Exploration and exploitation of crystal polymorph landscape

Hari Muddana

OpenEye, Cadence Molecular Sciences

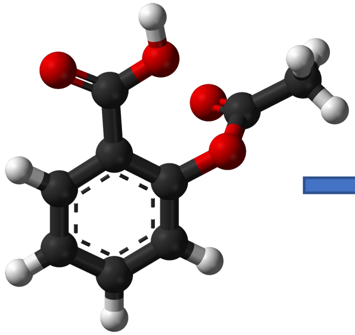
OpenEye SCIENTIFIC

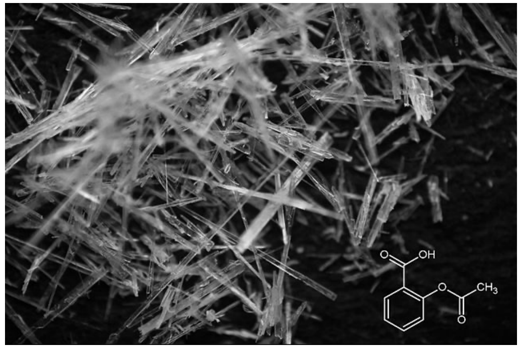
CADENCE MOLECULAR SCIENCES

Drug development pipeline

Target identification → Lead generation → Lead optimization → Pre-formulation → Clinical Trials

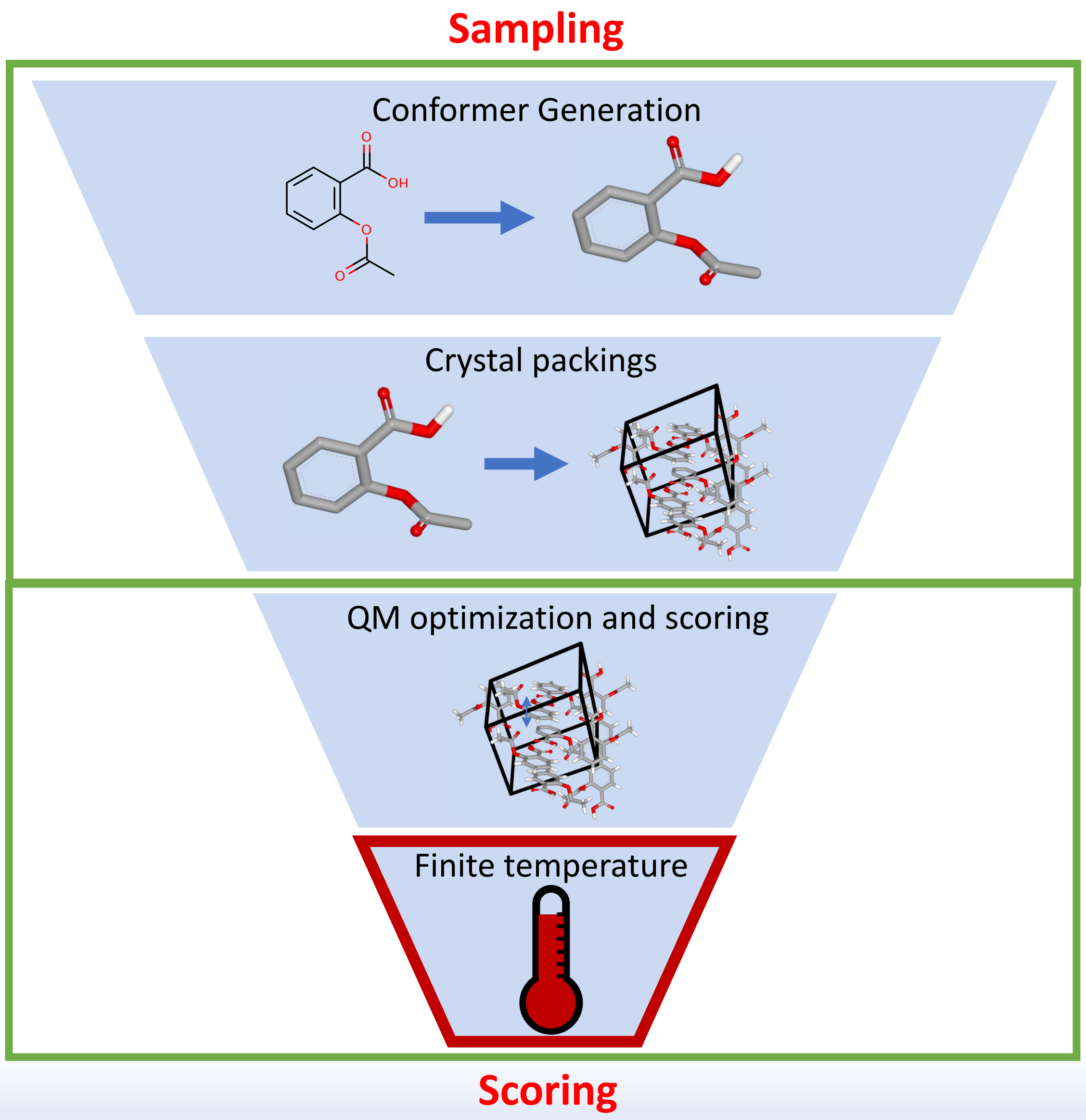
- **Pharmaceutical Formulation**
  - How to administer the drug? e.g. solid, injectable, etc.
  - Ensuring stability and delivery of drug

