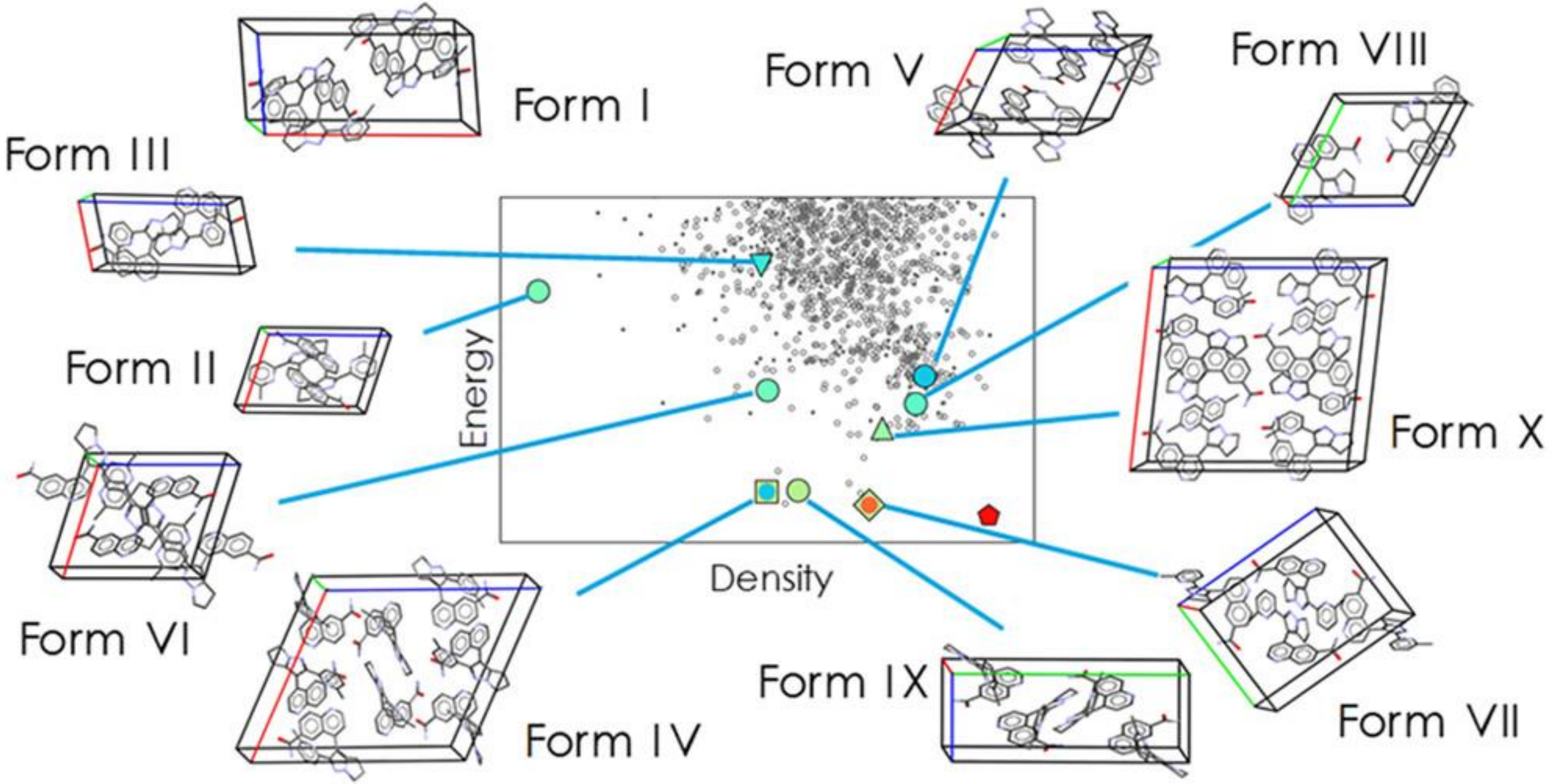
Introduction to Crystal Structure Prediction

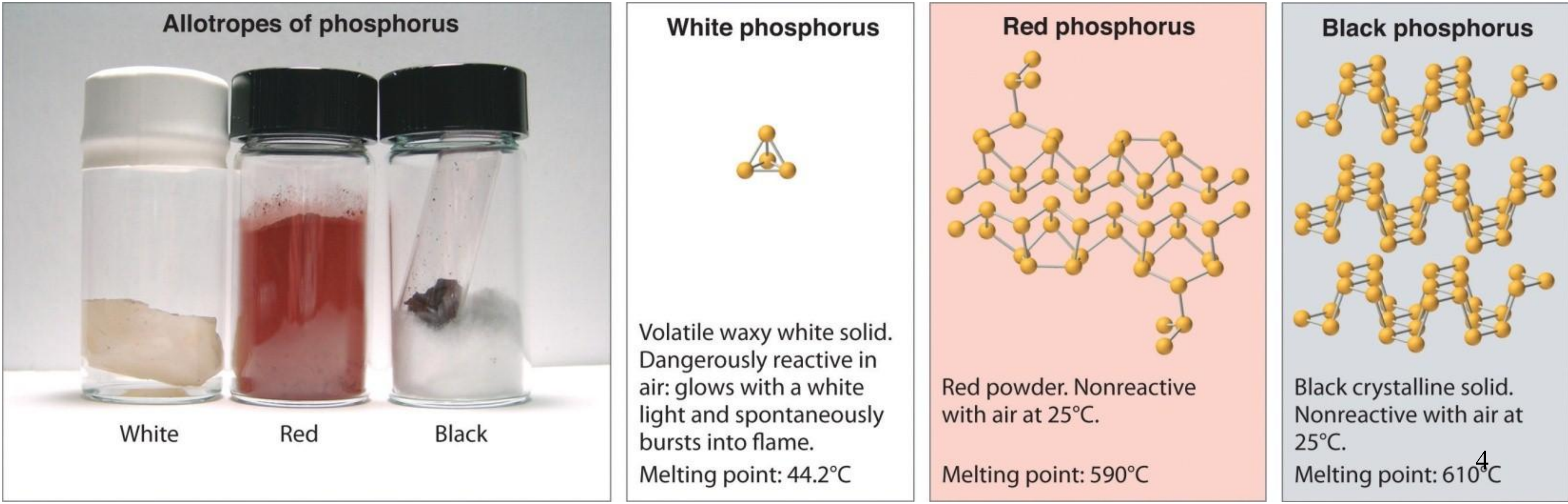


- Background
- Review of computational chemistry
  - Molecular mechanics
  - Electronic structure theory
  - Types of calculations and systems
- Crystal structure prediction (CSP)
  - Structure generation
  - Structure ranking
- Interpretation of results
  - Methods of crystal structure comparison
  - Thermal free energy corrections
- CCDC blind tests

- **Background**
- Review of computational chemistry
  - Molecular mechanics
  - Electronic structure theory
  - Types of calculations and systems
- Crystal structure prediction (CSP)
  - Structure generation
  - Structure ranking
- Interpretation of results
  - Methods of crystal structure comparison
  - Thermal free energy corrections
- CCDC blind tests

Polymorphism

- Distinct crystal structures for a compound of fixed composition



Allotropes of phosphorus

White phosphorus

Volatile waxy white solid. Dangerously reactive in air: glows with a white light and spontaneously bursts into flame.

Melting point: 44.2°C

Red phosphorus

Red powder. Nonreactive with air at 25°C.

Melting point: 590°C

Black phosphorus

Black crystalline solid. Nonreactive with air at 25°C.

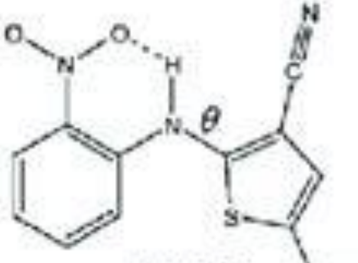
Melting point: 610°C

Polymorphs of ROY

(1) **R** P-1
mp 106.2 °C
θ = 21.7°

(2) **Y** $P2_1/c$
mp 109.8 °C
θ = 104.7°

(3) **ON** $P2_1/c$
mp 114.8°C
θ = 52.6°



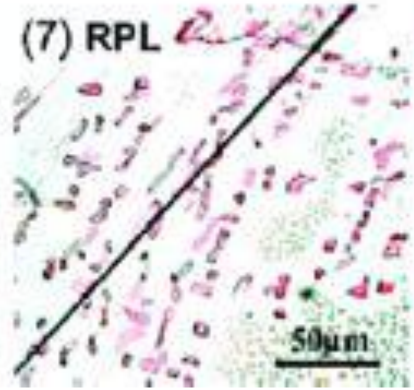
ROY

(4) **OP** $P2_1/c$
mp 112.7 °C
θ = 46.1°

(5) **YN** P-1, mp 99 °C
θ = 104.1°

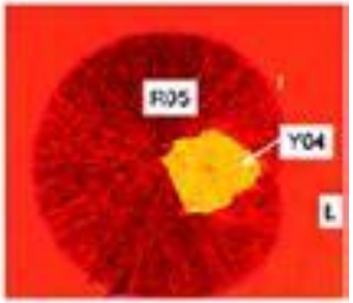
(6) **ORP** Pbca
mp 97 °C, θ = 39.4°

(7) **RPL**



(8) **Y04**

(9) **YT04** $P2_1/c$
mp 106.9°C
θ = 112.8°



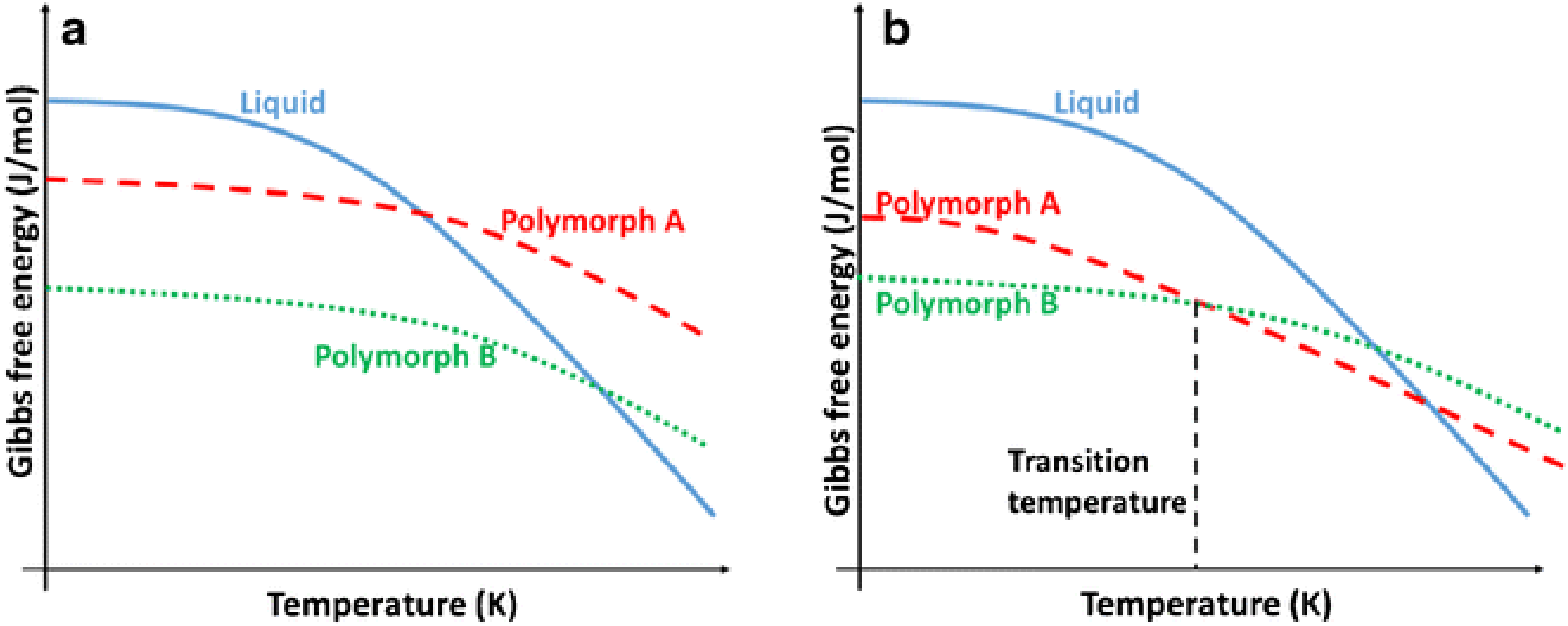
(10) **R05**

Importance in Pharma – Ritonavir

- 1992 – softgel formulation
- 1996-1998 HIV related deaths fell from 50,000/yr --> 18,000/yr in USA
- 1998 Product QC failures (solubility)
- Spontaneous nucleation of a new polymorph
- Reformulation took 1 yr and est. $250 million
- Regulatory filings now require demonstration of knowledge and control over polymorphism

Polymorphic relationships

- Monotropic or enantiotropic



Polymorphism frequency

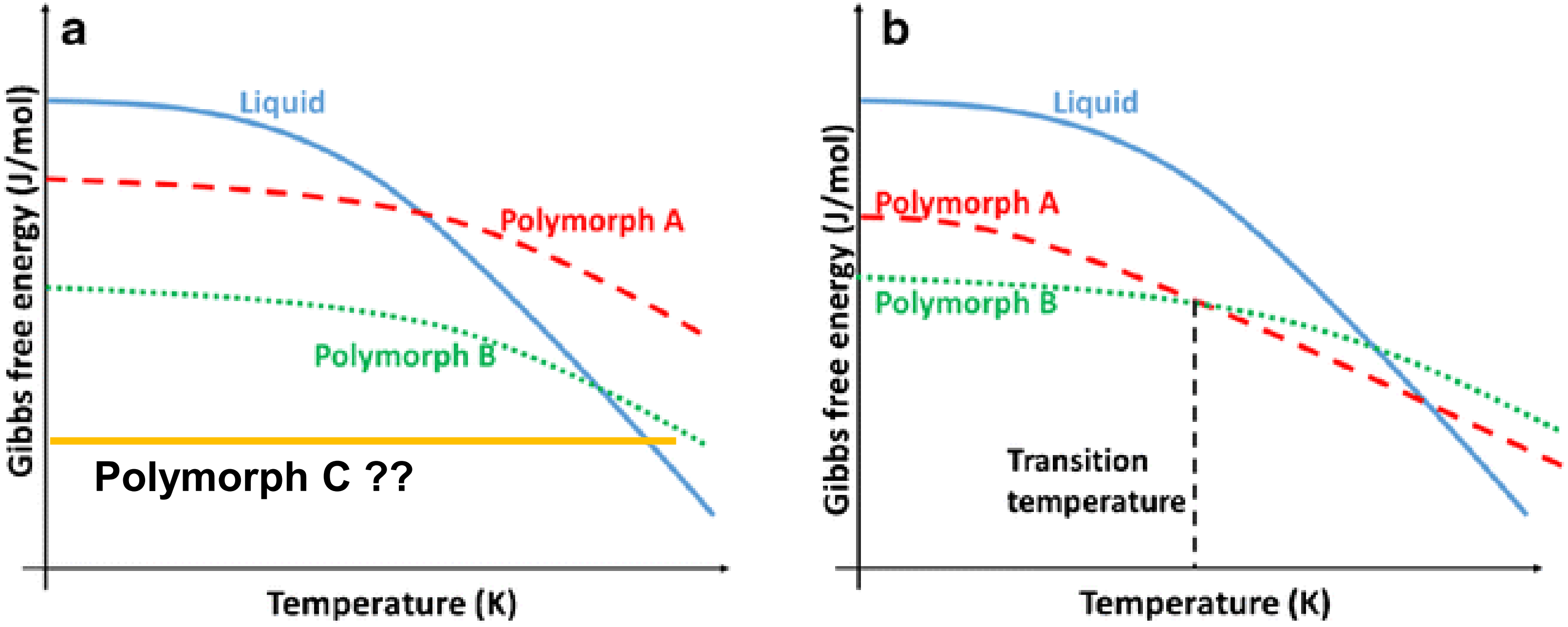
- “**Every compound has different polymorphic forms**, and that, in general, the **number of forms known** for a given compound is **proportional to the time and money spent in research** on that compound.”

