

[1] C Greenhill, M Sfragara, *Theo Comput Sci*, 2018.

Starting with enumeration

Counting via enumeration

- SMOG (1996), MOLGEN (1998), ASSEMBLE(2000), OMG (2012), PMG (2013), MAYGEN (2021), surge (2022)
- Until about 10-15 atoms

Orderly generation

- Find canonical sorting of (partial) molecular graphs
- Create graphs in canonical order

Counting without enumeration

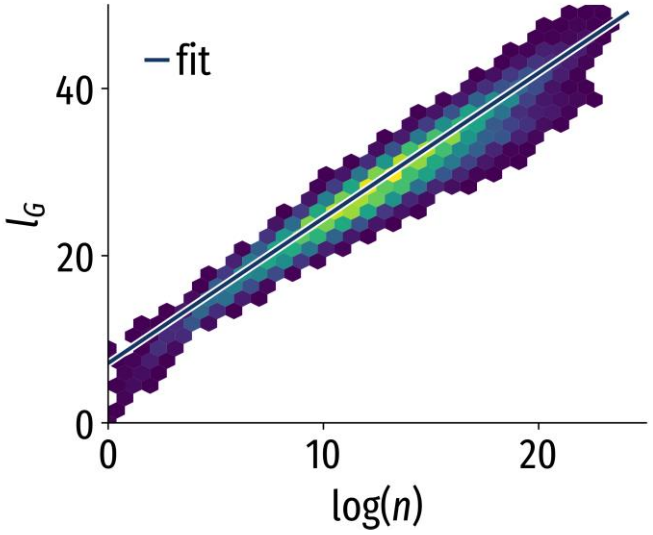
Goal

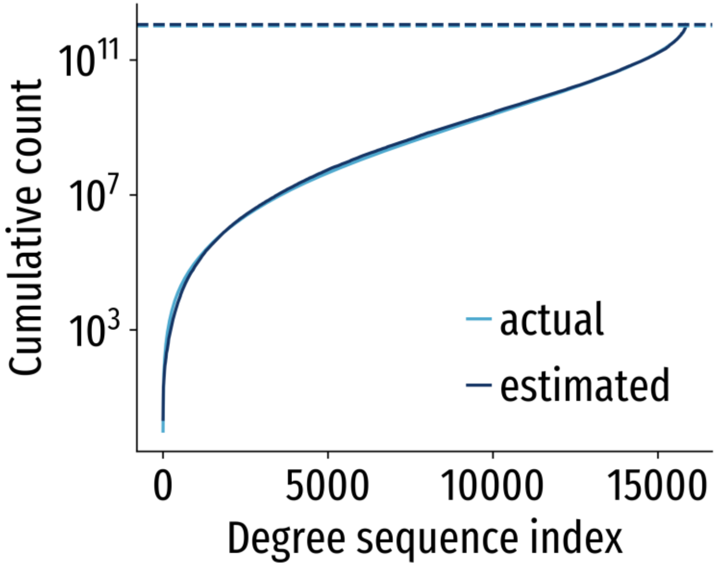
Estimate number n of loop-free multigraphs with given degree sequence.