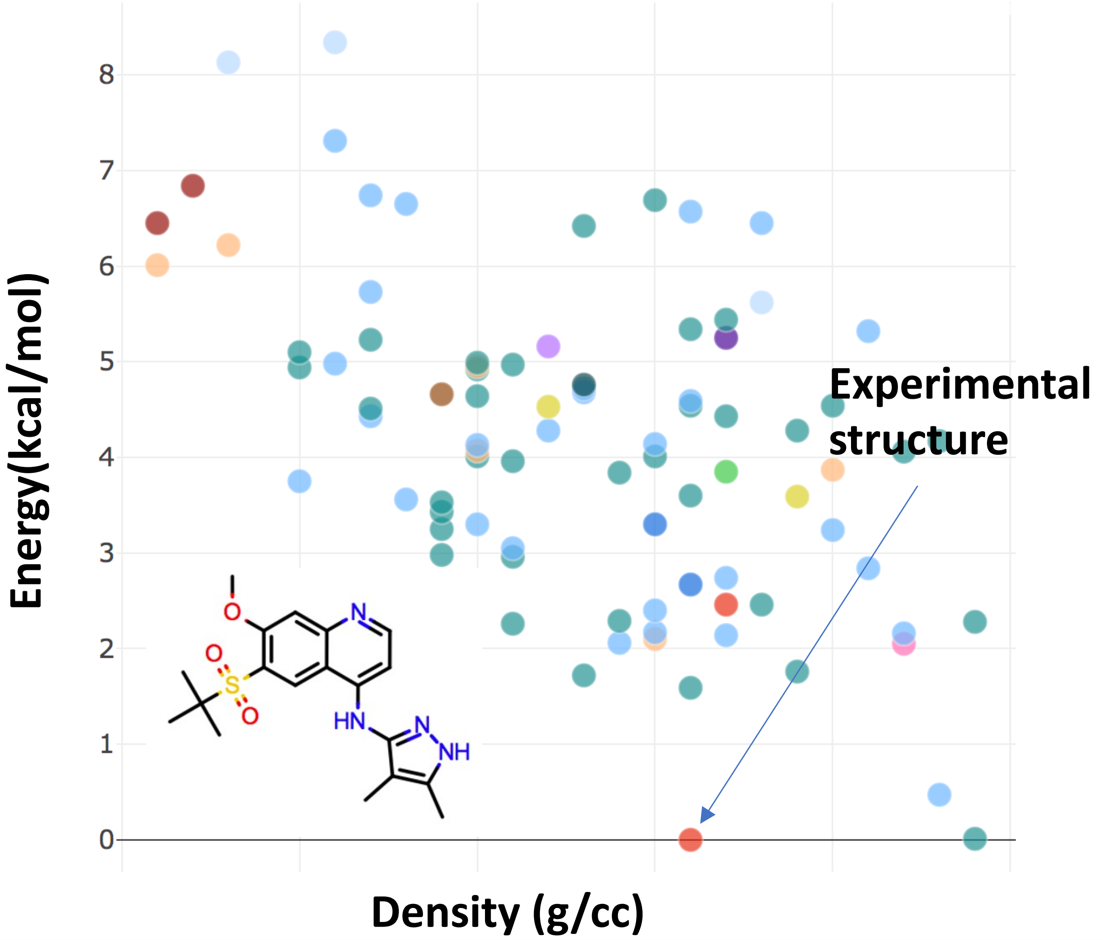




OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

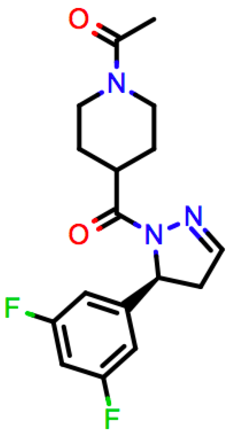
Crystal structure prediction workflow



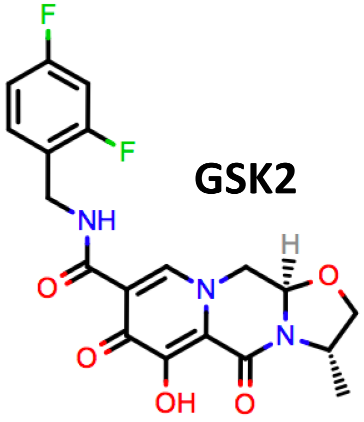


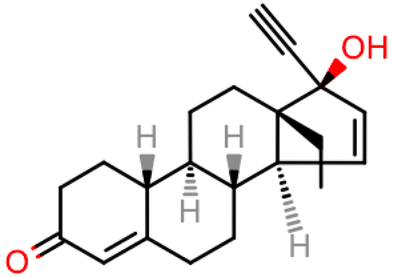
Z'=1 crystal structure predictions

GSK1

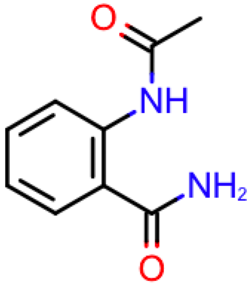


GSK2



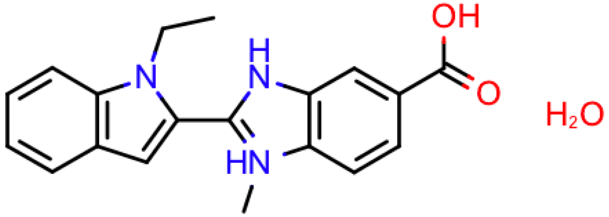


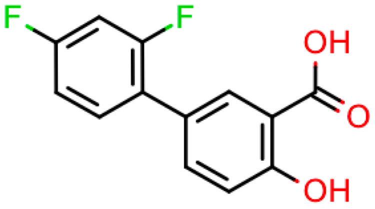
Gestodene



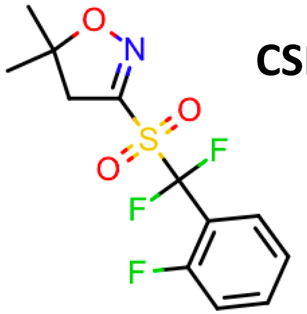
O-acetamidobenzamide

GSK6

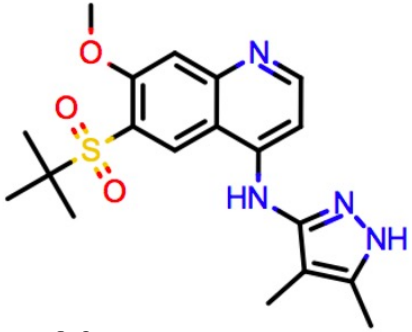