- When do you stop looking?

W C McCrone, Polymorphism, Physics and Chemistry of the Organic Solid State, Vol 2, pp. 725–767, New York, Wiley Interscience, 1965

Polymorphic relationships

- Monotropic or enantiotropic

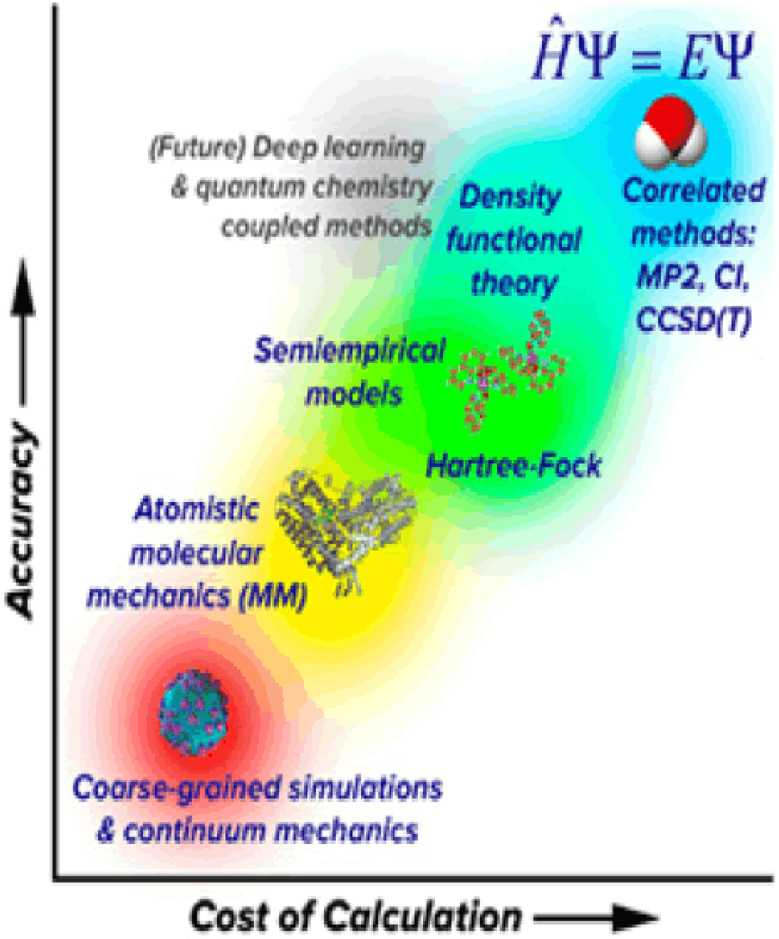


The CSP solution

- Use a computer to create hypothetical crystal structures of a given compound
- Determine the relative energy of those hypothetical structures
- The **lowest energy structure*** is the most stable polymorph
- Stop looking for polymorphs when you have found the thermodynamically stable form

- Background
- **Review of computational chemistry**
  - Molecular mechanics
  - Electronic structure theory
  - Types of calculations and systems
- Crystal structure prediction (CSP)
  - Structure generation
  - Structure ranking
- Interpretation of results
  - Methods of crystal structure comparison
  - Thermal free energy corrections
- CCDC blind tests

Review of computational chemistry

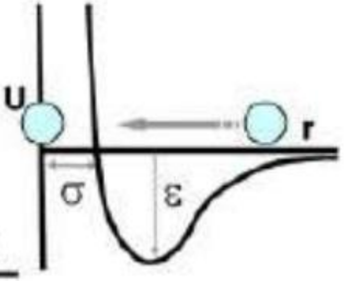

$\hat{H}\Psi = E\Psi$
(Future) Deep learning & quantum chemistry coupled methods
Density functional theory
Correlated methods: MP2, CI, CCSD(T)
Semiempirical models
Hartree-Fock
Atomistic molecular mechanics (MM)
Coarse-grained simulations & continuum mechanics
Accuracy
Cost of Calculation

Molecular mechanics

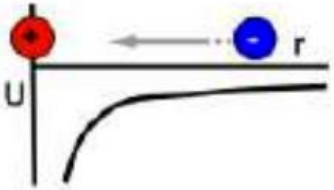
- MM, force-field (FF)
- Use classical physics equations to compute energy of an atomic system
  - Coulomb potential, Harmonic motion/Morse potential
  - $E_{tot} = E_{disp\text{-}rep} + E_{coul} + E_{bond} + E_{ang} + E_{tors}$
  - Can split into molecular geometry, and system geometry

Molecular mechanics

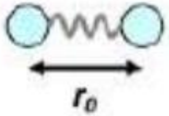
$$U = \sum_{i<j} \sum 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$



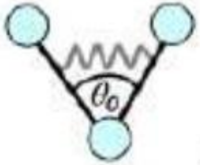
$$+ \sum_{i<j} \sum \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$



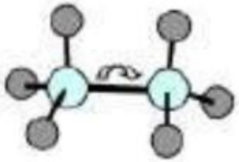
$$+ \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2$$



$$+ \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2$$



$$+ \sum_{torsions} k_\phi [1 + \cos(n\phi - \delta)]$$



Molecular mechanics

- Requires that the atoms of the system be assigned a *type* that is available within the particular MM method
  - E.g. Carbon = sp^3, aromatic, sp^2, sp
- Atom types have parameters that have been optimized based on some dataset (experimental data, computational data)

Molecular mechanics

- **AMBER** “Assisted model building with energy refinement”
  - Biological/biochem systems (nucleic acids, amino acids)
- **GAFF** “Generalized AMBER force field”
  - Computational data (MP2/6-31G*) on molecules from the CSD
- **MMFF** “Merck molecular force field”
  - Computational data (MP2/6-31G*) on organic molecules
- **OPLS** “Optimized potentials for liquid simulations”
  - Experimental properties of liquids (density, H_{vap}) and gases

Molecular mechanics

- Advantages
  - Fast - plug-and-chug parameters into simple equations
- Disadvantages
  - Accuracy - parameterization by atom type is often insufficiently precise
  - Missing atom types, or even elements

Molecular mechanics

- Tailor-made force fields
  - New MM parameters created for a given system by fitting to electronic structure theory data
  - Process is slower than using a generic MM method, but accuracy can be significantly improved for a fraction of the cost of doing *all calculations* with the higher level theory method

Electronic structure theory

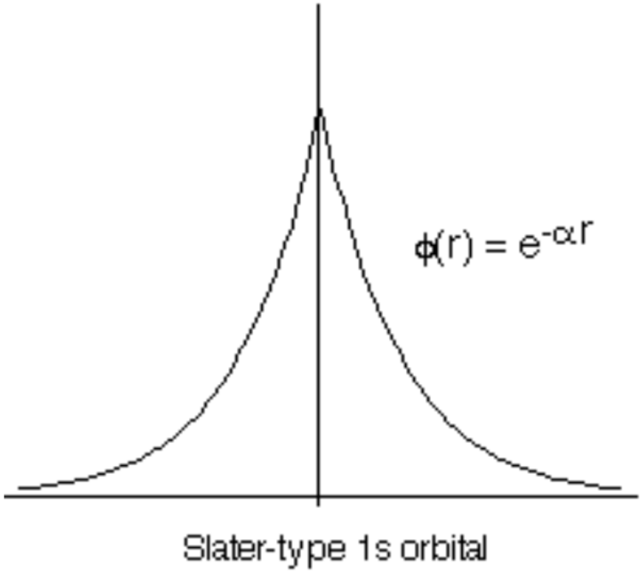
- Schrodinger equation
  - HΨ = EΨ
  - Ψ is the wavefunction – description of electron positions
  - Hamiltonian operator – operator that determines the total energy of that system

Electronic structure theory

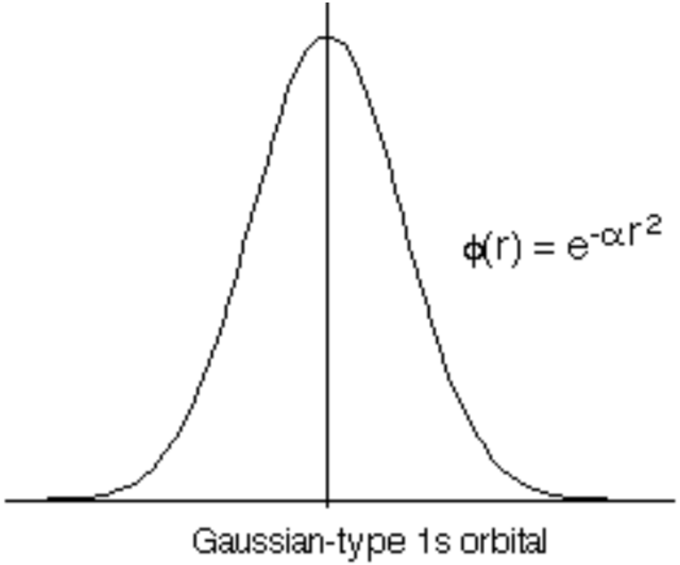
- How to compute?
  - Hamiltonian – “method” (HF, MP2, PBE, B3LYP...)
  - Wavefunction – “basis set” (6-31G*, cc-pVDZ...)
- The better the method and basis set represent the true system, the more accurate the results

Electronic structure theory

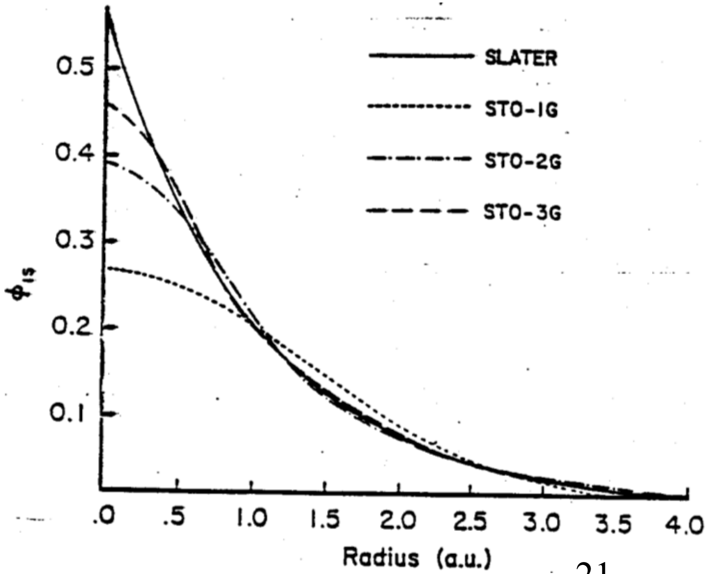
- Ψ descriptions



Slater-type 1s orbital



Gaussian-type 1s orbital



Electronic structure theory

- Hamiltonian forms (methods)
  - Wavefunction theory (HF, MP2, CCD(T), …)

    $$H = \nabla^2(r) + E_{coul}(r) + E_x(r_i, r_j)$$

    - Route to maximum accuracy is improving the basis set description of the system to describe e-e correlations
  - Density functional theory (PBE, BLYP, PW, …)

    $$H = \nabla^2(\rho) + E_{coul}(\rho) + E_{xc}(\rho)$$

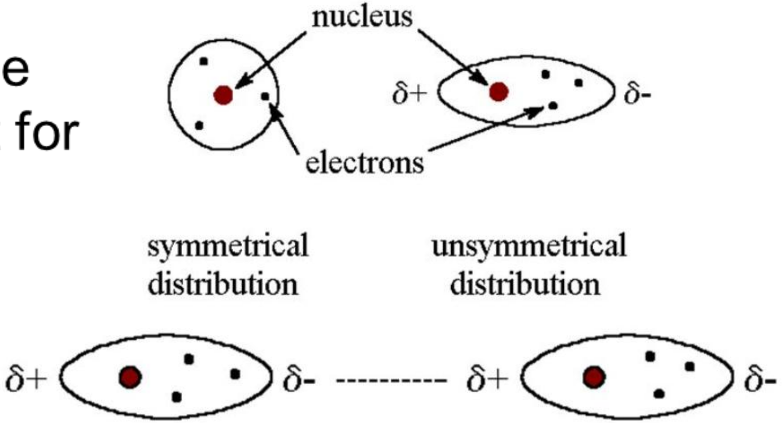
    - Route to maximum accuracy is improving the indeterminant form of the exchange-correlation functional

Electronic structure theory

- DFT (cont'n)
  - Hybrid DFT (B3LYP, PBE0,…) includes HF exchange

$$H = \nabla^2(\rho) + E_{coul}(\rho) + (\mathbf{1}-\boldsymbol{\alpha})\boldsymbol{E_x}(\boldsymbol{\rho}) + \boldsymbol{\alpha}\boldsymbol{E_x}(\boldsymbol{r_i},\boldsymbol{r_j}) + E_c(\rho)$$

