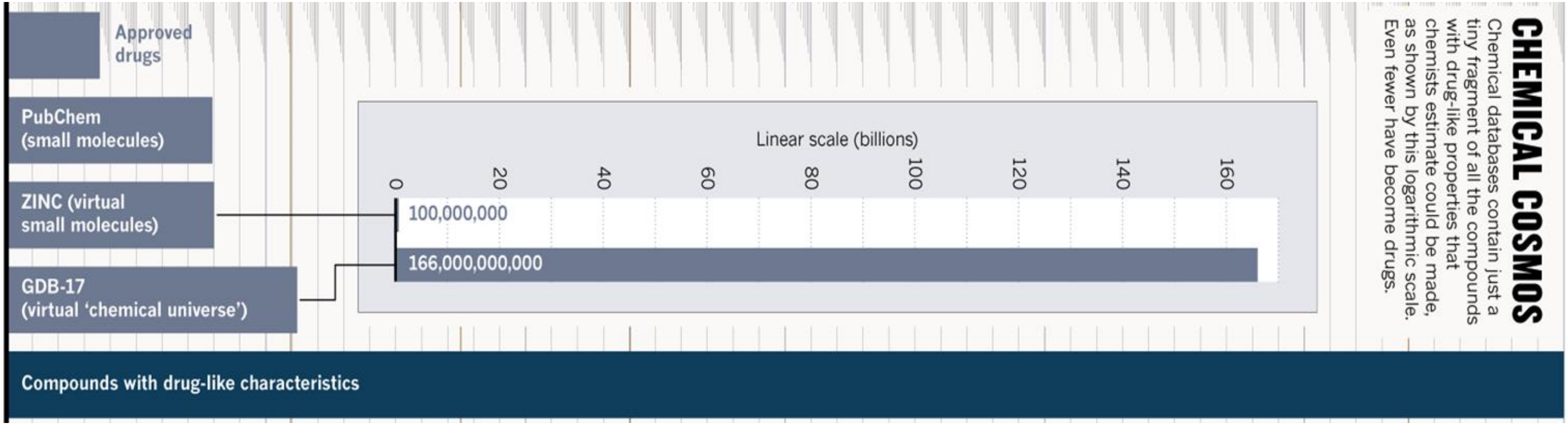
Chemistry at scale



Commercial databases

- 164 million molecules
- 15k added daily

Scale

- One person: 1 million compounds/second
- 10 billion people on earth
- 10^{26} universe ages to go through

https://doi.org/10.1038/549445a

Necessity vs Convenience

Necessity

- Only way to cover problem size
- Still open to systematic evaluation
- Often used as prefiltering step

- Complicated chemistry
- Tricky / error-prone reference calculations

Convenience

- Can be done more accurately
- Uneconomical/cumbersome reference method
- Often used as direct but optional substitute

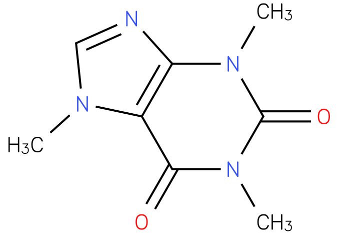
- Standard energy calculations of well-behaved systems
- Semi-empirical level sufficient

Summary

1. Almost none of chemical space has been explored.
2. Scaling is a key aspect to think about when comparing methods.
3. Chemical diversity drives molecular diversity.

Potential Energy Surfaces

Concept



$$\hat{\mathrm{H}}(\mathbf{R}_I, Z_I, n, \sigma) \rightarrow \Psi \rightarrow \text{Energy} \quad E(\mathbf{R}_I | Z_I, n, \sigma)$$