Diflunisal

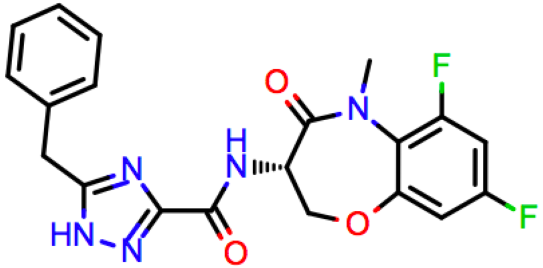


CSP7-XXXI



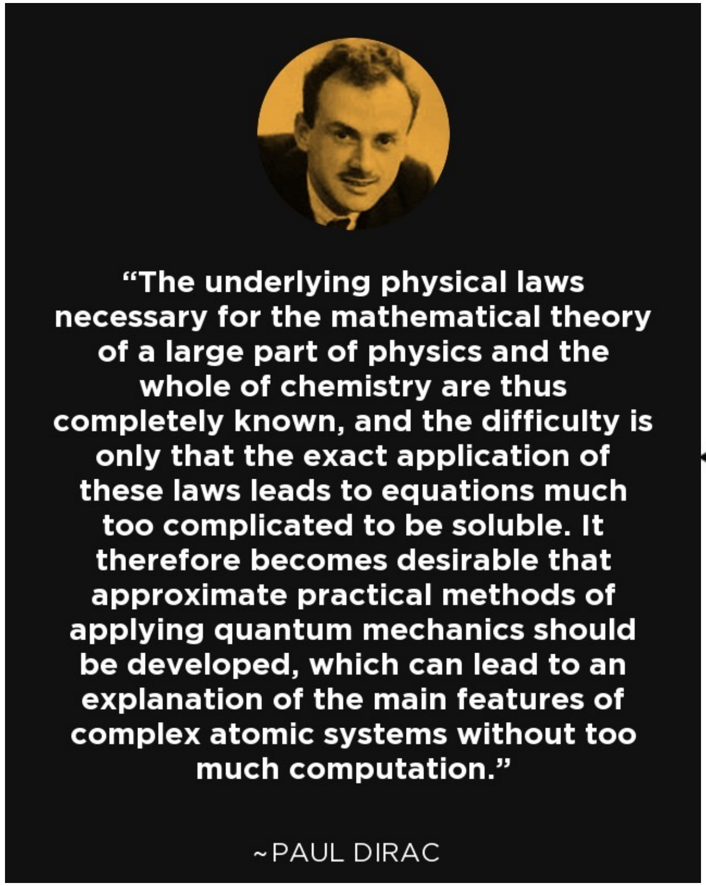
GSK4

GSK5

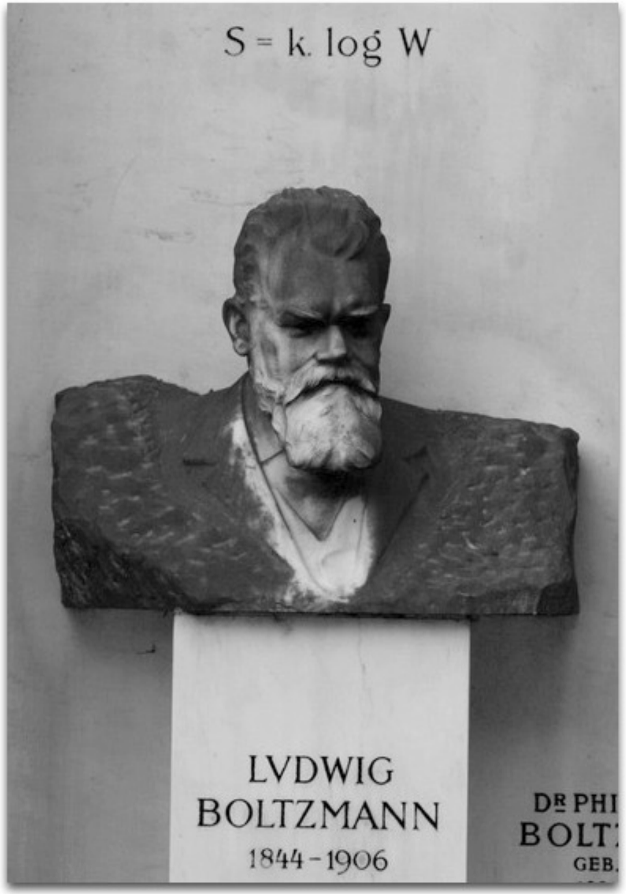


Two sides to physics-based molecular modeling

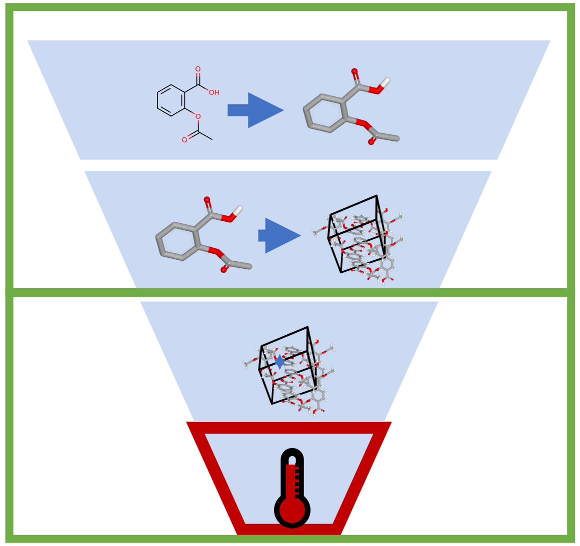
Energy Model



Sampling



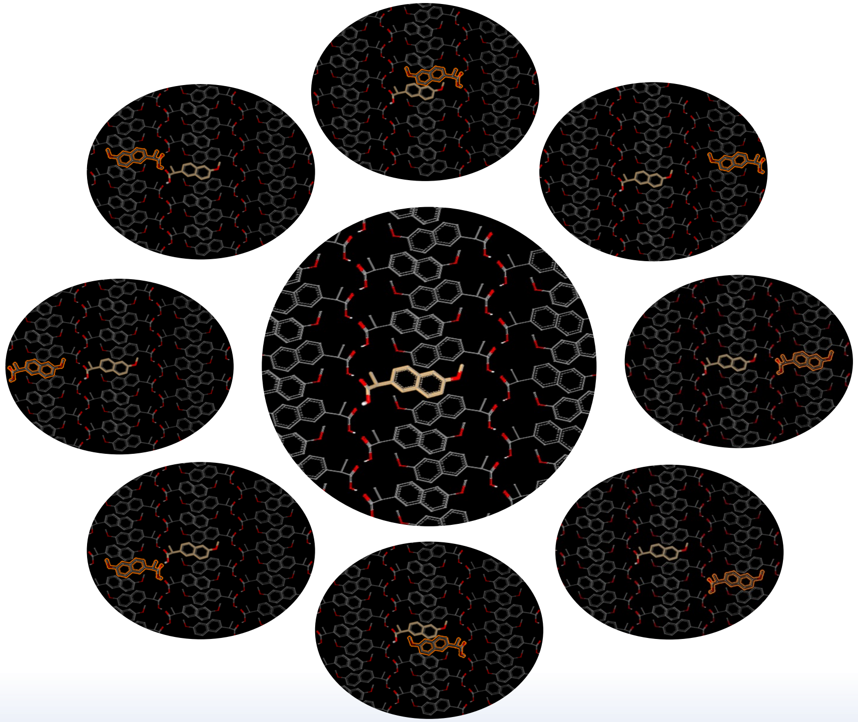
Sampling

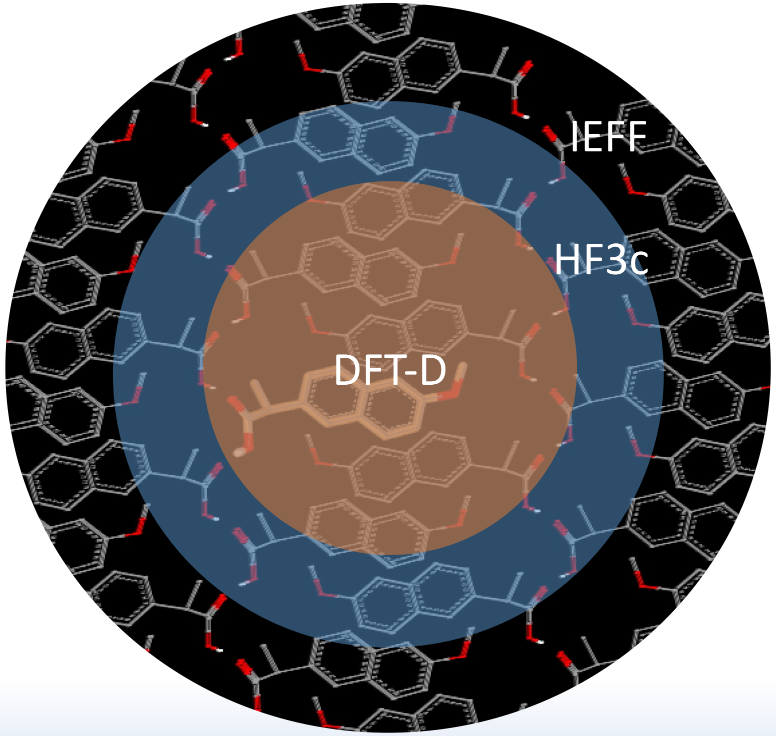


Energy model

Energy model

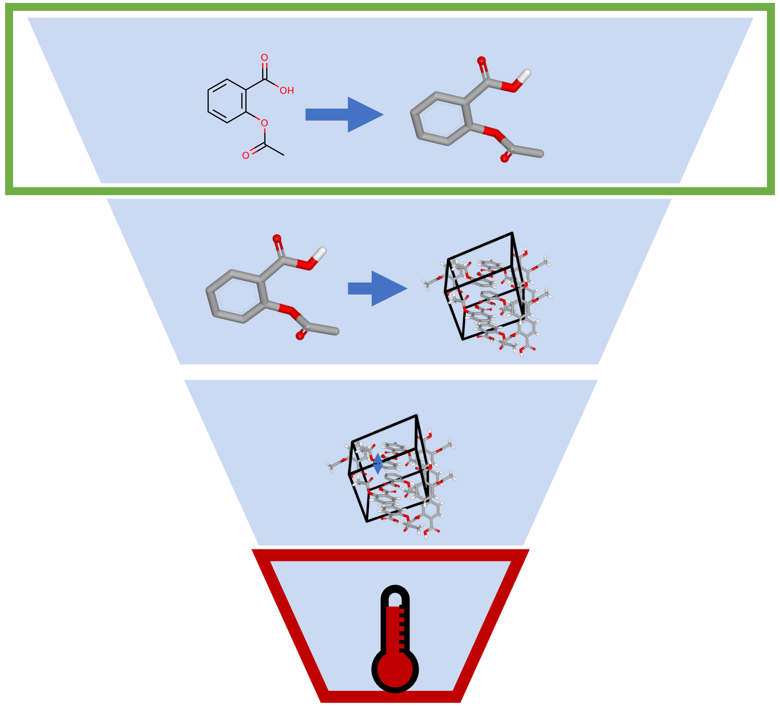
- Multipole force-field, IEFF
- Dimer expansion approach for optimizing and scoring crystals
    - Optimize 1000’s of crystal structures in a day