  - Dispersion corrections (D3, TS, XDM,…)
    - DFT does not consider long-range election correlations that account for phenomena like dispersion



Electronic structure theory

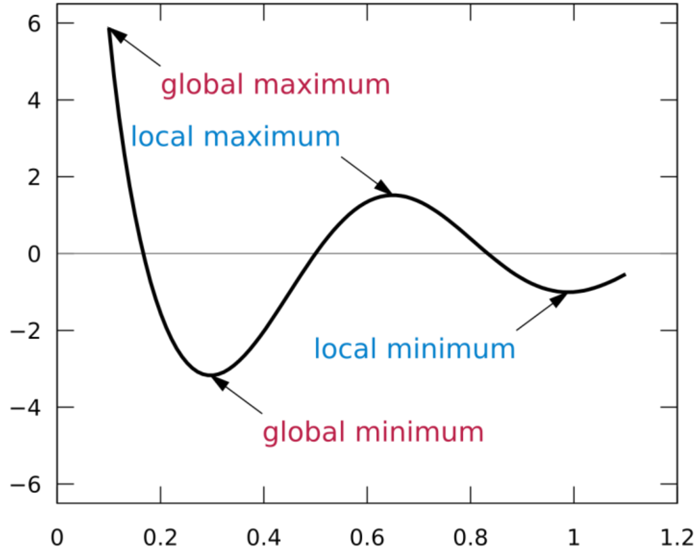
- Self-consistent field calculation (SCF)
  - Initial guess -> evaluate -> update ----> convergence
  - At convergence, the Ψ is the best representation for the system that can be obtained (for that method and basis set), and the energy of the *system in that state* is known

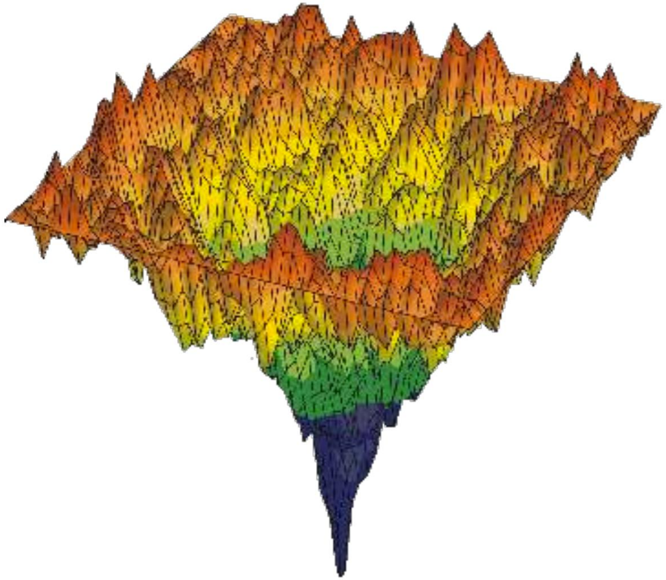
Types of calculations

- Single-point energy – static atom positions
  - MM: plug-n-chug
  - EST: SCF to optimize Ψ and obtain the energy of the system
- Geometry optimization – allow atoms to move into lowest energy position
  - Determine single-point energy of the system at various geometries (10s to 100s) and identify the minimum energy state (based on initial conditions)
  - Use some algorithm to direct changes in the geometry

Types of calculations

- Geometry optimizations (cont'n)
  - May need to start from different geometries (initial conditions) in order to reach the **global minimum**





Types of systems

- In-vacuo “gas phase” vs periodic boundary calculations
- In-vacuo (e.g. Gaussian) – finite atom count (Ψ)
- Periodic-boundary – solids, infinite atom count (Ψ)
  - Calculations done on the unit cell with Fourier transform tricks to account for the infinite nature of a crystal
    - EST: planewave basis set + k-point sampling
    - MM: Ewald summation

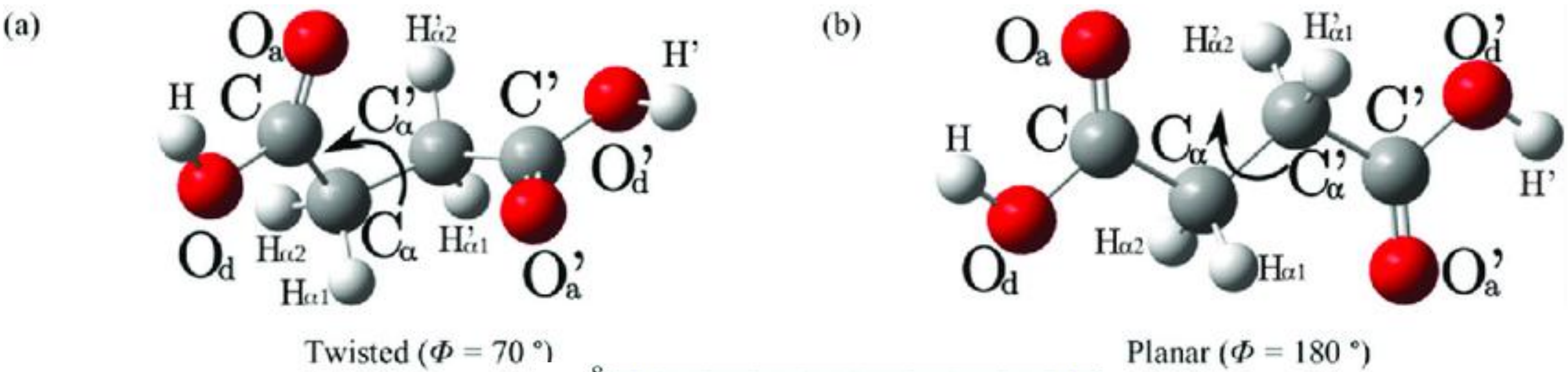
- Background
- Review of computational chemistry
  - Molecular mechanics
  - Electronic structure theory
  - Types of calculations and systems
- **Crystal structure prediction**
  - Structure generation
  - Structure ranking
- Interpretation of results
  - Methods of crystal structure comparison
  - Thermal free energy corrections
- CCDC blind tests

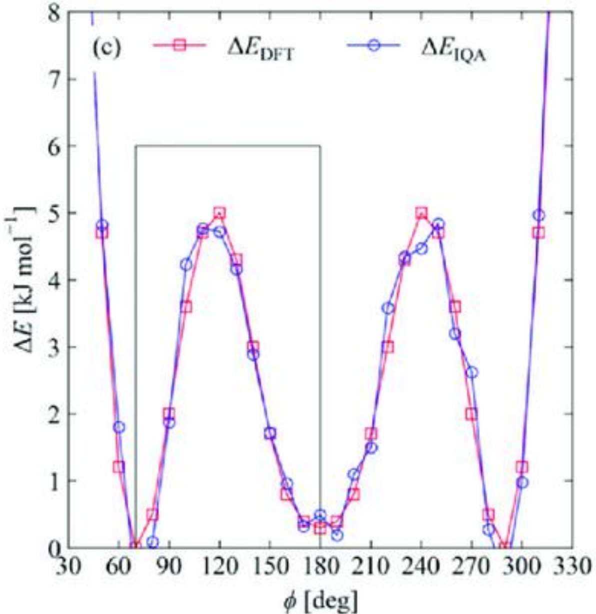
Structure generation

1. Molecular conformation
2. Crystal structure search parameters
3. Structure generation algorithms

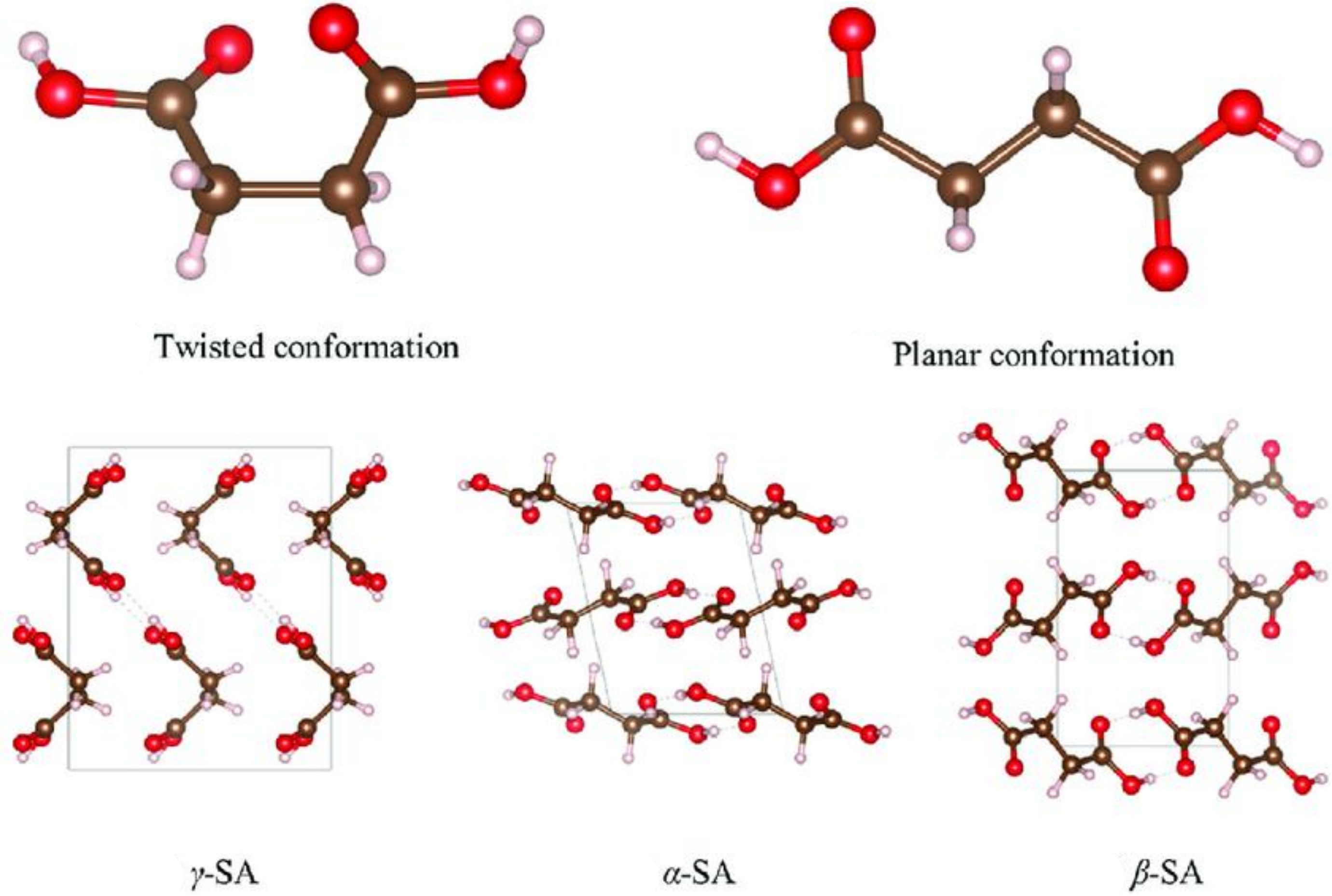
Structure generation

- Conformation searching commonly done with hybrid DFT in-vacuo optimizations
- Starting point is important
  - Energy barrier won’t be overcome on optimization
  - intra-/inter-molecular H-bonding





Lucaioli P, Nauha E, Gimondi I, Price LS, Guo R, Iuzzolino L, Singh I, Salvalaglio M, Price SL, Blagden N. *CrystEngComm*. **2018,** *20*, 3971-3977.
la Vega AS, Duarte LJ, Silva AF, Skelton JM, Rocha-Rinza T, Popelier PL. *PCCP*. **2022**, 24, 11278-11294.



Twisted conformation

Planar conformation

γ-SA

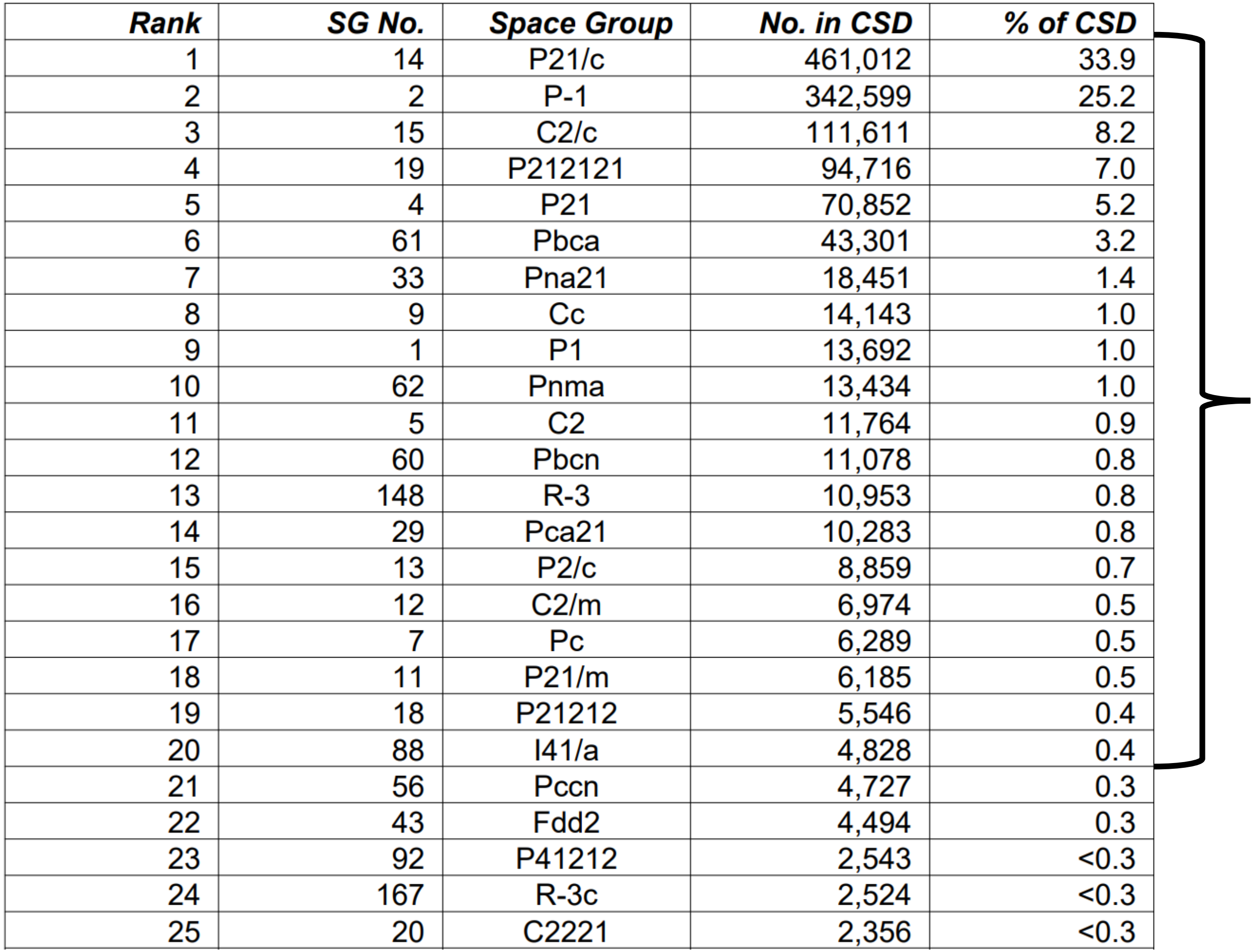
α-SA

β-SA

Lucaioli P, Nauha E, Gimondi I, Price LS, Guo R, Iuzzolino L, Singh I, Salvalaglio M, Price SL, Blagden N. *CrystEngComm*. **2018,** *20*, 3971-3977.
la Vega AS, Duarte LJ, Silva AF, Skelton JM, Rocha-Rinza T, Popelier PL. *PCCP*. **2022**, 24, 11278-11294.

