UNIVERSITY OF LIVERPOOL

The Leverhulme Research Centre for Functional Materials Design

LEVERHULME TRUST

Geometric Modeling in the Prediction of Molecular Crystal Structures
The Close-Packing Principle Revisited

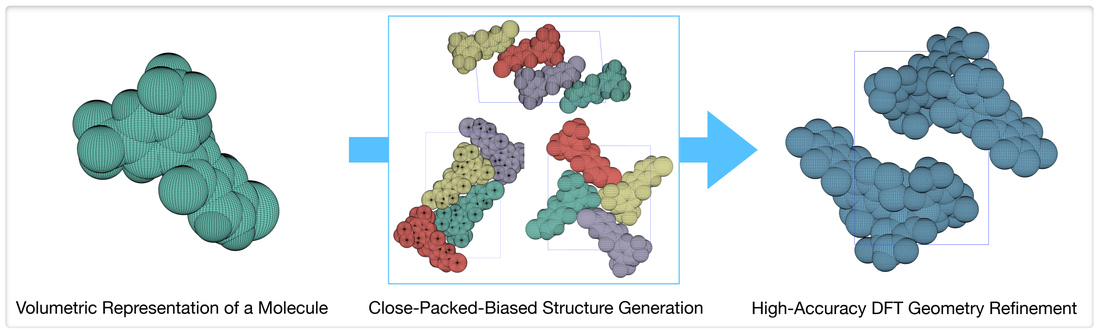
Miloslav Torda*, Xenophon Evangelopoulos* and Andrew I Cooper*
* Department of Chemistry, University of Liverpool, United Kingdom

milotorda.net

Towards Efficient Ground State Prediction of Molecular Crystals
Crystal Structure Prediction (CSP) is a critical tool in materials science, but its reliance on computationally intensive molecular quantum chemistry calculations makes it slow and difficult to scale. This bottleneck hinders rapid materials discovery, especially for autonomous labs.

Kitaigorodsky's Close-Packing Principle
"The mutual arrangement of the molecules in a crystal is always such that the 'projections' of one molecule fit into the 'hollows' of adjacent molecules."[1]

Fibre-Orbifold-Constrained Programming
We present a geometry-driven bias for crystal structure prediction that uses close-packing as a universal prior, sharply reducing search space and compute.



1. Fibered-Orbifold-Constrained Program

Find $\mathcal{K}_{\max} = \text{argmax}_{\mathcal{K}_G : G \in [G]_f} \rho\left(\mathcal{K}_G\right)$

over $[G]_f = \{H \mid \mathcal{F}(H) \cong \mathcal{F}(G)\}$

where $\mathcal{F}(G) = \mathbb{E}^3/G$ is a *Geometric Fibered Orbifold*

such that $\mathcal{K}_G = \bigcup_{g \in G} gK$, $\text{int}\left(g_i K\right) \cap \text{int}\left(g_j K\right) = \emptyset$, $\forall\ g_i, g_j \in G,\ g_i \neq g_j$

where $\rho\left(\mathcal{K}_G\right) = \dfrac{N\ \text{vol}(K)}{\text{vol}(U)}$ is the *Geometric Packing Density*.

- K - Closed Subset of $\mathbb{R}^3$
- G - Crystallographic Space Group
- U - Unit cell.
- N - Number of elements of factor group G/L where L is the lattice subgroup of G.

Entropic Trust Region Packing Algorithm[2]

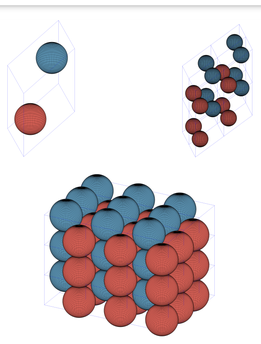
- Stochastic relaxation: $\tilde{\theta} = \text{argmax}_{\theta \in \Theta} J(\theta)$
  - $J(\theta) := E[F \mid \theta] = \int_{x \in \mathcal{X}} F(x) dP(\theta)$ - *Expected fitness*
    - F - *Fitness* of a packing configuration
    - $\mathcal{X}$ - Configuration space
    - $S = \left\{ dP(\theta) \mid \theta \in \Theta \subseteq \mathbb{R}^n \right\}$ - Statistical manifold with a dually flat Riemannian structure parametrized by θ.
    - $dP(\theta)$ - Extended Multivariate von Mises distributed probability measure
- Solution to the relaxed problem is a non-euclidean trust region search with the update equations

$$\theta^{t+1} = \theta^t + \Delta^t \frac{\widetilde{\nabla} J(\theta^t)}{\| \widetilde{\nabla} J(\theta^t) \|_{\mathcal{J}_{\theta^t}}}$$

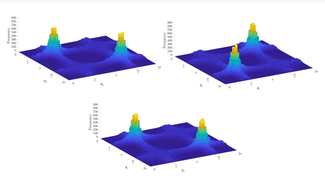
where $\widetilde{\nabla} J(\theta) = \mathcal{J}_\theta^{-1} \nabla_\theta J(\theta)$ is the *natural gradient* of the e*xpected fitness* $J(\theta)$.