    - Compute Entropy of crystals at QM level within a day
    - Parallelization reaching 100K processors





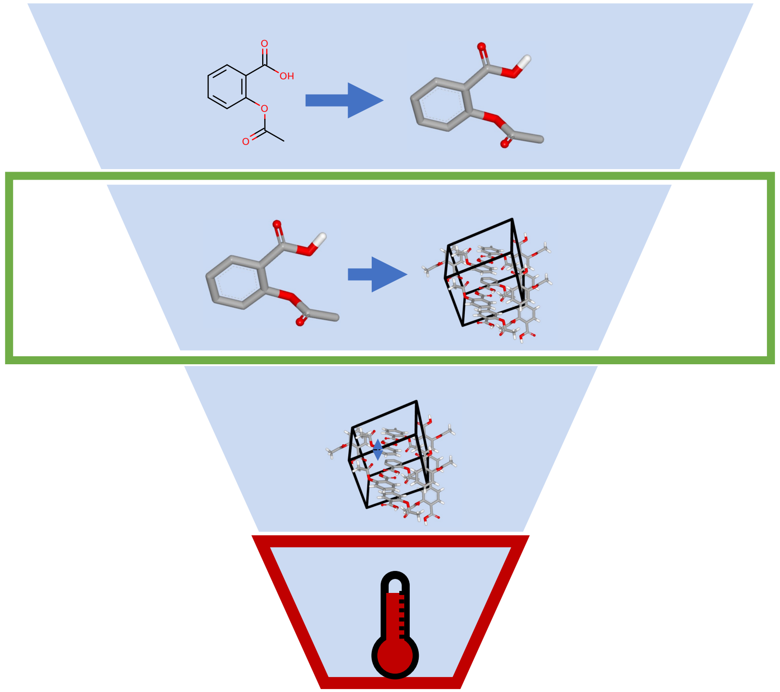
OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Conformer Sampling

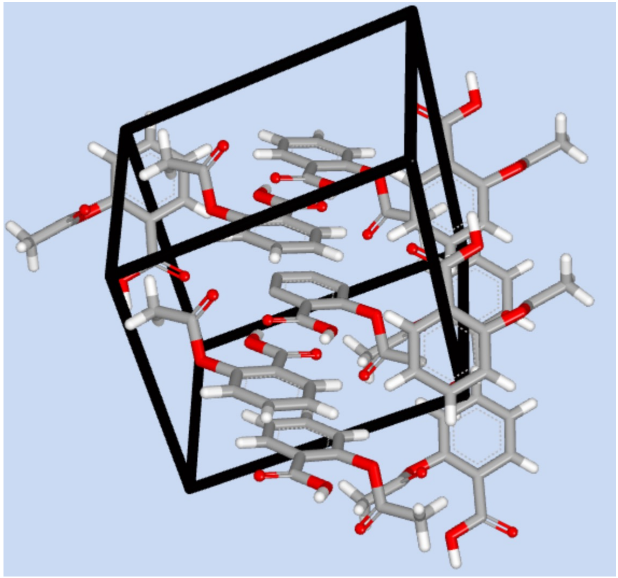


- Custom torsion rules for Omega conformer generation
  - Automated fragmentation
  - Torsion energy scanning and rule generation
- Multi-stage hierarchical sampling for highly flexible molecules
  - Identification of conformers that pack efficiently
  - Finer sampling of conformer space