Structure generation

- Search space for crystal structures of a **molecule** is vast
  - Unit cell dimensions (a, b, c, α, β, γ)
  - Position and orientation of the **molecule** within the unit cell
    - Relative position and orientations of **molecular** components if multi-component
  - **Conformer** used
    - Number of rotate-able bonds
  - **Space groups**
    - Commonly, the space groups covering 90-95% of the CSD are used
    - Consideration of molecular symmetry, steriochemistry
  - **Asymmetric unit**
    - Z' = 1 is the default level, Z' = 2 is a multifold increase in the search space (nxm)



Rank	SG No.	Space Group	No. in CSD	% of CSD
1	14	P21/c	461,012	33.9
2	2	P-1	342,599	25.2
3	15	C2/c	111,611	8.2
4	19	P212121	94,716	7.0
5	4	P21	70,852	5.2
6	61	Pbca	43,301	3.2
7	33	Pna21	18,451	1.4
8	9	Cc	14,143	1.0
9	1	P1	13,692	1.0
10	62	Pnma	13,434	1.0
11	5	C2	11,764	0.9
12	60	Pbcn	11,078	0.8
13	148	R-3	10,953	0.8
14	29	Pca21	10,283	0.8
15	13	P2/c	8,859	0.7
16	12	C2/m	6,974	0.5
17	7	Pc	6,289	0.5
18	11	P21/m	6,185	0.5
19	18	P21212	5,546	0.4
20	88	I41/a	4,828	0.4
21	56	Pccn	4,727	0.3
22	43	Fdd2	4,494	0.3
23	92	P41212	2,543	<0.3
24	167	R-3c	2,524	<0.3
25	20	C2221	2,356	<0.3

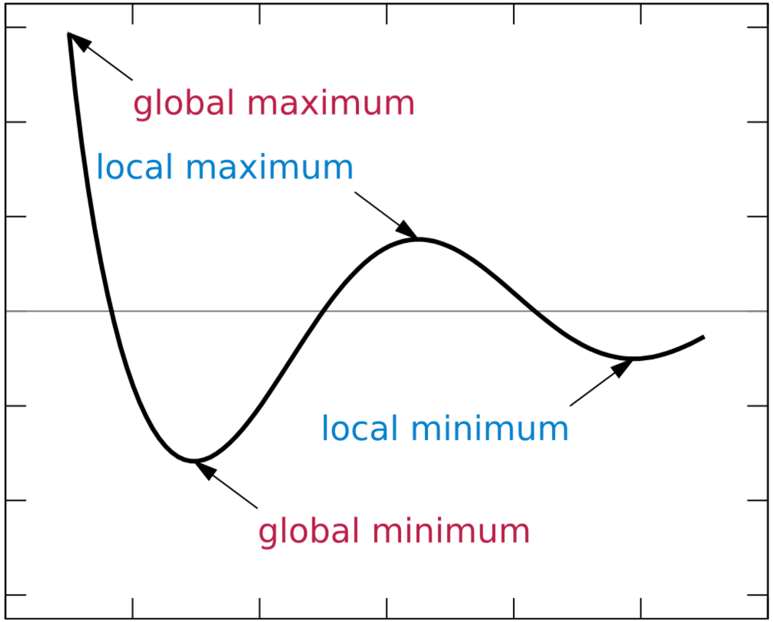
93.4%

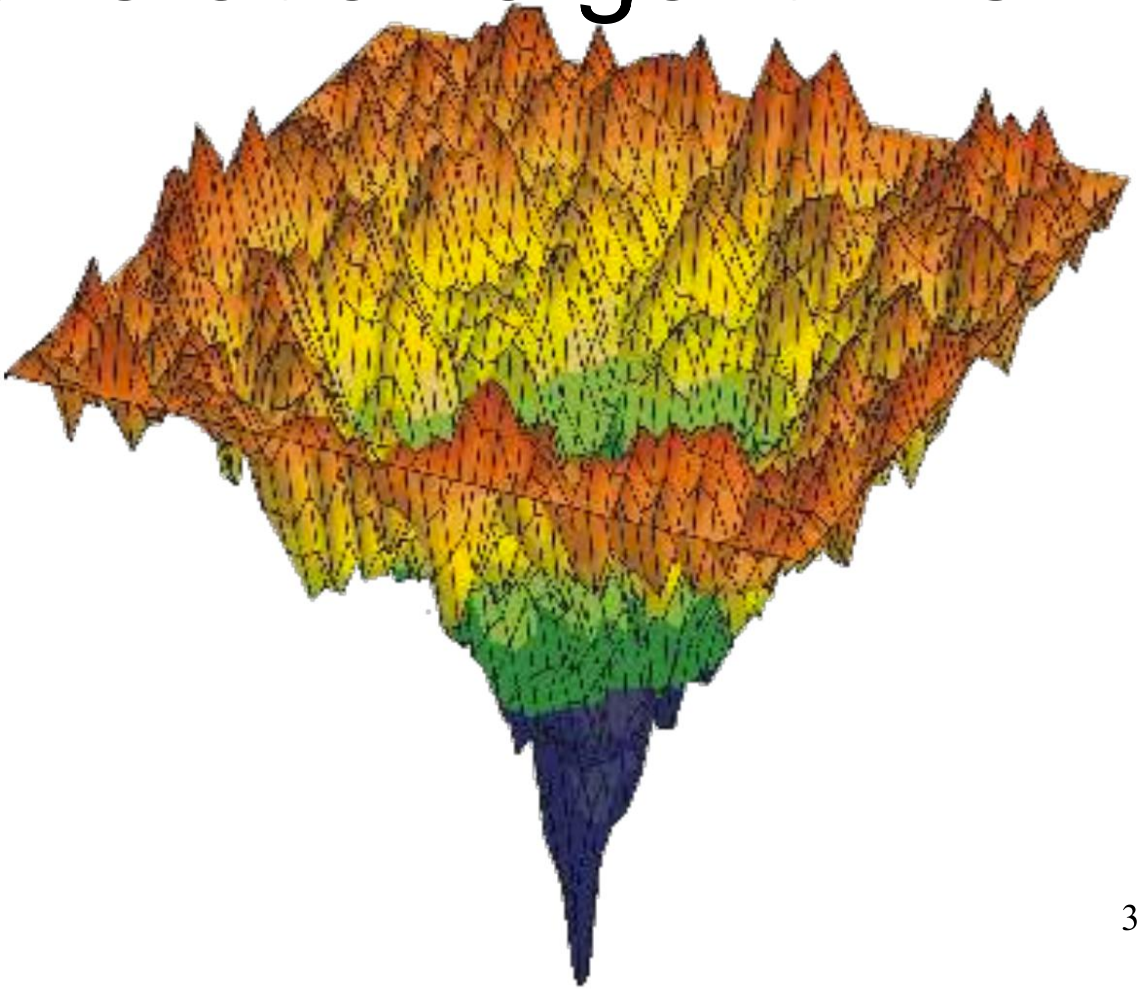
https://www.ccdc.cam.ac.uk/media/CSD-Space-Group-Statistics-Space-Group-Frequency-Ordering-2025.pdf

Structure generation algorithms

- Random
  - Inefficient with large search space
  - Theoretically explores the entire search space (given enough time)
- Biased
  - Will not cover the entire search space
  - More quickly identifies low-energy regions of the search space
  - e.g. simulated annealing, parallel tempering, genetic algorithms, particle-swarm optimization…
  - Requires scoring function (energy calculation)
  - Need tricks to get out of local minima

Structure generation algorithms



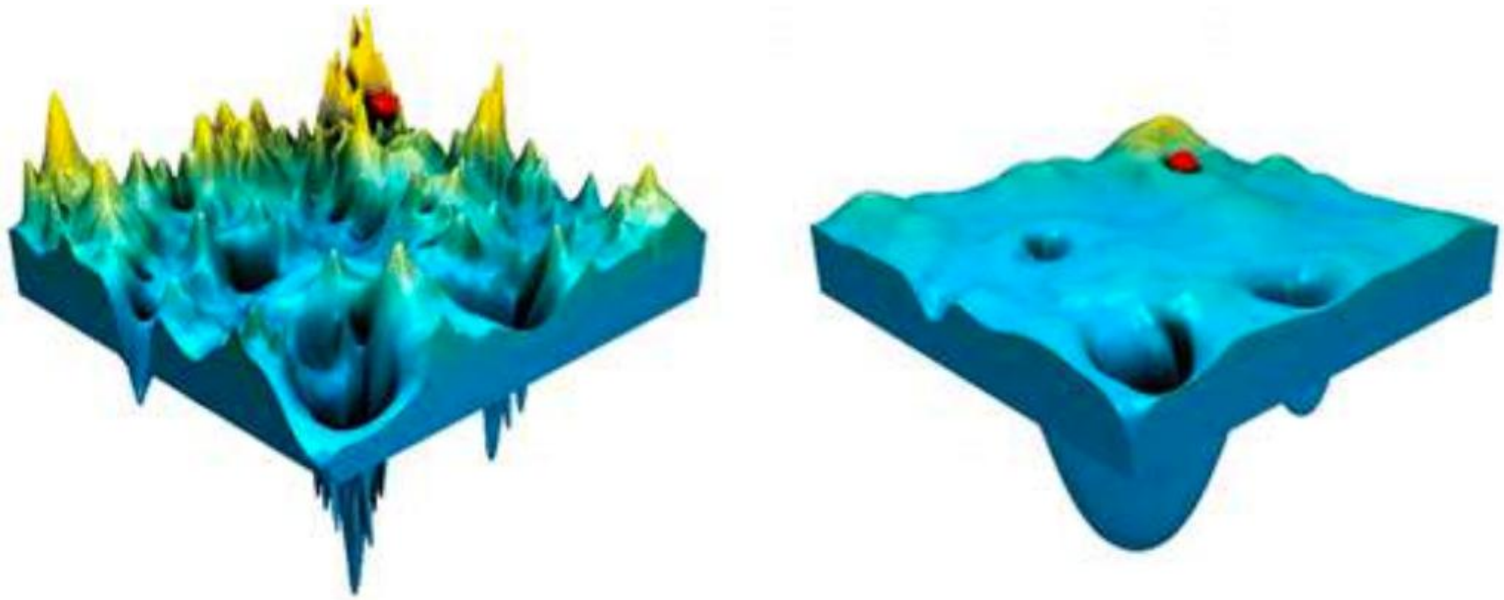


Structure generation with biasing algorithms

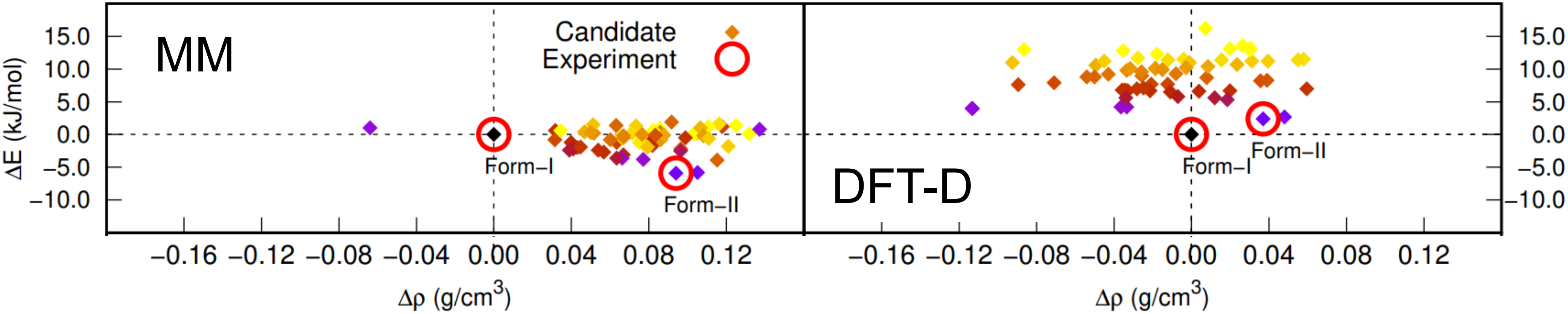
- Need to compute the energy of every generated crystal structure in order to bias the algorithm
  - Must be fast
- MM methods used
- Poor-moderate accuracy

Structure generation with biasing algorithms

- Different methods will give different PES



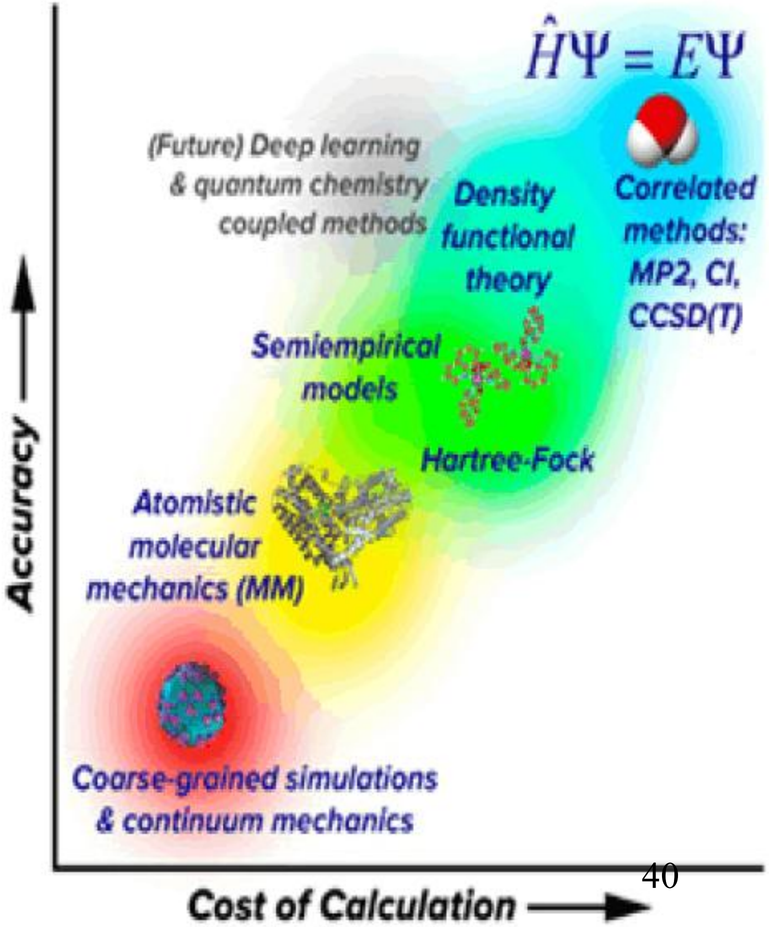
Structure generation



Form I is more stable than Form II

Structure ranking

- What level of theory is required for sufficiently accurate results, and can we develop tricks to make it faster?