- $\nabla_\theta J(\theta) = \mu_{F_{1-\frac{1}{q}}} - \mu$: Standard (Euclidean) gradient with respect to the Natural parametrization θ.

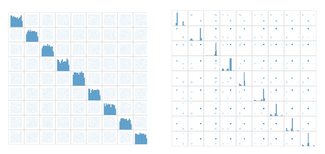
μ	Expectation parametrization of $dP(\theta)$ dual to θ.
$\mu_{F_{1-\frac{1}{q}}}$	Expectation parametrization of truncated probability distribution derived from the $(q-1)$-th q-quantile of the fitness F denoted $F^{\theta}_{1-\frac{1}{q}}$.
$\mathcal{J}_\theta$	Fisher metric tensor at point θ.
Δ^t	Step size.



A packing where G is of type $P\bar{1}$ and K is the unit sphere. (Top left) a single unit cell and (top right) eight unit cells. (Bottom) densest $P\bar{1}$ packing packing of the unit sphere with packing density of $\frac{\pi}{\sqrt{18}}$.



Histograms of projections along (Top) θ_1, (middle) θ_2 and (bottom) θ_3 coordinates of 100'000 samples generated from the trivariate Extended Multivariate von Mises Distribution.



1000 realizations of the Extended Multivariate von Mises distribution defined on an 9-dimensional torus from a single run of the Entropic Trust Region Packing Algorithm applied to the packings of the unit sphere in space group $P1$. (Left) Initial distribution and (right) output distribution.

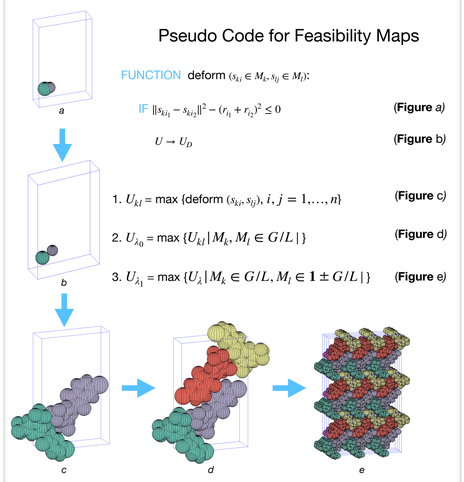
2. Close-Packings of Volumetric Models of Molecules

Feasibility Maps

- Non-overlap constrains are given by
  $\|s_{ki_1} - s_{ki_2}\|^2 - (r_{i_1} + r_{i_2})^2 \geq 0$
  where $s_{ki} = Ug_k c_k + g_k m_i,\ g_k \in G$ and $M_k = \left\{ (m_i, r_i), i = 1, \ldots, n \right\}$
  - c_k - Fractional coordinates of the centroid of molecule M_k
  - m_{ki} - Euclidean coordinates of i-th atom of molecule M
  - r_{ki} - van der Waals radius of i-th atom of molecule M
- We map every sampled configuration to a configuration without any constraint violation through the following deformation of the unit cell U:

$$U_D = UD$$
$$D = argmin\ vol(U_{D_l}) \mid l = \{1, \ldots, 7\}$$

$$D_1 = \begin{bmatrix} d & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & d \end{bmatrix}, D_2 = \begin{bmatrix} d & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & 1 \end{bmatrix}, D_3 = \begin{bmatrix} d & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix}, D_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & d \end{bmatrix}, D_5 = \begin{bmatrix} d & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, D_6 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & 1 \end{bmatrix}, D_7 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix}$$



- Additionally given a set of feasible configurations $S_1, S_2, \ldots, S_n$ where $vol(U^{S_i}) \leq vol(U^{S_j}) \mid i < j$, for each k-th decile $S_{\lceil k\frac{n}{10} \rceil}$ we define the following non-linear program