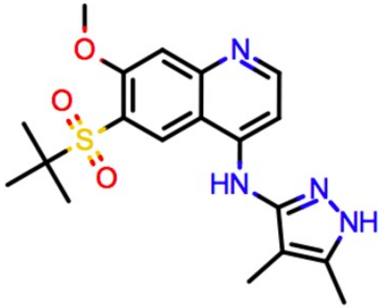
Packing sampling

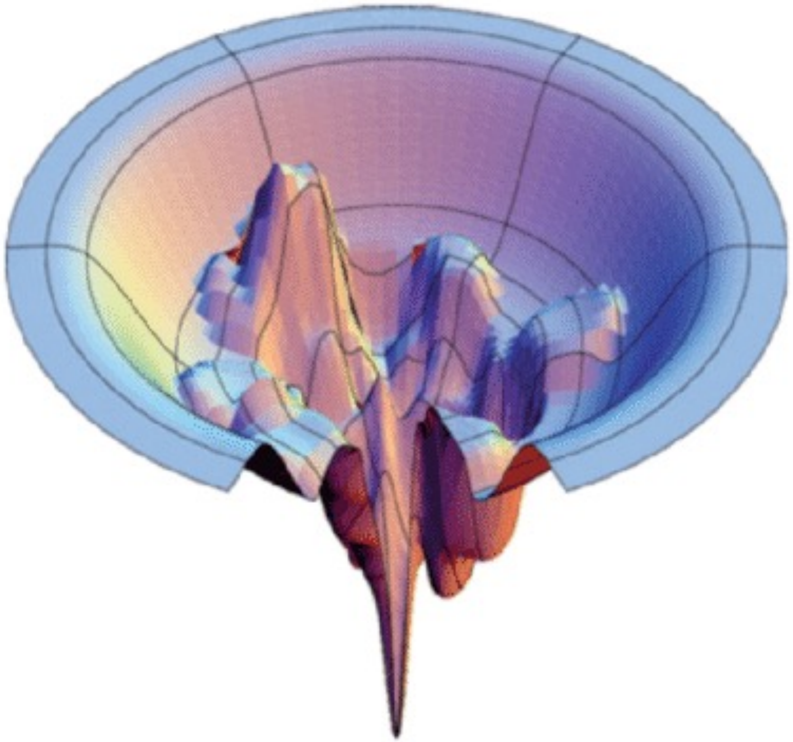


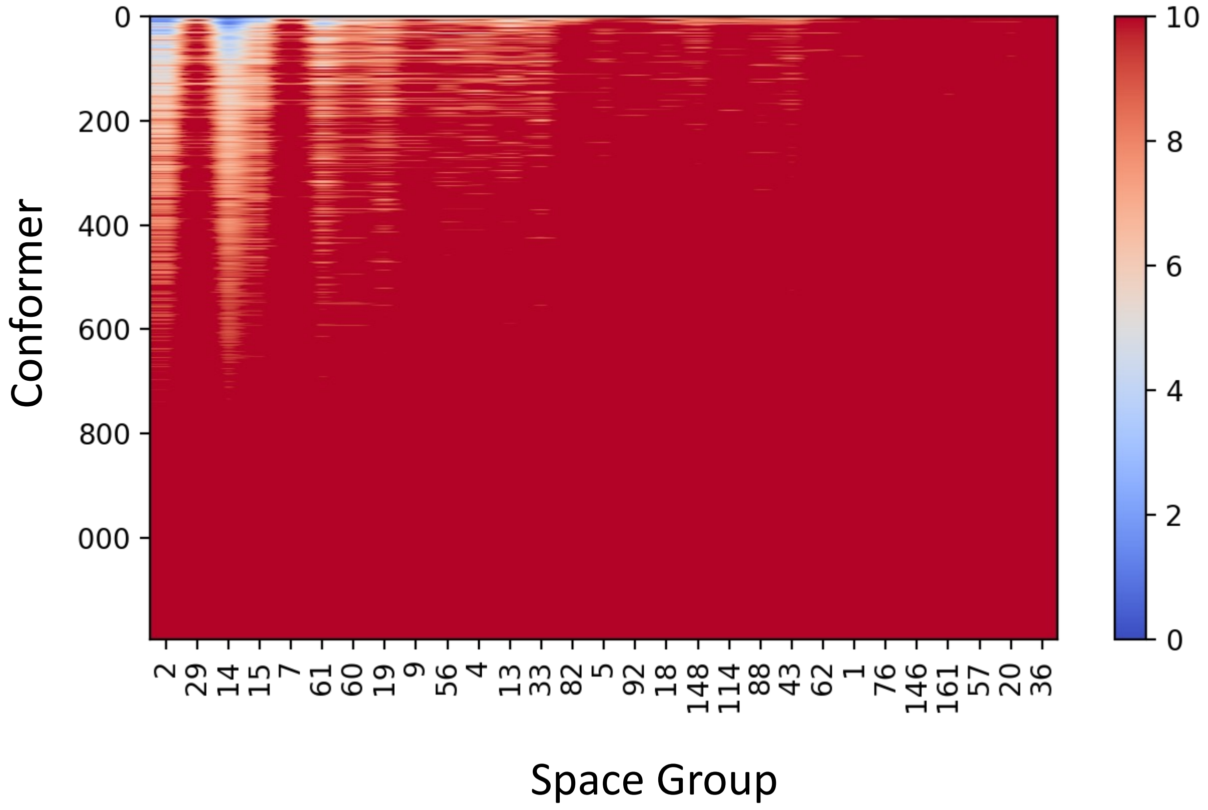
- Search parameters
  - Conformer
  - Space group
  - Orientation of the asymmetric unit
  - Unit cell dimensions