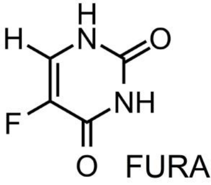


$\hat{H}\Psi = E\Psi$
(Future) Deep learning & quantum chemistry coupled methods
Density functional theory
Correlated methods: MP2, CI, CCSD(T)
Semiempirical models
Hartree-Fock
Atomistic molecular mechanics (MM)
Coarse-grained simulations & continuum mechanics
Accuracy
Cost of Calculation

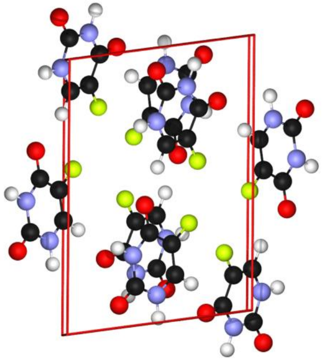
Structure ranking

- Select a cutoff energy in the MM CSP landscape to run DFT-D optimization/single-point energy calculations on
- Often use a funnel with a series of cutoffs
  - +1 million MM calculations in structure generation
  - -> 500k more tailored/advanced MM optimization
  - -> 250k semi-empirical (DFTB, HF-3c)
  - -> 1k (hybrid)DFT-D single point
- Still a risk of leaving an experimentally observed structure behind at any of the steps

Structure ranking

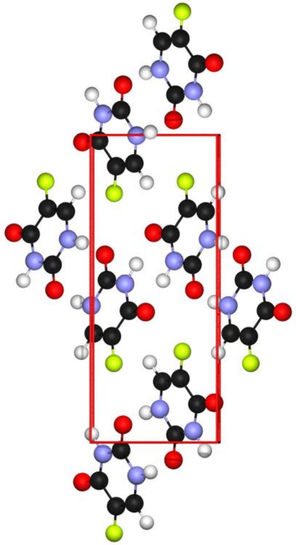


FURACL



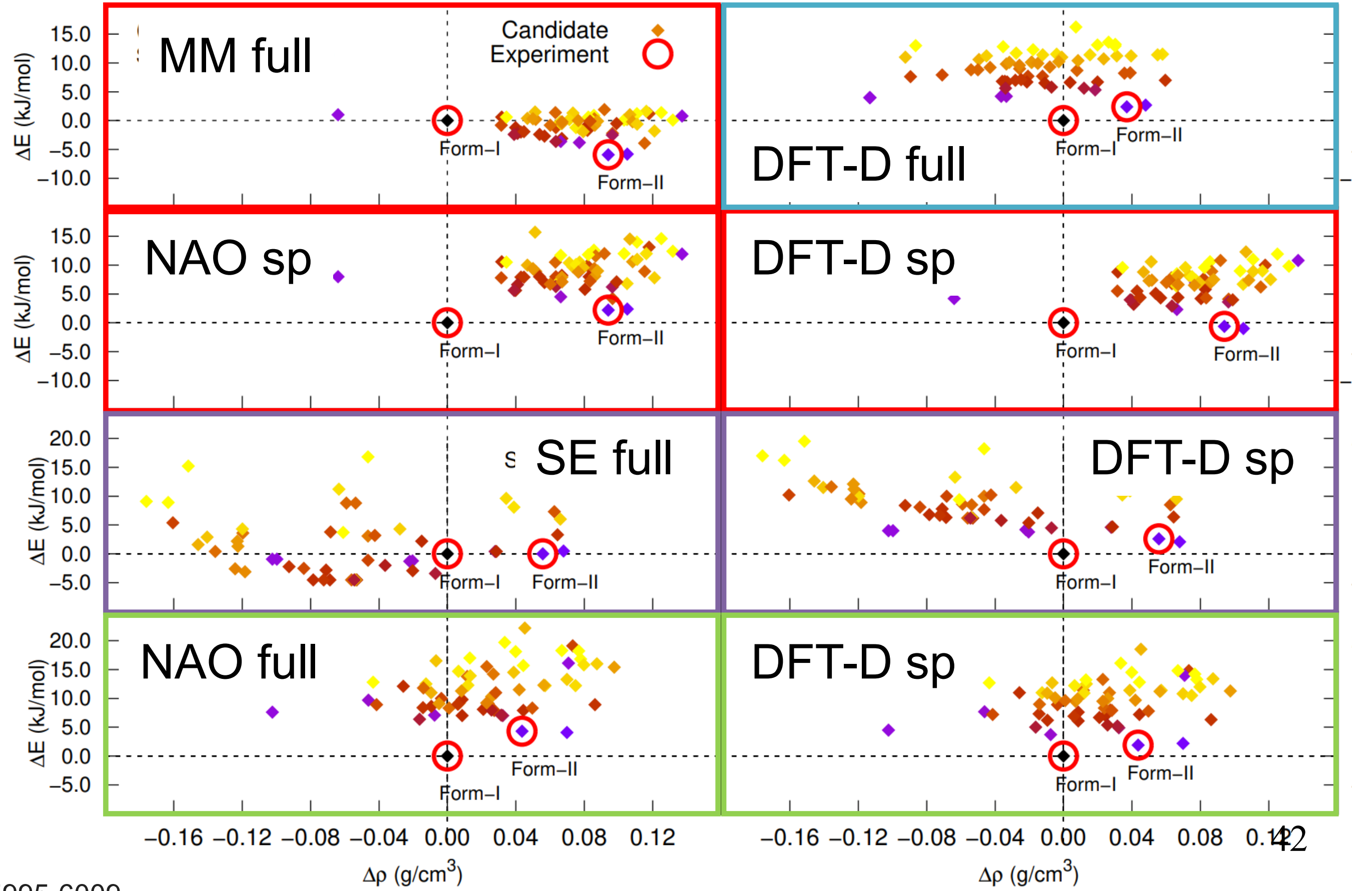
Form-I

stable



Form-II

metastable

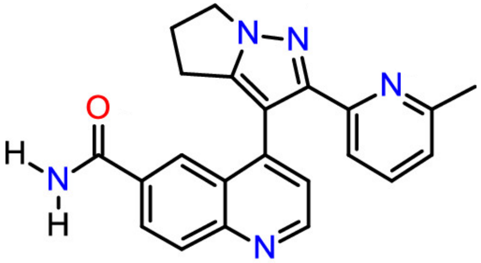


LeBlanc LM, Johnson ER. *CrystEngComm*. **2019,** *21*, 5995-6009.

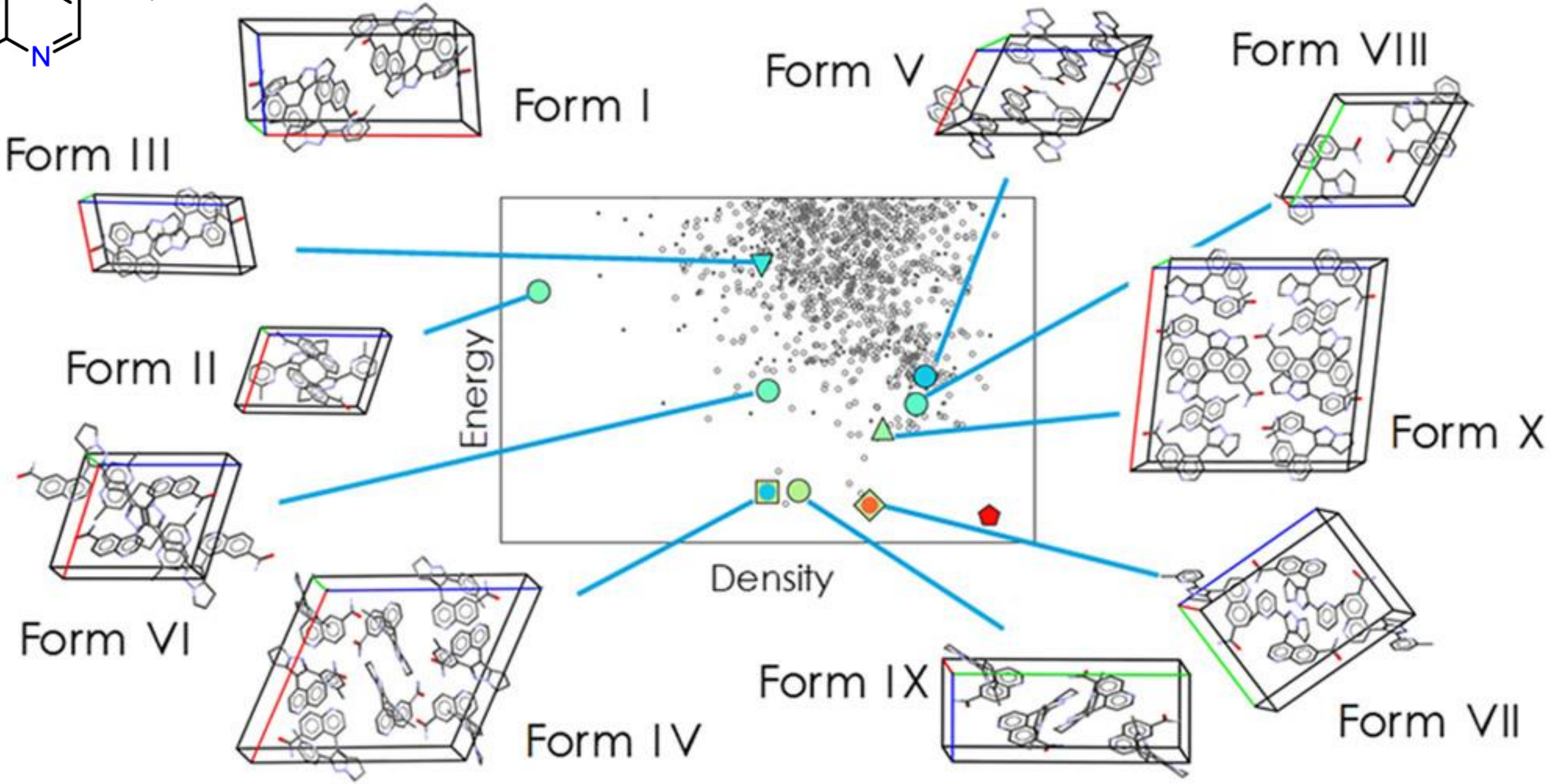
- Background
- Review of computational chemistry
  - Molecular mechanics
  - Electronic structure theory
  - Types of calculations and systems
- Crystal structure prediction
  - Structure generation
  - Structure ranking
- **Interpretation of results**
  - Methods of crystal structure comparison
  - Thermal free energy corrections
- CCDC blind tests

CSP protocol benchmarking

- How to assess CSP methods?
    1. Does it find the experimental structure(s)?
    2. Does it correctly rank the stability of the experimental structures?
        a) Is the experimental thermodynamic form the lowest energy structure?
- Step 1 – identify whether the experimental structures were generated i.e. crystal structure comparisons
    - But we have a lot of structures to compare!