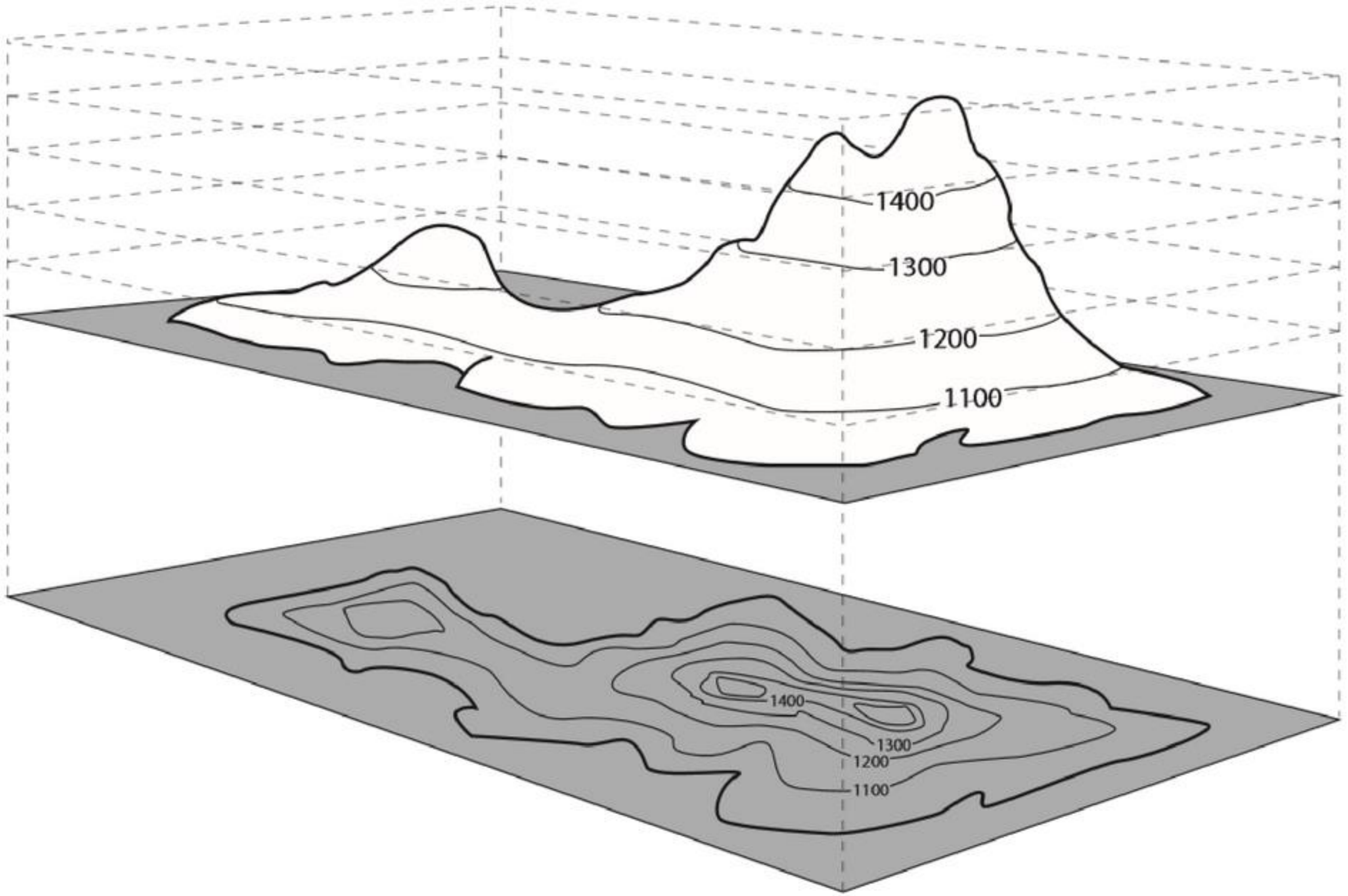
Numbers of electrons

Electronic spin

Nuclear charges

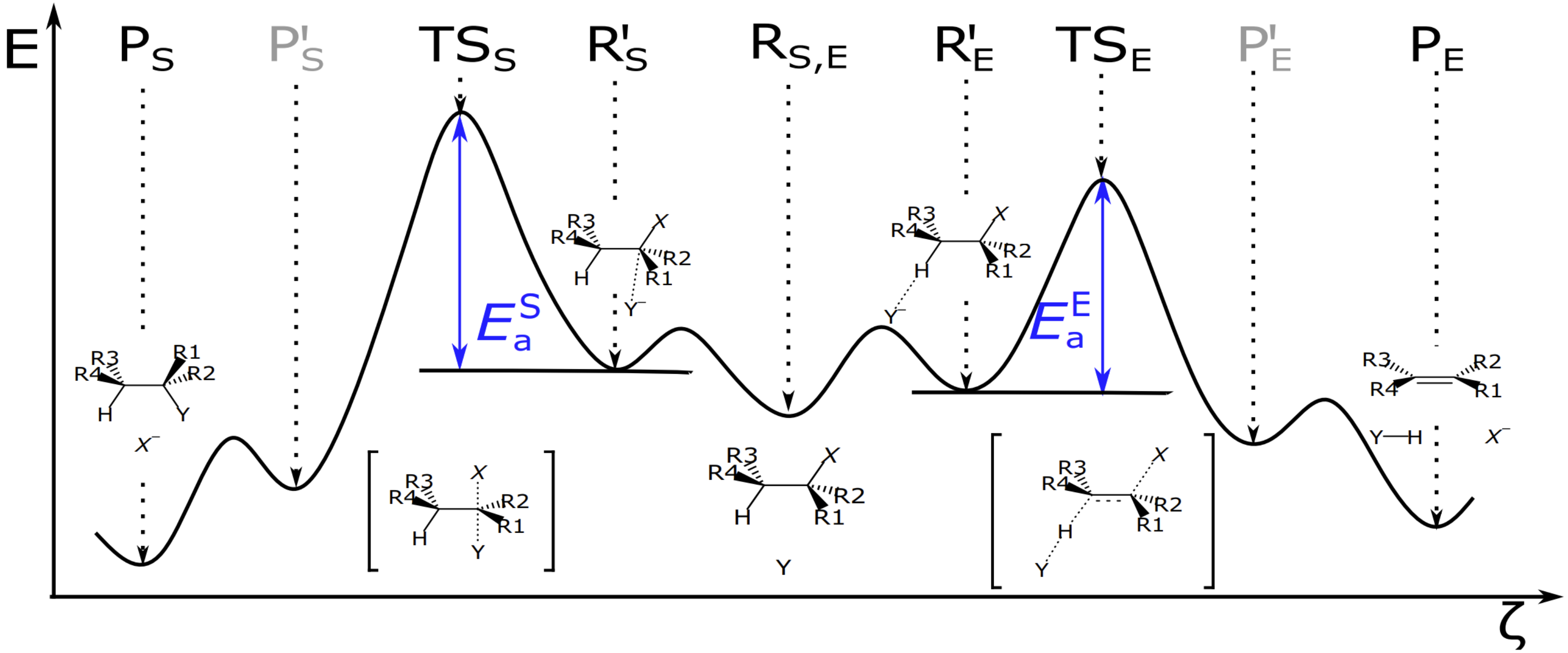
Nuclear coordinates

Topographical map



https://mapswithnoroads.com

Critical points



- Local minima — Meta-stable, ensembles
- Global minima — Most stable
- Saddle points — Barriers / access