Crystal polymorph landscape

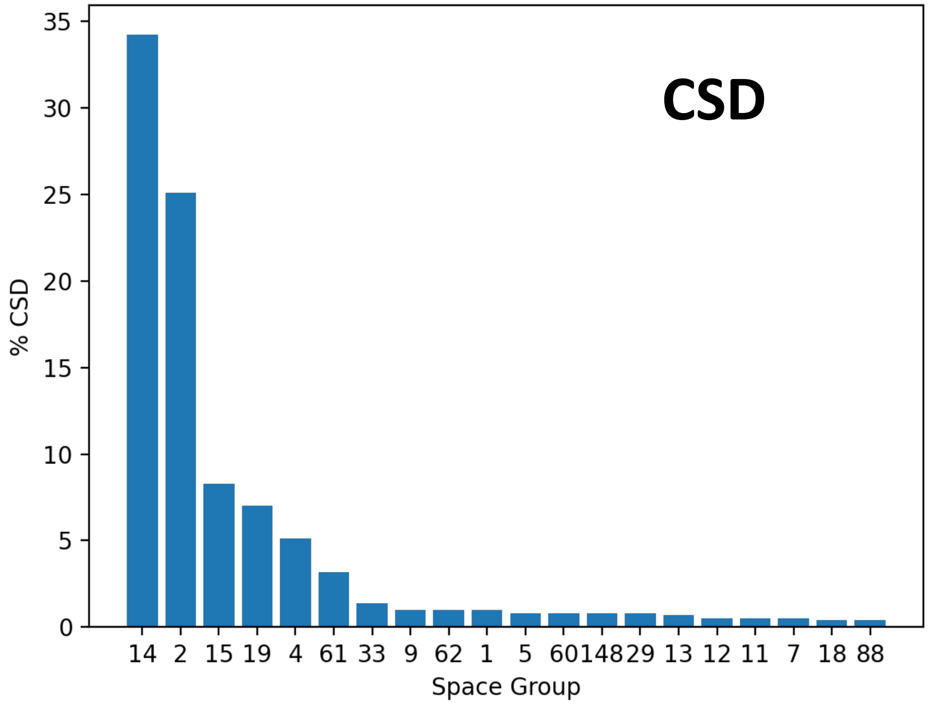


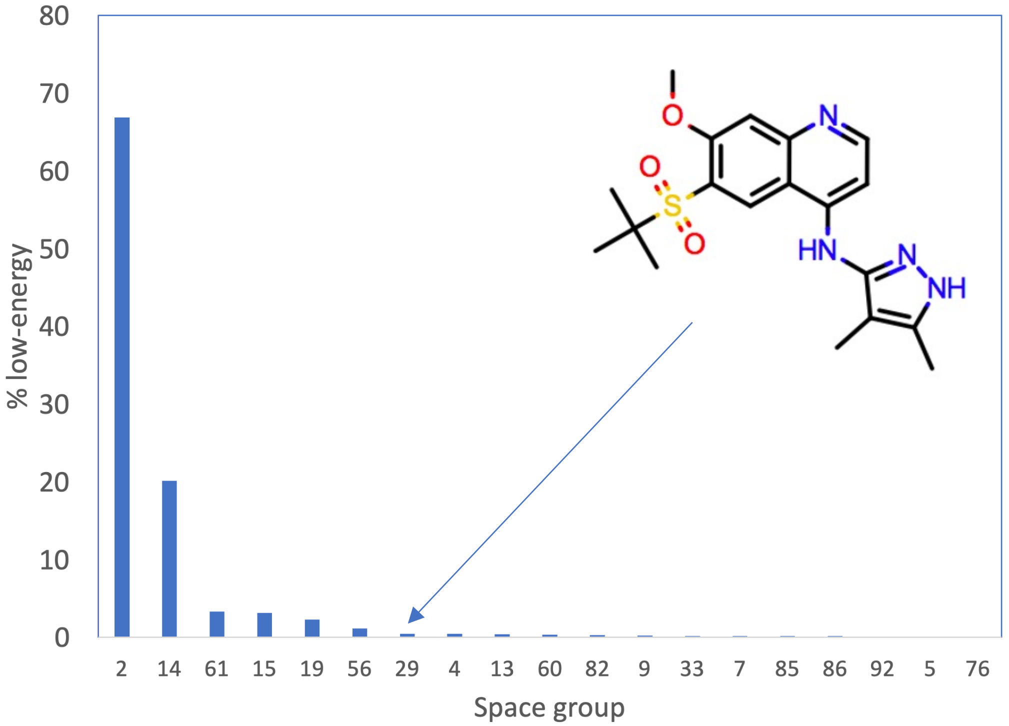




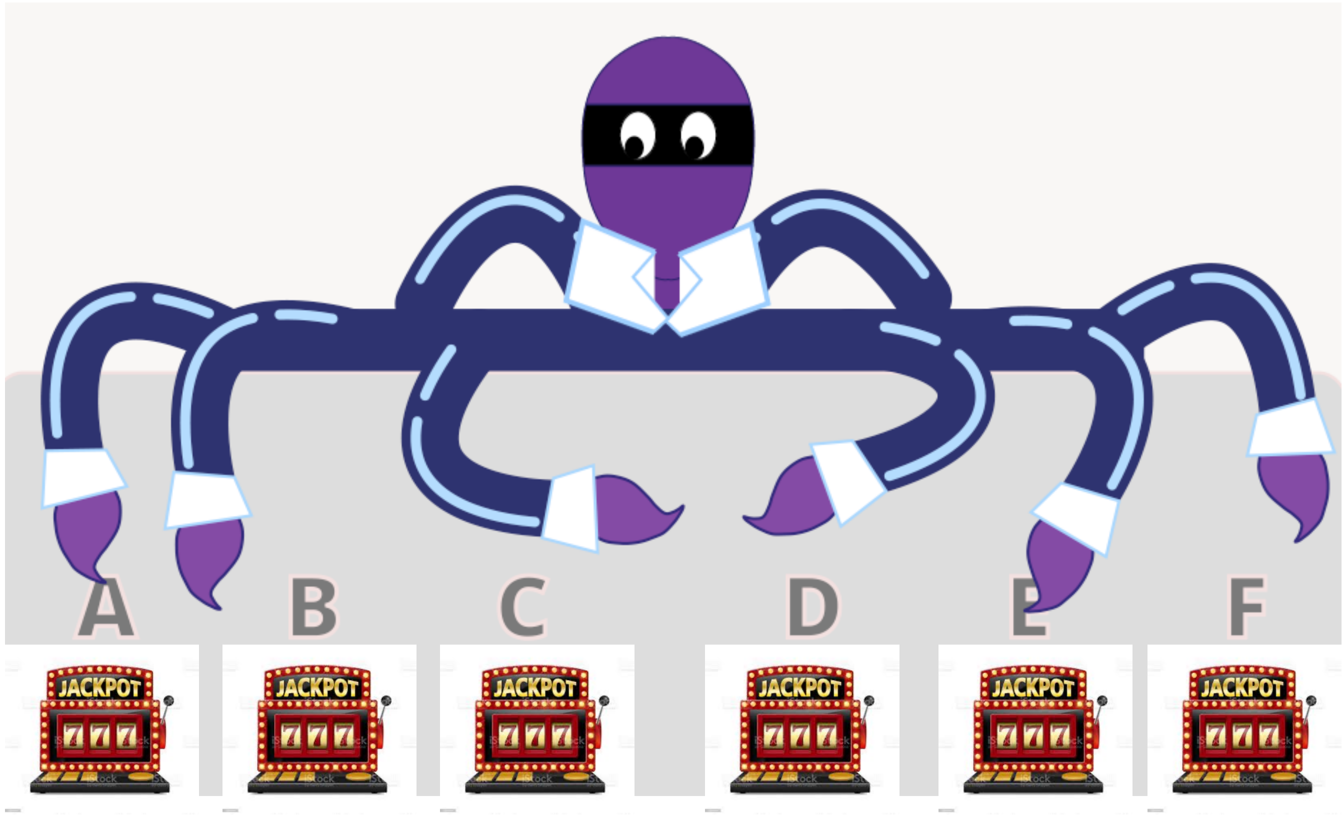
OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Space group distribution