Galunisertib



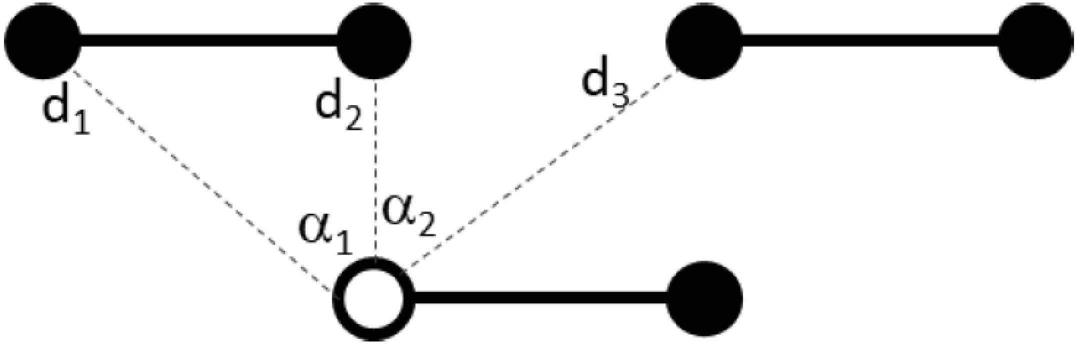
Bhardwaj RM, *et al., J. Am. Chem. Soc.* **2019**, *141*, 13887-13897

Methods of crystal structure comparison

- Atomic position-based methods
  - Advantage of being more accurate with changes in temperature (CSP_0 to SC-XRD structure) without tricks
- PXRD-based methods
  - Advantage of being able to compare to experimental PXRD

Methods of crystal structure comparison

- Atomic position-based methods

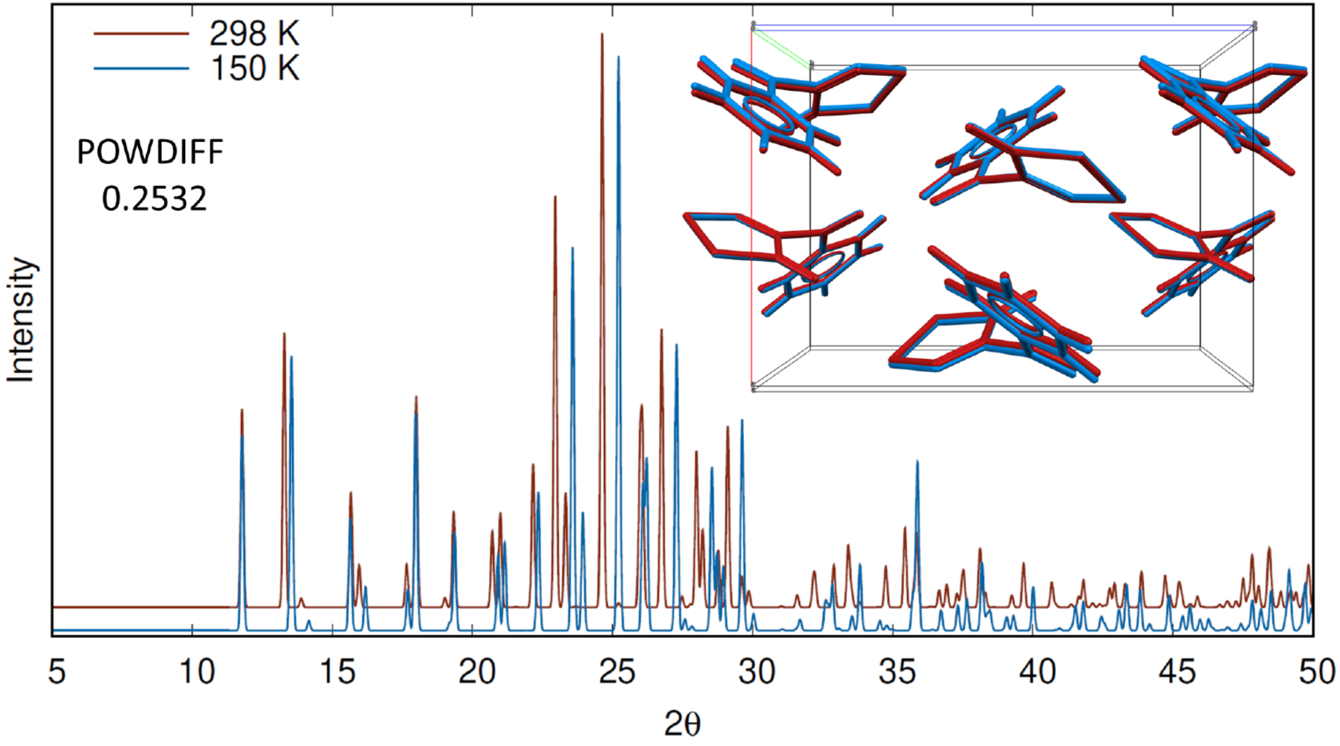


- COMPACK (CCDC) N/M and RMSD

Chisholm JA and Motherwell S, *J. Appl. Cryst.* **2005**, *38*, 228-231

Methods of crystal structure comparison

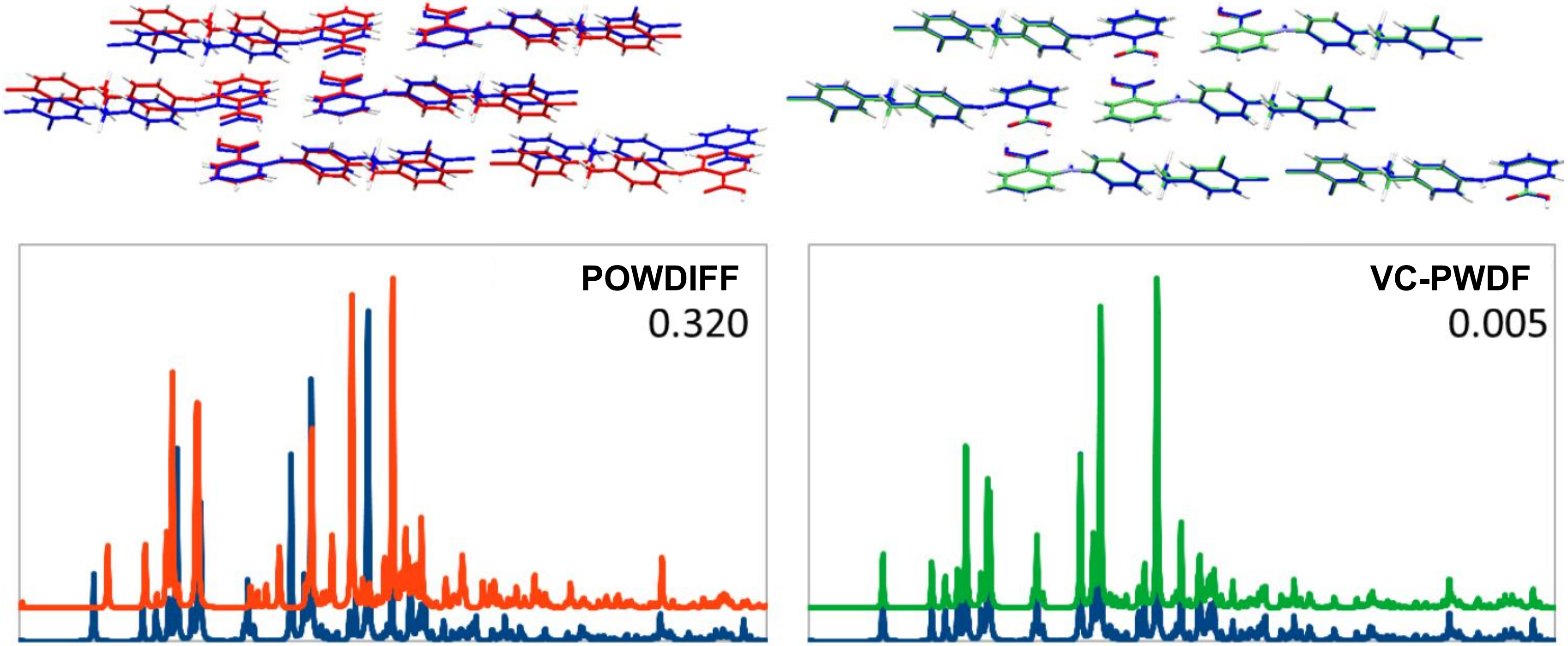
- PXRD-based methods



20/20 match
RMSD(20)
0.108 Å

Methods of crystal structure comparison

- PXRD-based methods - tricks

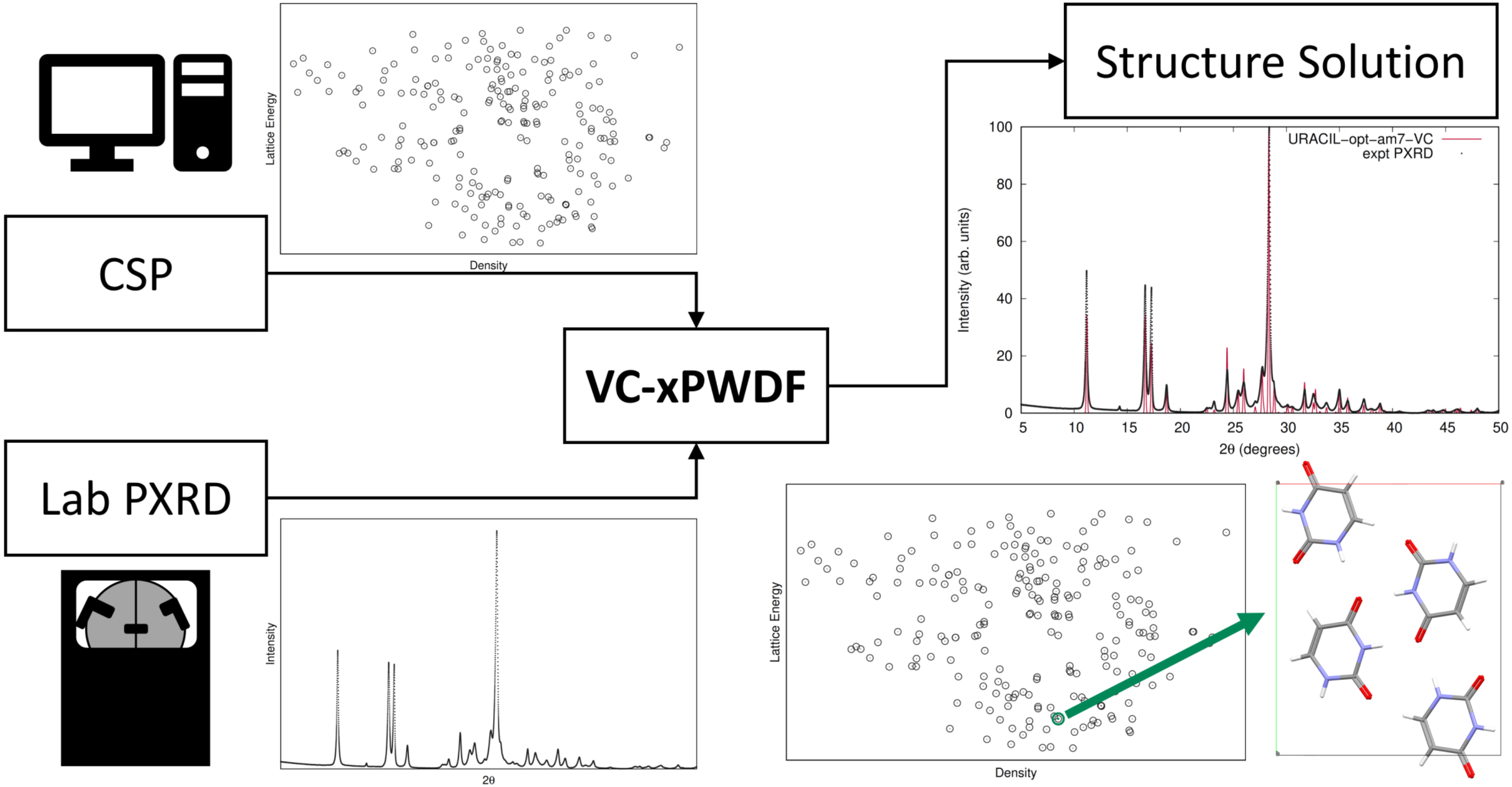


Mayo RA, Otero de la Roza A, Johnson ER, *CrystEngComm*, **2022**, *24*, 8326–8338.

Methods of crystal structure comparison

- Comparison methods are used throughout the CSP protocol
  - Identifying duplicates in structure generation
  - Identifying duplicates post-optimization
  - Reduce number of repeated structures passing through to more expensive computations