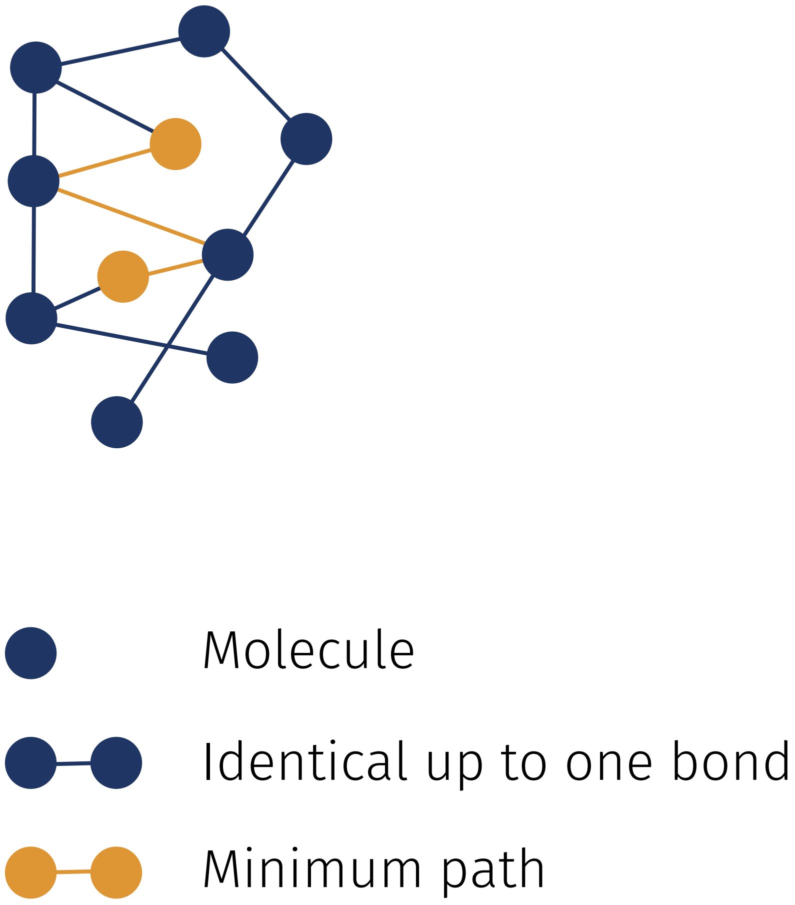
Average Path Length l_G

Sample from random molecule pairs

$$l_G \sim \log n$$







● Molecule

●—● Identical up to one bond

●—● Minimum path

Estimating Average Path Length

Pure degree sequence

Every valency exists only once.

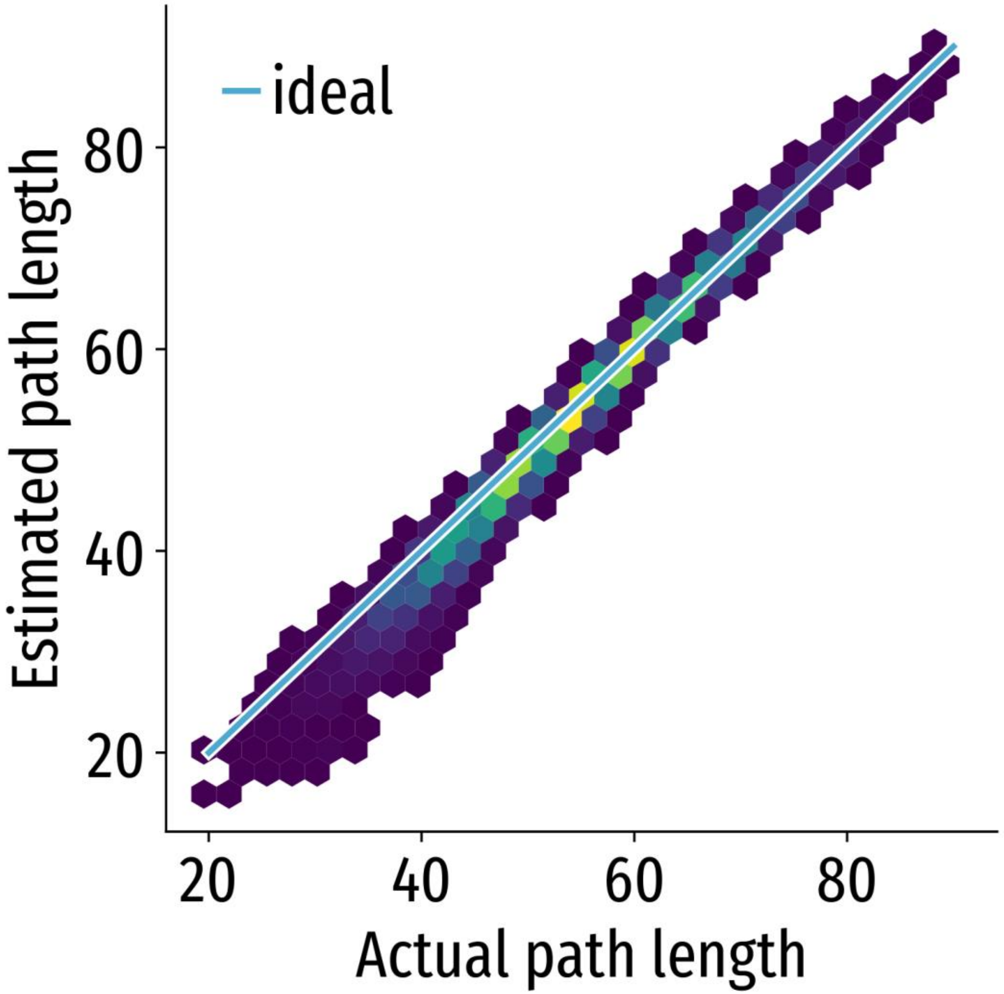
O=C=S O=C=O S=C=S

Non-pure estimate

- Assumes that random graphs are almost never symmetric
- All modifications independent
- Combinatorial product