Dynamic multi-armed bandit Problem



P_A P_B P_C P_D P_E P_F

P_X = Probability of getting a reward by sampling X
X = {conformer, sg}

- Reward
  - 1 – if the sample is a low-energy structure
  - 0 – if the sample is a high-energy structure
- No reward for finding the same low-energy structure again
  - Evolving probability
    - $P_x(t) < P_x(t\text{-}1)$
- Finite total reward for each arm
- Other applications
  - Ad marketing, e.g. Facebook Ads
  - Web design, e.g. Google optimize
  - Clinical trials

Sampling strategies

- Balanced (null model)
  - No prior knowledge or learning
  - Sample all combinations of conformers and space groups uniformly
- Reinforcement learning
  - Thompson (Bayesian) sampling
    - Probability of sampling an arm is proportional to the probability of the arm being optimal
    - Sample according to the posterior probability distribution
  - Dynamic probability matching
    - Probability of sampling an arm is proportional to the observed mean probability of reward
    - Most recent samples are considered for calculating the mean probability (dynamic)

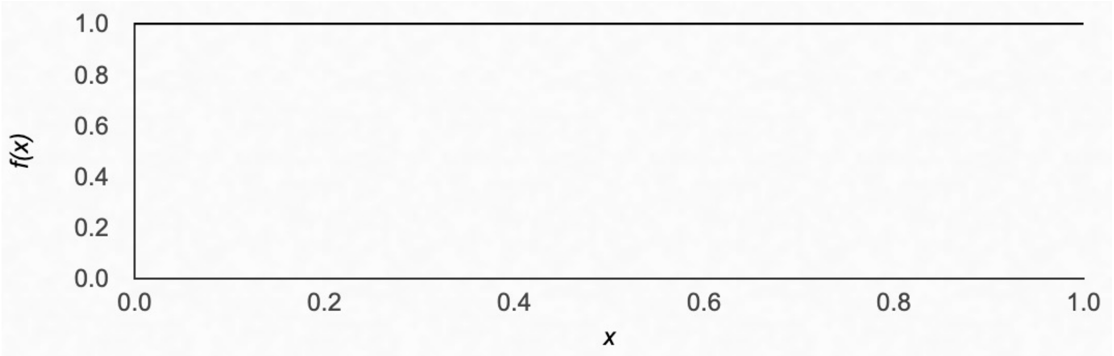
Thompson vs. Probability matching

- A/B Testing
  - A has a probability of 0.7 for reward
  - B has a probability of 0.3 for reward
- Thompson
  - Always pick “A”
  - Average reward = 0.70
- Probability matching
  - Pick A 70% of the time, and B 30% of the time
  - Average reward = 0.7*0.7 + 0.3*0.3 = 0.58

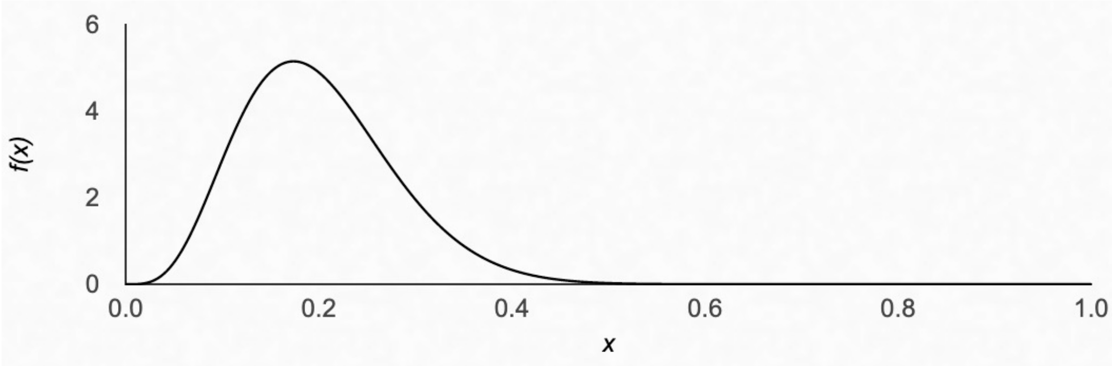
Thompson Sampling (Bayesian)

- Prior distribution
  - $P_x^1 = Beta(\alpha = 1, \beta = 1)$
  - Uniform distribution
- Posterior distribution
  - $P_x^n = Beta(\alpha^{n-1} + r^n, \beta^{n-1} + 1 - r^n)$
  - r^n = reward for n^{th} sample
- Arm selection
  - For each arm:
    - Generate a sample p_x from P_x^n
  - Select 'x' with max. p_x