Crystal structure determination



CSP

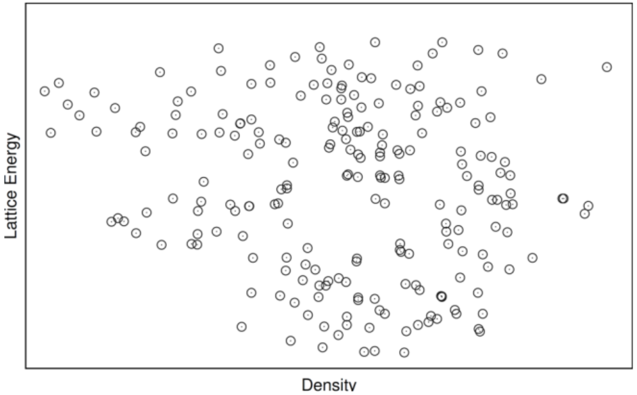
Lab PXRD

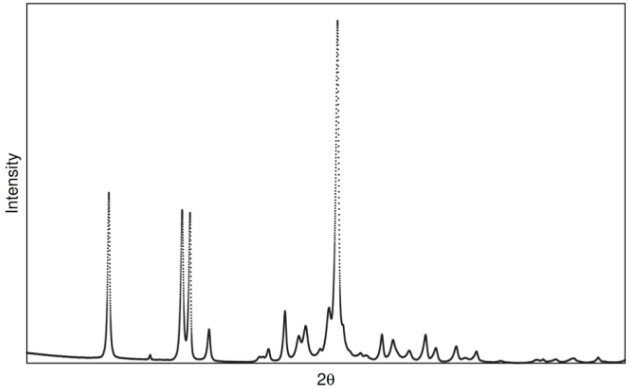
VC-xPWDF

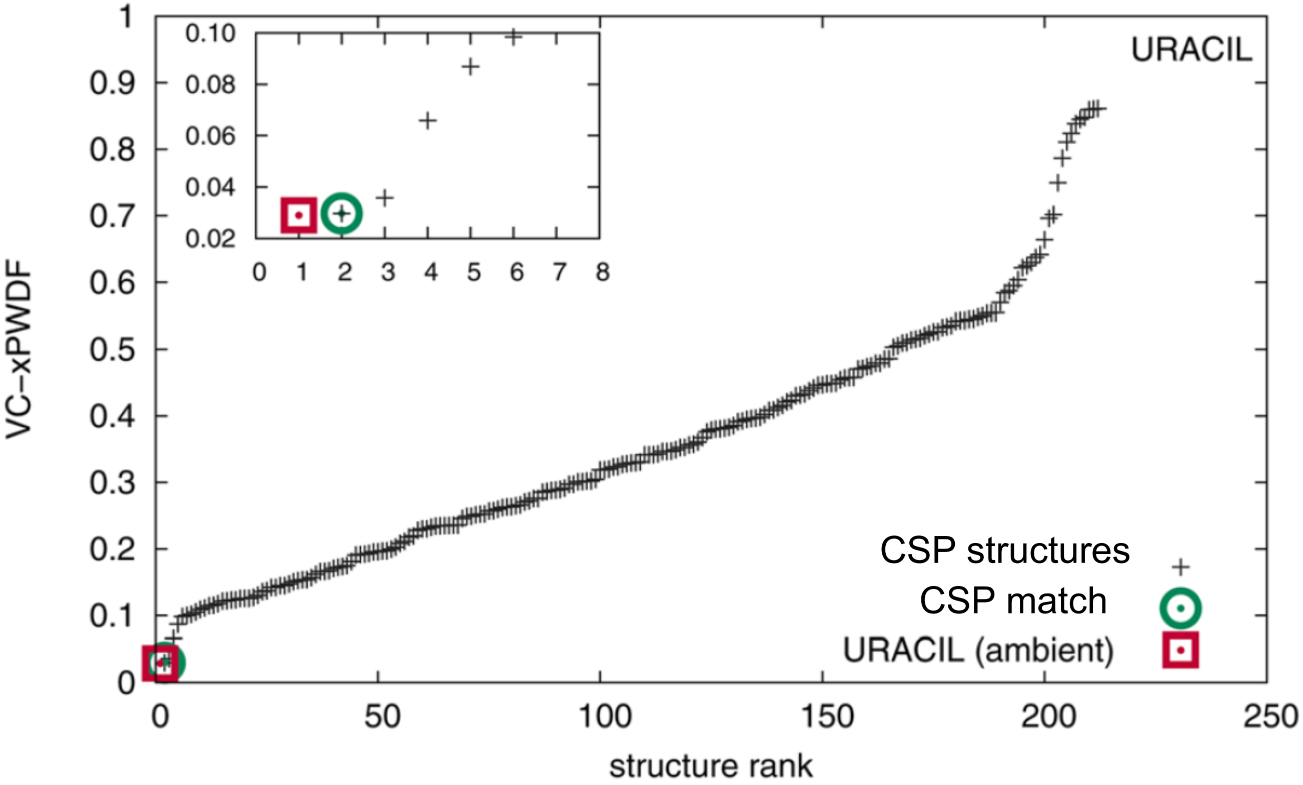
Structure Solution

Mayo RA, Marczenko KM, Johnson ER. *Chem. Sci.*, **2023**, *14*, 4777-4785

Crystal structure determination

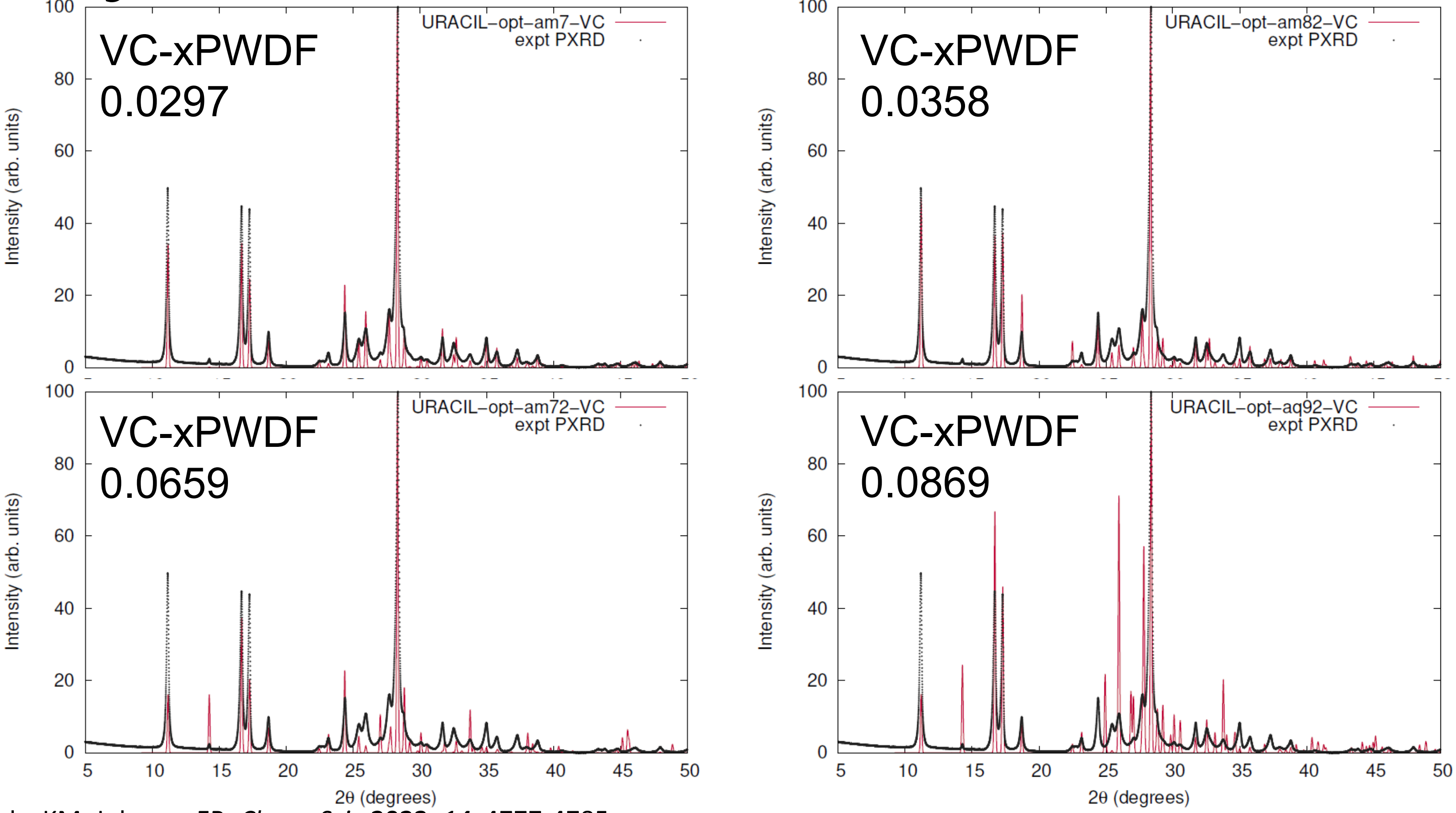






Mayo RA, Marczenko KM, Johnson ER. *Chem. Sci.*, **2023**, *14*, 4777-4785

Crystal structure determination



Mayo RA, Marczenko KM, Johnson ER. *Chem. Sci.*, **2023**, *14*, 4777-4785

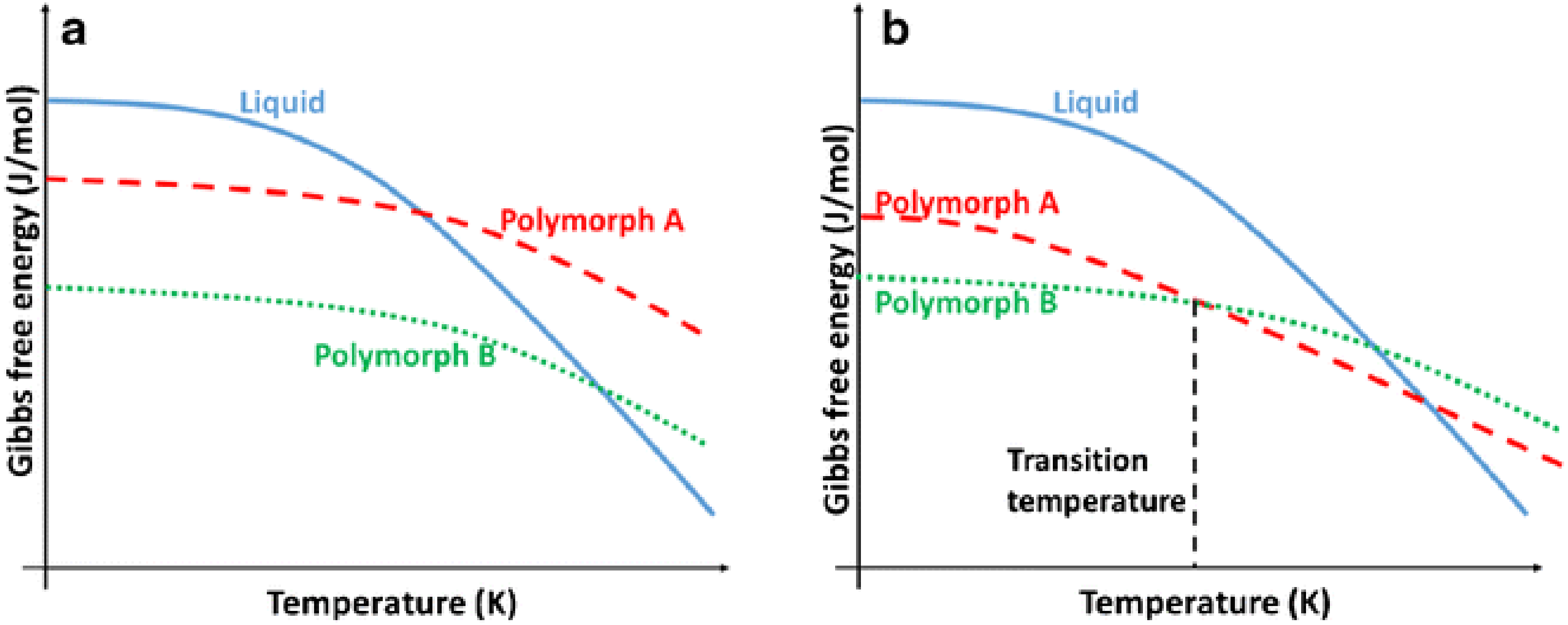
Temperature effects

- CSP “structure-energy landscape” = static lattice electronic energies
- Sometimes called the “CSP_0” landscape, ideas of 0 Kelvin temperature

Temperature effects

- Monotropic vs enantionotropic polymorps

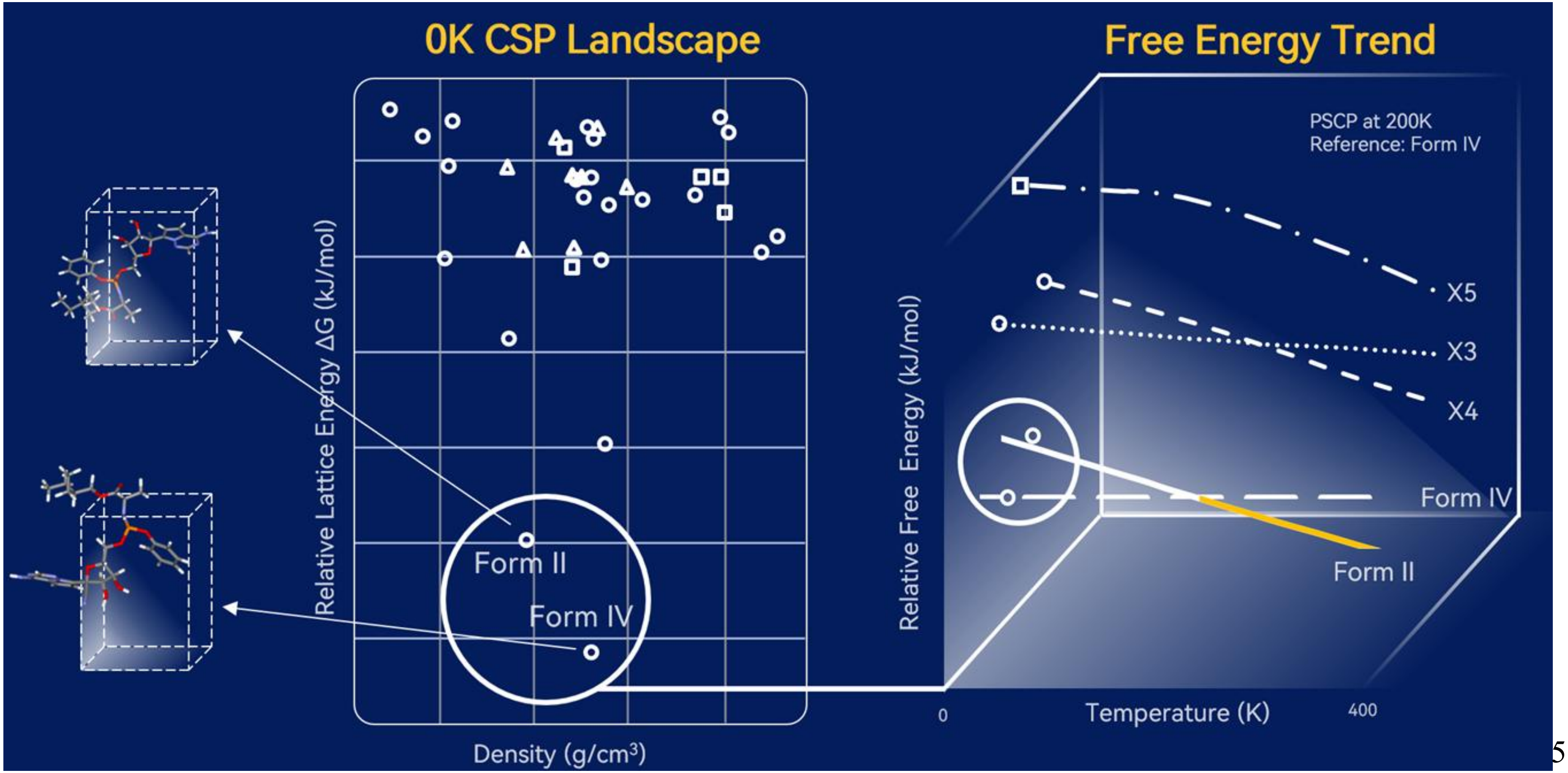
$$\Delta G = \Delta H - T\Delta S$$



Free energy corrections

- Methods for free energy correction to ambient conditions – computationally demanding!
- Phonon calculations yield S contribution
  - (Quasi)-Harmonic approximation, (Q)HA
  - Pseudo-supercritical path, PSCP
  - Molecular dynamics, MD

Temperature effects



Sivakumar S, *et al. RSC Advance.* **2021**, *11*, 17408.