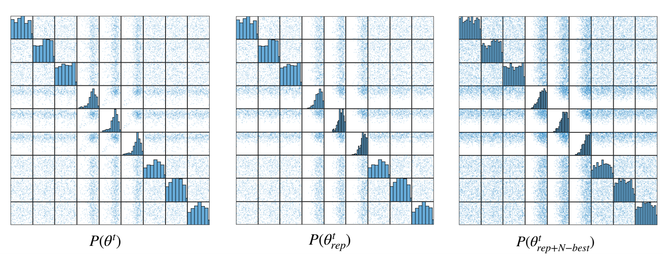
minimize $vol(U^{S_{\lceil k\frac{n}{10} \rceil}})$
subject to the non-overlap constraints

and solve the program via *Sequential-Quadratic Programming.*

Correction of the Search Gradient Respecting Feasibility Maps

- Repairing the unfeasible configurations slightly shifts the underlying probability distribution from $P(\theta^t)$ to $P(\theta^t_{\text{rep}})$
- *Entropic Trust Region Search Gradient* correction:
  1. Transfer $\widetilde{\nabla} J(\theta^t_{\text{rep}}) \in T_{\theta^t_{\text{rep}}} S$ to the tangent space $T_{\theta^t} S$ using the p*arallel transport map*
     $$\mathcal{P}_\gamma : T_{\theta^t_{\text{rep}}} S \to T_{\theta^t} S$$
     along the curve γ from θ^t_{rep} to θ^t. Since S is dually flat, the parallel transport $\mathcal{P}_\gamma$ is independent of the choice of γ.
  2. Adjust $\widetilde{\nabla} J(\theta^t_{\text{rep}}) \in T_{\theta^t}$ by $\mathcal{J}^{-1}_{\theta^t}(\mu^t_{\text{rep}} - \mu^t)$ to align with the Feasibility Map.
- Additionally, we introduce a *tournament selection* to increase competition within the packing population by merging the N-best configurations found in previous iterations with the repaired configurations from the current iteration.
- Final form of the search gradient respecting feasibility map transformations:

$$\widetilde{\nabla} J(\theta^t) = \mathcal{J}^{-1}_{\theta^t_{\text{rep + N-best}}}(\mu^t_{\text{rep + N-best}_{\rho_{1-\frac{1}{q^*}}}} - \mu^t_{\text{rep + N-best}}) + \mathcal{J}^{-1}_{\theta^t_{\text{rep}}}(\mu^t_{\text{rep + N-best}} - \mu^t_{\text{rep}}) + \mathcal{J}^{-1}_{\theta^t}(\mu^t_{\text{rep}} - \mu^t)$$

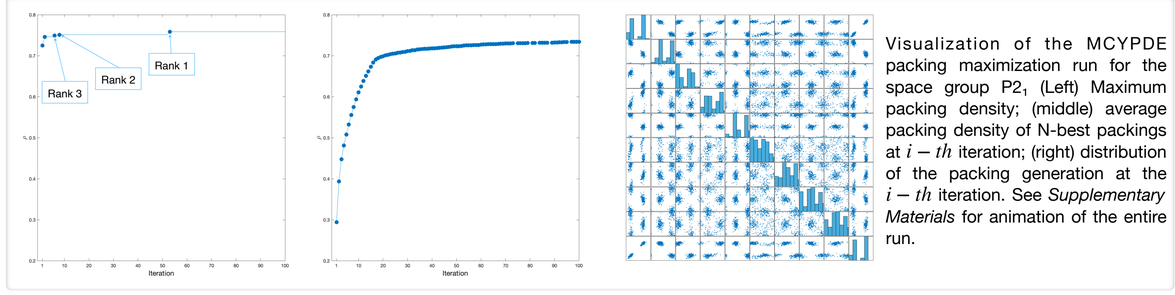


Visualization of the Extended Multivariate von Mises distribution defined on an 9-dimensional torus at iteration t and parametrized by θ^t_*. (Left) The original distribution at point θ^t; (middle) the shifted distribution at point θ^t_{rep}, obtained by feasibility maps from $P(\theta^t)$; (right) the distribution at point $\theta^t_{rep+N-best}$, resulting from merging N-best configurations from tournament selection for all $s \mid s < t$ with the repaired configurations from $P(\theta^t_{rep})$.

3.Towards Efficient Ground State Prediction of Molecular Crystals

- Selected four benchmarking structures from the '*Dataset: CSP-generated crystal structures of 1,000+ rigid organic molecules*'[3]
  - The global energy minimum structure is also Global molecular density maximum structure.
  - CSP computations were performed on ARCHER2, the UK National Supercomputing Service.
- Close-Packing searches were performed on a MacBook Pro with a M2 Max SoC.
  - Total Number of Cores: 12 (8 performance and 4 efficiency)
- Structural comparisons were conducted using CSD software's implementation of COMPACK[4].
- Structure ranking is given by *Geometric Packing Density*.
  - Rank i structure is the is the i-th densest packing such that there is no structural match between Rank i and Rank j for all $j = i-1, i-2, \ldots, 1$.