$$N_P(d) = \prod_v \prod_i \binom{\sum_{j>i} c_j}{c_i}$$

$$l_G(d) = \left(1 + \left[\sum_i d_i\right]^{-1} \log N_P^L\right) l_G(d_U)$$



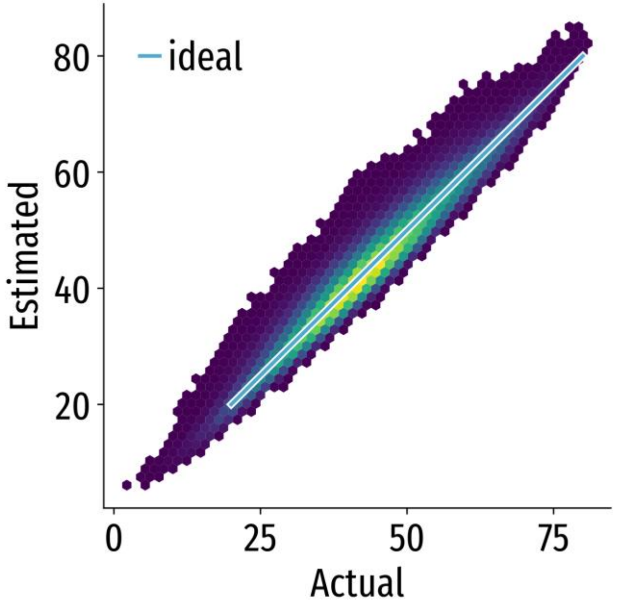
Asymptotic Scaling

Average Path Lengths become expensive

- Heuristics become less efficient
- More sampling
- Converging slowly

Asymptotic Scaling [1]

- Needs to be calibrated to molecules



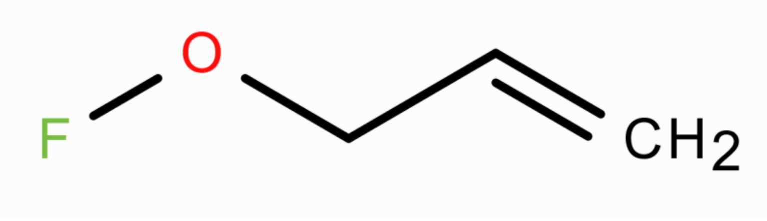
$$G = \frac{M!}{(M/2)!2^{M/2}k_1!\cdots k_n!} \exp\Bigg(\left(y_1 - \frac{1}{2}\right)\frac{M_2}{M} + \left(x_2 - \frac{1}{2}\right)\frac{M_2^2}{2M^2} + \frac{M_2^4}{4M^5} - \frac{M_2^2 M_3}{2M^4} + \left(x_3 - x_2 + \frac{1}{3}\right)\frac{M_3^2}{2M^3} + (an+b)/M + (cn+d)M + e\Bigg)$$

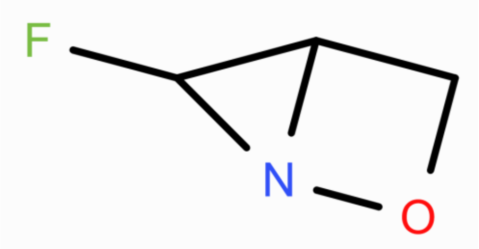
$$M_r = \sum_i^n [k_i]_r$$

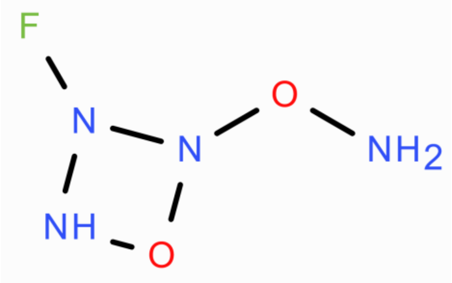
[1] C Greenhill, B McKay, *SIAM*, 2013.