- Background
- Review of computational chemistry
  - Molecular mechanics
  - Electronic structure theory
  - Types of calculations and systems
- Crystal structure prediction
  - Structure generation
  - Structure ranking
- Interpretation of results and applications
  - Thermal free energy corrections
  - Methods of crystal structure comparison
- **CCDC blind tests**

CCDC blind tests

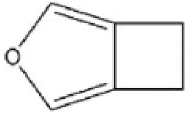
- Provide participants with molecular 2D diagram
- Participants do CSP and submit the lowest energy structure as the predicted crystal structure
- CCDC compares the submitted structure from each group to the known experimental structure that was not shared with the participants

* See references at the end of the deck for all CCDC blind test citations

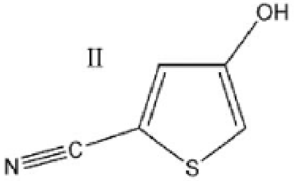
CCDC blind tests 1-4

- First run in 1999
- For BT1-4, goal was to submit the known crystal structure
- Mostly simple, rigid molecules
- Participants only used MM methods
- Low success rate

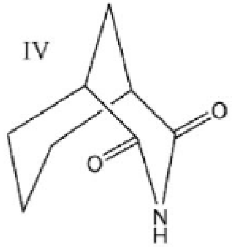
CCDC blind tests 1-4



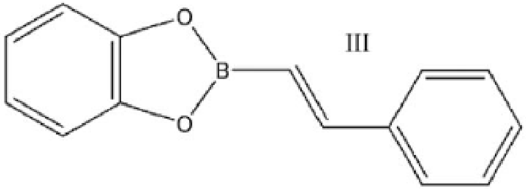
I



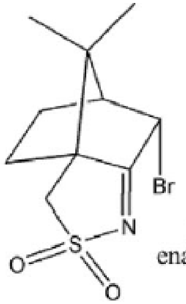
II



IV

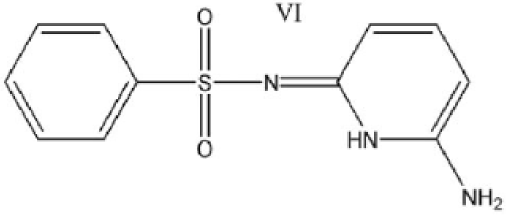


III



V

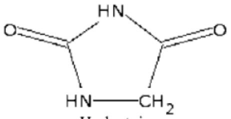
Pure enantiomer



VI

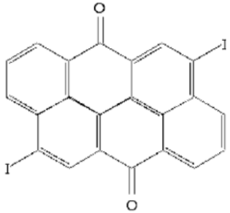
VII Propane

(VIII)



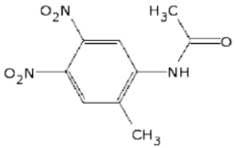
Hydantoin

(IX)



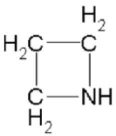
2,9-Di-iodo-anthanthrone

(X)



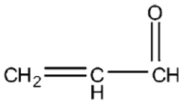
2-Acetamido-4,5-dinitrotoluene

(XI)



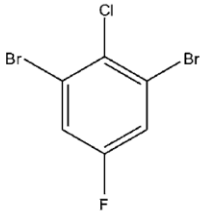
Azetidine

(XII)



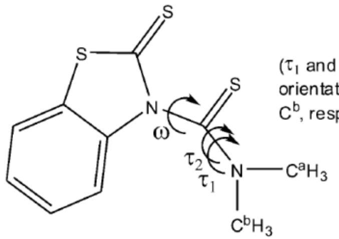
2-Propenal

(XIII)



1,3-Dibromo-2-chloro-5-fluorobenzene

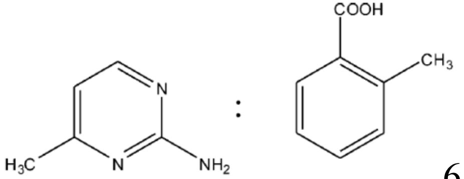
(XIV)



(τ_1 and τ_2 refer to the orientations of C^a and C^b, respectively)

N-(Dimethylthiocarbamoyl)benzothiazole-2-thione

(XV)

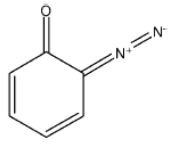


2-Amino-4-methylpyrimidine: 2-methylbenzoic acid

BT5

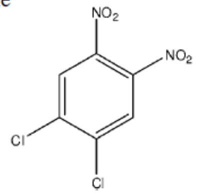
- Submit top 3 structures
- Group 11 used DFT-D and predicted all targets correctly

(XVI)



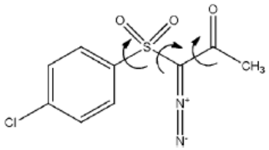
2-Diazo-3,5-cyclohexadiene-1-one

(XVII)



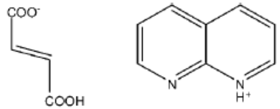
1,2-Dichloro-4,5-dinitrobenzene

(XVIII)



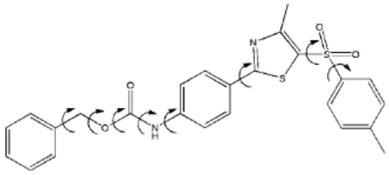
(1-((4-Chlorophenyl)sulfonyl)-2-oxopropylidene)diazenium

(XIX)



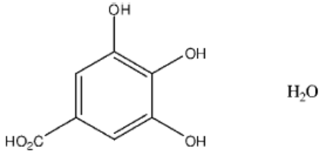
1,8-Naphthyridinium fumarate

(XX)



Benzyl-(4-(4-methyl-5-(*p*-tolylsulfonyl)-1,3-thiazol-2-yl)phenyl)carbamate

(XXI)



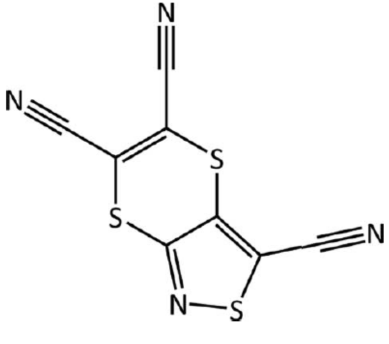
Gallic acid monohydrate

BT6 (2014-2015)

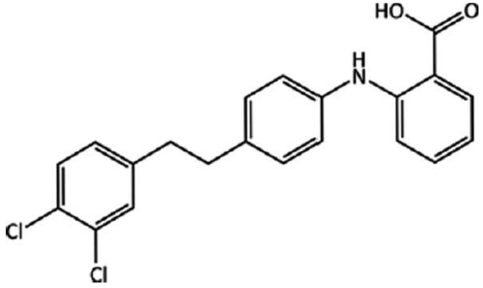
- Goal to submit a landscape (100 structures), not just a couple structure
- Many more participants, many more methods used
- More complex molecules, polymorphic, Z’=2 cases, multi-component
- DFT-D continues to be a top performer
- Few instances of free energy corrections

BT6 (2014-2015)

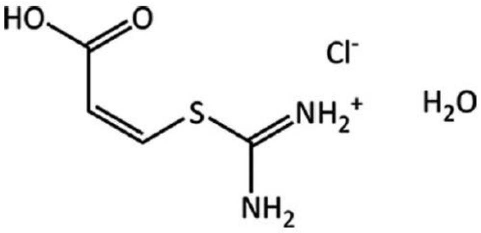
(XXII)



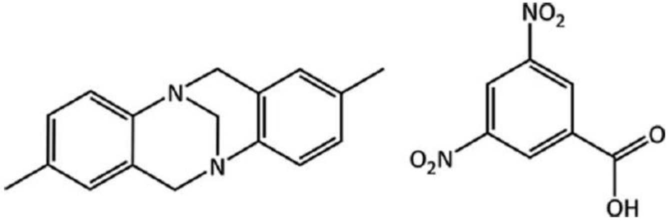
(XXIII)



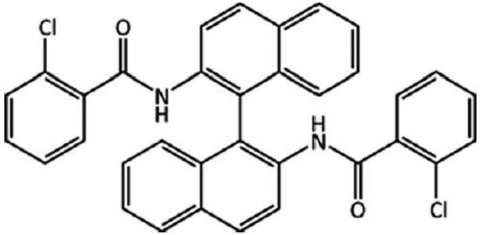
(XXIV)



(XXV)

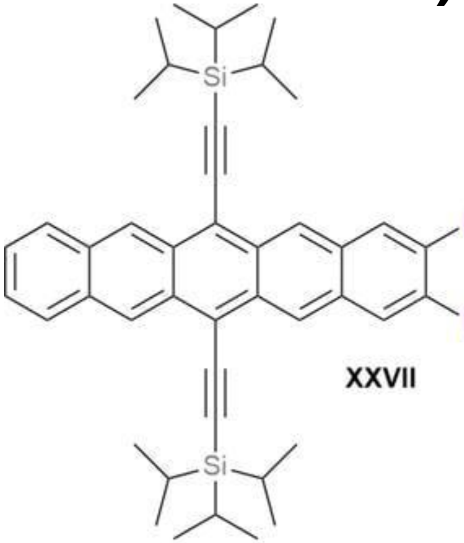


(XXVI)

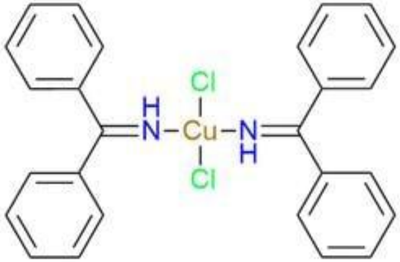


BT7 (2021-2023)

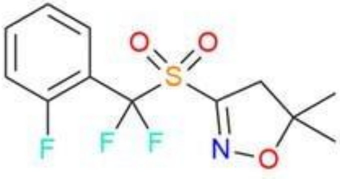
- Two stages
  - Structure generation
  - Structure ranking
- 28 groups participated



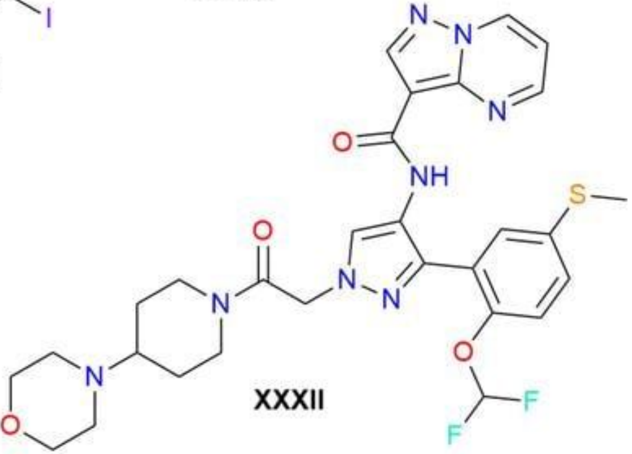
XXVII



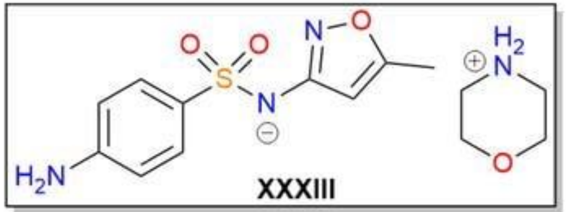
XXVIII



XXXI



XXXII



XXXIII

BT7 (2021-2023)

- Molecular complexity was significantly increased
- Two groups successfully predicted all experimental structures
- Thermal free energy corrections were necessary to correctly rank 2 of the polymorphic systems
- Significant issues with COMPACK comparisons
  - XXIX – insufficient cluster size (required 70 molecules)
  - XXVIII – highly branched

Outstanding challenges

- Disordered crystal
- High Z’ (>2) searches are not routine
- Rare space groups
- High complexity systems

References and additional reading

Molecular mechanics

https://wiki.lct.jussieu.fr/workshop/images/4/44/School_cttc2019_mm_md_compressed.pdf

Comparison methods

Mayo RA, Otero de la Roza A, Johnson ER, *CrystEngComm*. **2022**, *24*, 8326-8338.
(Mayo RA, Johnson ER. CH9 Quantitative Crystal Structure Comparison *in* Advances in Organic Crystal Chemistry, Springer, 2025 – in press)

Thermal free energy corrections

PSCP – Yang M, *et al*. *Cryst. Growth Des*. **2020**, *20*, 5211–5224
Weatherby J, *et al*. *J. Chem. Phys*. **2022**, *156*, 114108.

CCDC CSP blind tests

1.Lommerse JPM, *et al. Acta Cryst*. **2000**,*B56*, 697–714.
2.Motherwell WDS, *et al*. *Acta Cryst*. **2002**, *B58*, 647-661.
3.Day GM, *et al. Acta Cryst*. **2005**, *B61*, 511-527.
4.Day GM, *et al. Acta Cryst*. **2009**, *B65*, 107-125.
5.Bardwell DA, *et al. Acta Cryst*. **2011**, *B67*, 535-551.
6.Reilly AM, *et al. Acta Cryst*. **2016**, *B72*, 439-459.
7.Huniset LM, *et al*. *Acta Cryst*. **2024**, *B80*, 517-574.

CSP workflow overview

1 Crystal Structure Search

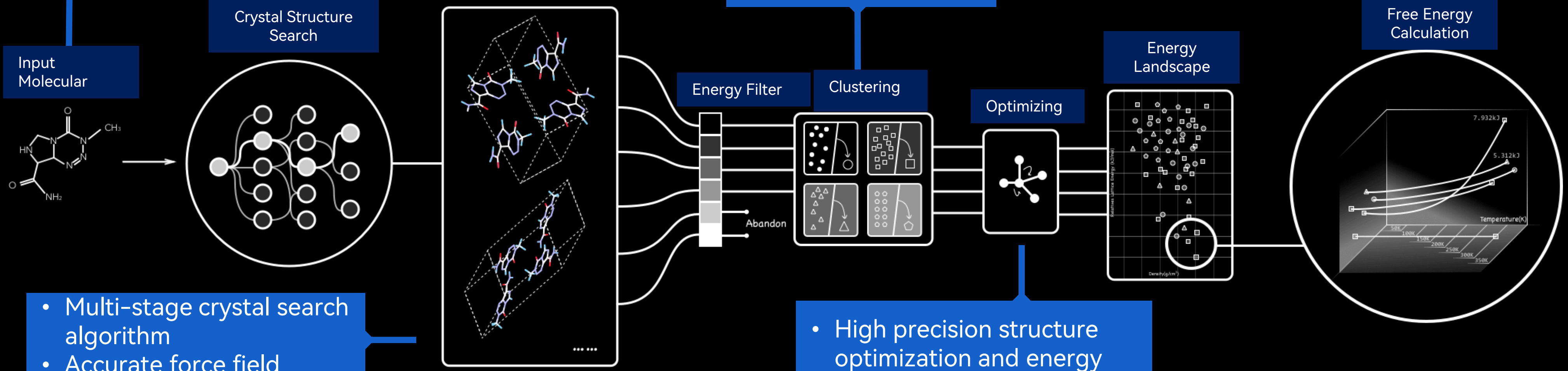
- Input: 2D diagram
- Automatic conformation analysis

2 Energy Ranking

- Ultra-fast crystal structure clustering

3 Stability Evaluation

- Drug product stability
- Dissolution
- Bioavailability



Input Molecular

Crystal Structure Search

Structure Pool

Energy Filter

Abandon

Clustering

Optimizing

Energy Landscape

Free Energy Calculation

- Multi-stage crystal search algorithm
- Accurate force field training

- High precision structure optimization and energy ranking

2~8 Weeks

Ref: Cryst. Growth Des, 2018, 18, 11, 6891−6900