MCYPDE - Space Group $P2_1$



Visualization of the MCYPDE packing maximization run for the space group $P2_1$ (Left) Maximum packing density; (middle) average packing density of N-best packings at $i-th$ iteration; (right) distribution of the packing generation at the $i-th$ iteration. See *Supplementary Materials* for animation of the entire run.

- Rank 1 structure attained at ~ 7.5 min

Reference	Comparison	Molecules in Common	RMS	ρ
MCYPDE	Rank 1	30 out of 30	0.247	0.7584
MCYPDE	Rank 2	8 out of 30	1.790	0.7514
MCYPDE	Rank 3	10 out of 30	0.756	0.7510

HOHLAR - Space Group $P\bar{1}$

- Rank 3 structure attained at ~ 30 min

Reference	Comparison	Molecules in Common	RMS	ρ
HOHLAR	Rank 1	9 out of 30	0.430	0.7468
HOHLAR	Rank 2	3 out of 30	0.148	0.7432
HOHLAR	Rank 3	30 out of 30	0.802	0.7412

BDEHAN - Space Group $P2_1/c$

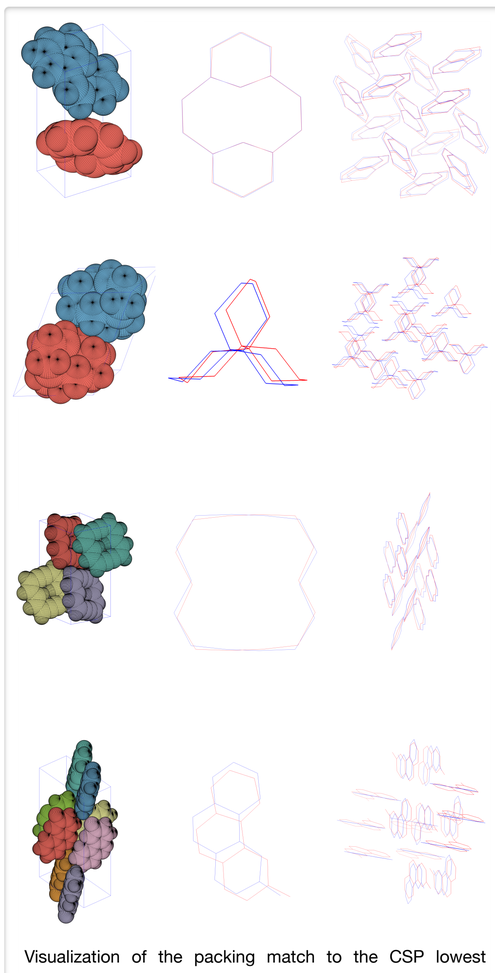
- Rank 2 structure attained at ~ 1.25 min

Reference	Comparison	Molecules in Common	RMS	ρ
BDEHAN	Rank 1	1 out of 30	0.000	0.7604
BDEHAN	Rank 2	30 out of 30	0.641	0.7430
BDEHAN	Rank 3	2 out of 30	0.180	0.7406

SIZZUW - Space Group C2/c

- Rank 1 structure attained at ~ 15 min

Reference	Comparison	Molecules in Common	RMS	ρ
SIZZUW	Rank 1	30 out of 30	0.665	0.7670
SIZZUW	Rank 2	5 out of 30	1.380	0.7472
SIZZUW	Rank 3	6 out of 30	2.211	0.7451



Visualization of the packing match to the CSP lowest energy structure (Left) A single unit cell displayed in a space-filling representation, where colors indicate symmetry operations modulo lattice translations. (Middle and right) Overlay in a wireframe representation: the global energy minimum (blue) and the matching packing (red). (Middle) Asymmetric unit of the configuration; (right) a 15-molecule cluster.

References

1. A. I. Kitaigorodsky, Molecular Crystals and Molecules, Academic Press New York, 1973.
2. M. Torda, J. Y. Goulermas, R. Púček and V. Kurlin, *Entropic Trust Region for Densest Crystallographic Symmetry Group Packings*, SIAM Journal on Scientific Computing, 2023.
3. C. Taylor, P. Butler and G. M. Day, Predictive Crystallography at Scale: Mapping, Validating, and Learning from 1,000 Crystal Energy Landscapes, Faraday Discussions, 2024.
4. J. A. Chisholm and S. Motherwell, *COMPACK: A Program for Id*entifying Crystal structure similarity using distances, Journal of Applied Crystallography, 2004.

Supplementary Materials

https://milotorda.net/lrc-symposium-supplementary-materials/

The Cambridge Crystallographic Data Centre · diamond · Hartree Centre Science & Technology Facilities Council · ICNS International Center for Nano-Systems · Imperial College London · King Abdullah University of Science and Technology · University of Southampton

The authors thank the Leverhulme Trust for funding this research via the Leverhulme Research Centre for Functional Materials Design