An Open Science Grid Implementation of the Steady State Genetic Algorithm for Crystal Structure Prediction

K N Varela [1,2], G I Pagola [1,2], M B Ferraro[1,2], A M Lund [3,4], A M Orendt [3], J C Facelli [4]

[1]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física. Buenos Aires, Argentina.

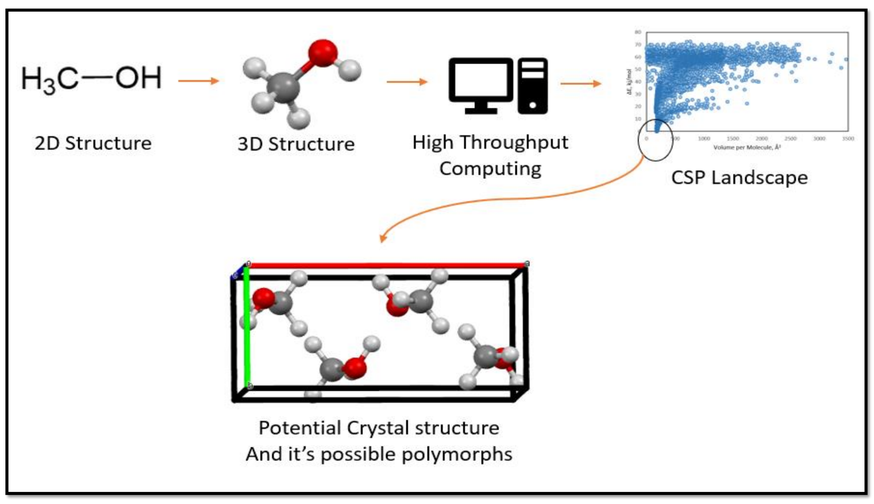
[2] CONICET - Universidad de Buenos Aires, Instituto de Física de Buenos Aires (IFIBA). Buenos Aires, Argentina.

[3]Center for High Performance Computing, The University of Utah. Salt Lake City.

[4] Department of Biomedical Informatics, The University of Utah. Salt Lake City, Utah. USA Utah, USA

INTRODUCTION

Cristal Structure Prediction (CSP) based only of the chemical diagram



CSP Importance

- Development of new compounds and materials
- Crystal and co-crystal design
- Materials science
- Drug design
- Solid state chemistry

CSP Challenges

- Computational complexity (High computational cost)
- Vast configurational space (Large number of possible configurations)
- There is a need for new and improved algorithms that can leverage emerging high-performance computer architectures to enable the use of the most accurate energy models.

METHODOLOGY

MGAC-QE-OSG[1]

MGAC[2]: Modified Genetic Algorithm for Crystals - "Steady State" GA (SSGA)

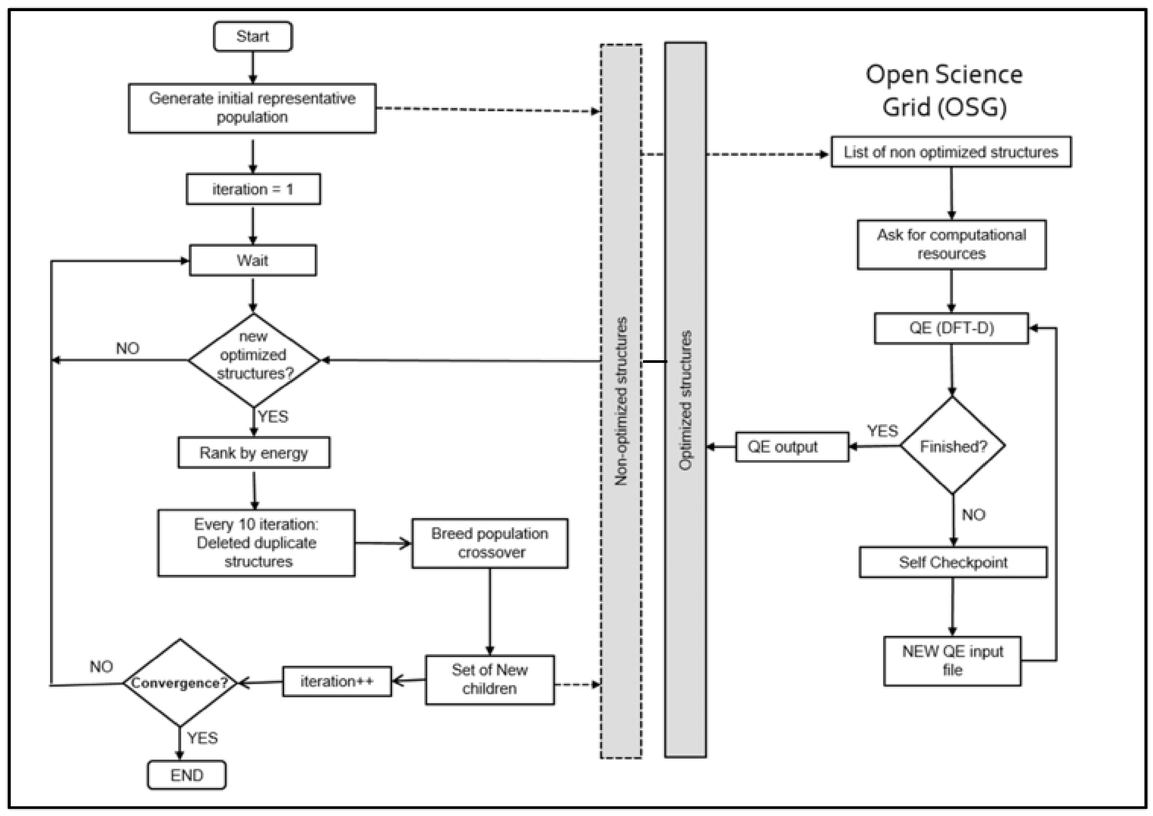
- No generations
- General definition of the GA genome that eliminates the need of searching individually for each of the 230 possible space groups
- Selection criterion for the best individuals: energy of the candidate crystal