$(\alpha = 1, \beta = 1)$



$(\alpha = 5, \beta = 20)$

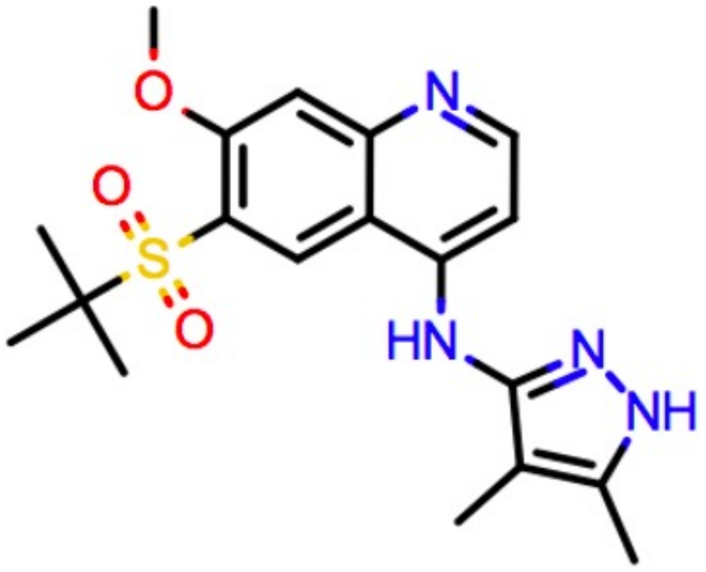


Dynamic probability matching

- Common decision-making strategy used by adult humans
  - Intuitive but not optimal
  - Also used by cockroaches, fish, and pigeons
  - Kids, non-human mammals, and (statisticians?) use the Bayesian strategy (Bayes)
- Estimate P_x^n as the mean probability of reward in the past
  - $P_x^n = \frac{\sum_{i=1}^{n} r^i}{n}$
- (Dynamic) Estimate based on recent history of outcomes
  - $P_x^{m,n} = \frac{\sum_{i=m}^{n} r^i}{(n-m)}$
- Sampling is proportional to the estimated probability
  - $n_x \propto P_x^{m,n}$

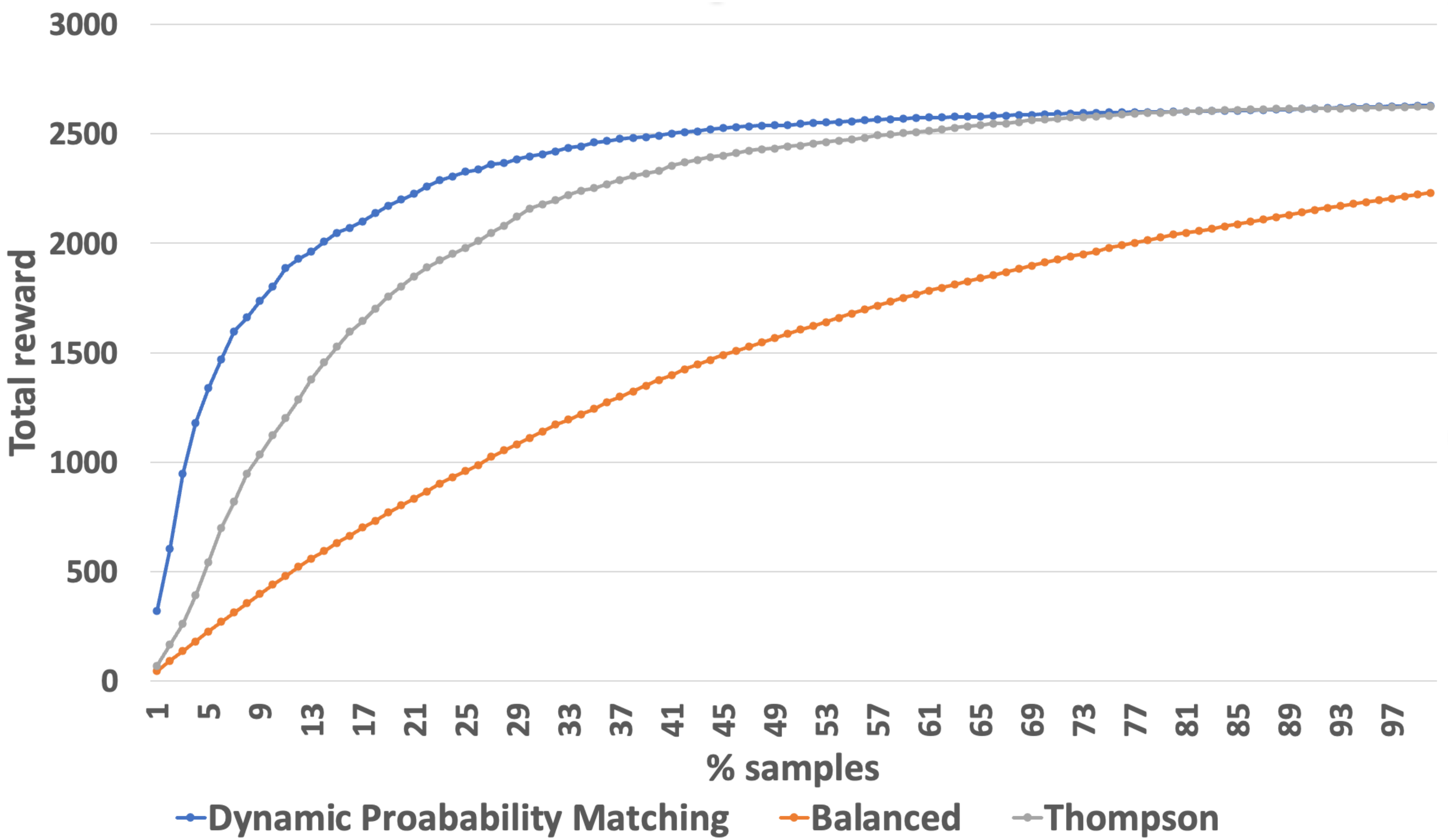
Saldana, C., Claidière, N., Fagot, J. *et al.*, *Sci Rep* **12**, 13092 (2022).

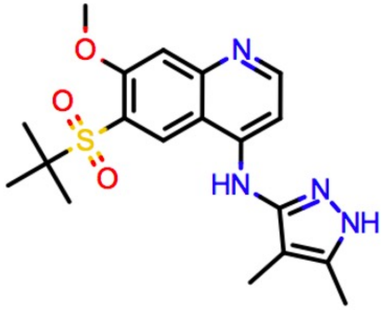
Proof-of-concept: Case study



- Number of arms
  - {conformer, sg} -> **38495**
  - 99.3% of the arms give no reward
- Total landscape
  - **~31 million** minima
  - Prebuilt using exhaustive sampling
- Total reward
  - # of low-energy structures = **2781**

Proof-of-concept: Case study





Standard error << 1% (too small to show)

Conclusions

- Thompson sampling outperforms balanced sampling
  - Thompson sampling is slow to adapt to evolving probabilities
  - Converges to optimal sampling in the long run
- Dynamic probability matching outperforms Thompson sampling
  - Higher early return-on-investment
  - Order of magnitude boost in sampling over balanced sampling
- Efficient sampling of conformer/sg space
  - Z’=2, co-crystals, and highly flexible drugs

Acknowledgements

- OpenEye
  - Caitlin Bannan
  - Varsha Jain
  - Grigory Ovanesyan
  - Tom Darden
  - Geoff Skillman
  - Anthony Nicholls
- GSK
  - Colin Edge
  - Luca Russo
  - Alan Graves (Nimbus)
  - Deborah Loughney
  - Eric Manas (Treeline)