Examples

10 atoms, CHONF, at least 3 hydrogens and one fluorine



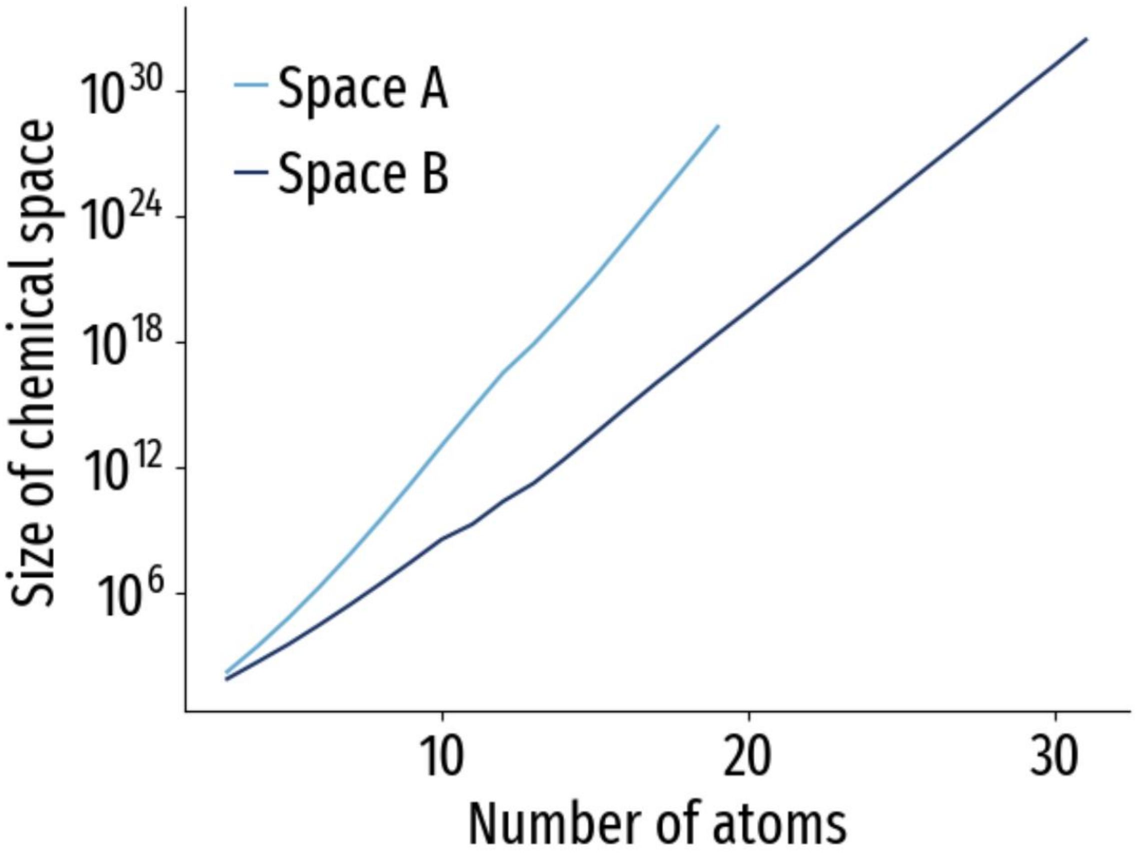




Available conditions in this approach

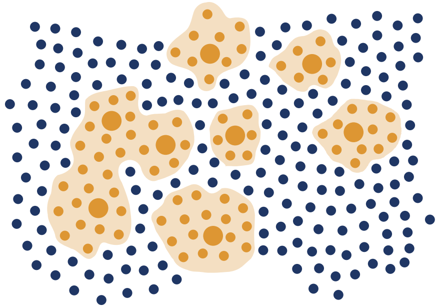
- Number of atoms
- Atoms per element (and combinations)
- Valences
- Via rejection sampling:
  - Bonds (count and bond orders)
  - Ring presence / membership
  - Stability
  - Substructures
  - …

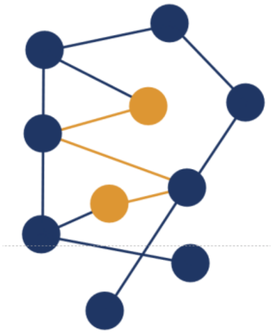
Scaling

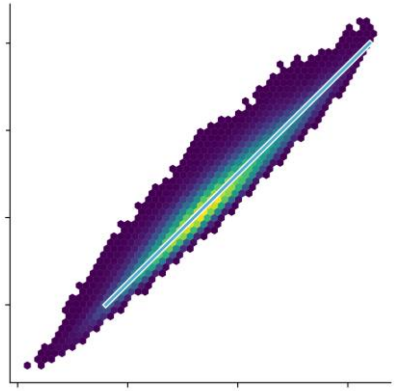


Space	Valence	Multiplicity	Example
A	1	5	F, H, Cl, Br, I
	2	2	O, S
	3	2	N, P
	4	3	C, S, Si
	5	2	N, P
	6	1	S
B	1	5	F, H, Cl, Br, I
	2	2	O, S
	3	2	N, P
	4	1	C
	5	1	P

Summary







Thanks

Ali Banjafar
Sarah Engel
Sana Qureshi
Nicolas Grimblat

Randomized	Known distribution: statistical statements
Regions	No one-by-one iteration
Bias reduction	Datasets and predictive power
Seeding	Maximally spanning datasets or Monte-Carlo acceleration

✉ vonrudorff@uni-kassel.de

nablachem.org/talks

ferchault

𝕏 @ferchault