- Attractive basins

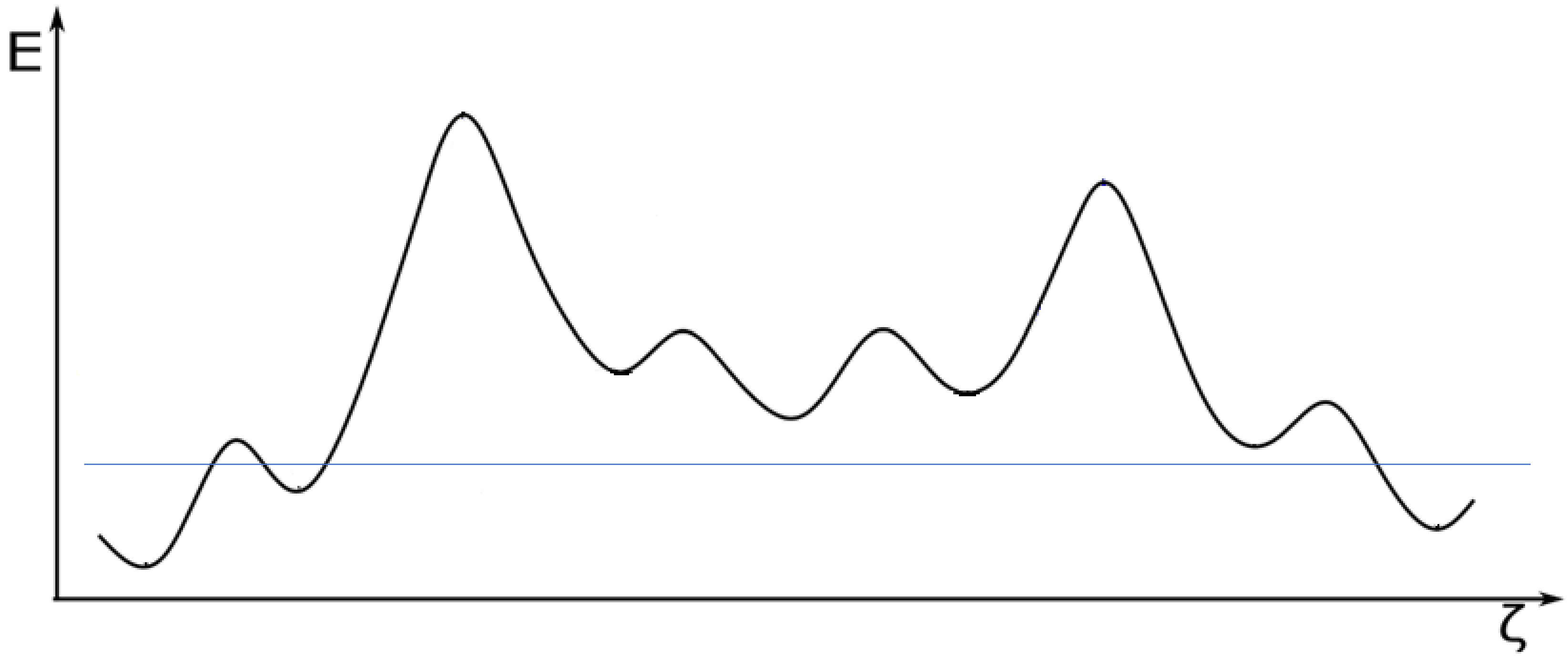
https://doi.org/10.1088/2632-2153/aba822

Properties

Curvatures:
- Vibrational frequencies
- Normal modes
- Vibrational energy levels

Barriers:
- Tunneling
- Thermalisation

Thermally accessible regions:
- Ensemble of configurations



Geometry Optimization

Frequent problem class

Energy

- Find most stable molecular geometry	Compare conformers
- Find transition state geometries	Identify reaction pathways

Residuals

- Fitting experimental data	Model observations
- Potential fitting	Simplify calculations
- Machine learning	Surrogate models

Definitions

Solution coefficients x
- Molecular geometries
- Fitting coefficients
- Model coefficients

Scalar objective function f
- Energy
- Residual norm
- Here: smooth, i.e. differentiable function

$$f(x_1, x_2, \ldots, x_n) = f(\mathbf{x}) = y$$

Domain X
- Valid parameter range
- Any solution within accepted

$$\mathbf{x}_0 \equiv \underset{\mathbf{x}\in X}{\operatorname{argmin}} f(\mathbf{x})$$
$$= \{\mathbf{x} | \mathbf{x}, \mathbf{y} \in X : f(\mathbf{x}) \leq f(\mathbf{y})\}$$

Target x_0
- Maximise or minimise y (over domain)