Quantum Espresso: www.quantum-espresso.org/

- DFT-D optimization

Open Science Grid (OSG): www.osg-htc.org/

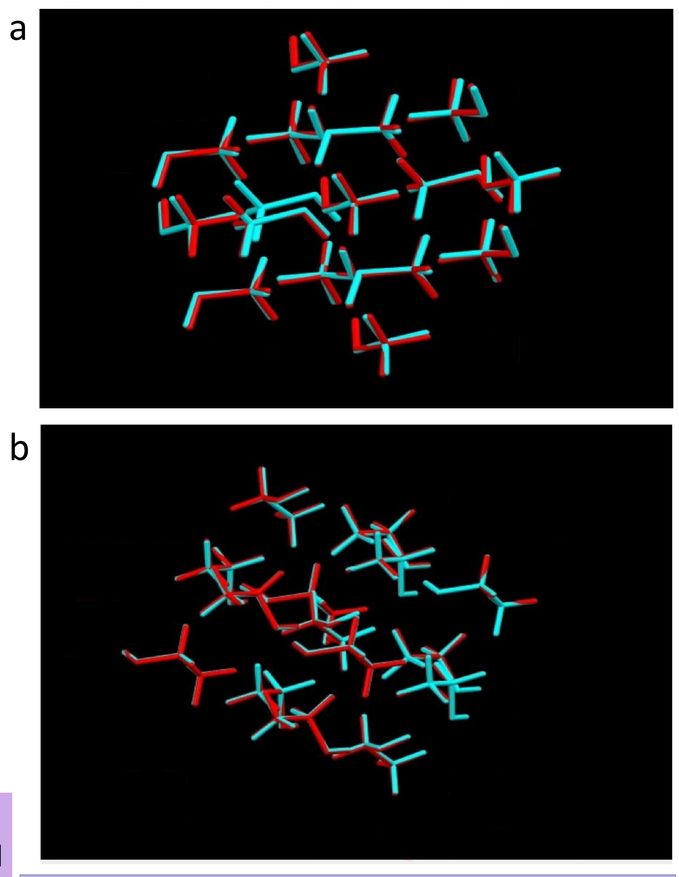
- Asynchronous distribution of DFT(QE) calculations.
- High performance and efficency

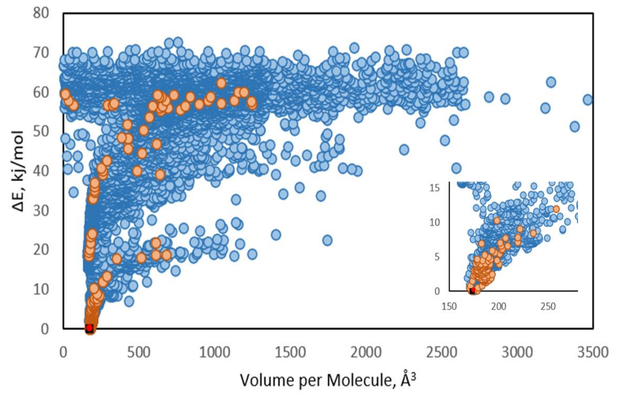


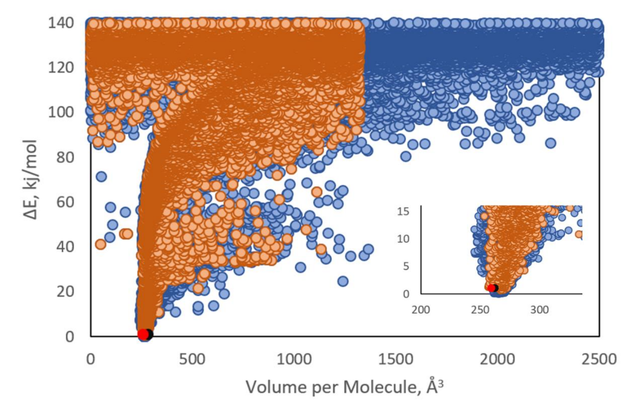
MGAC-QE-OSG flow chart

- Two lists of individuals, one for optimized structures and another for non-optimized structures.
- On the left (virtual machine provided by the Center of High-Performance Computing (CHPC) at the University of Utah): creation and crossover of individuals, elimination of duplicate structures, and verification of convergence.
- On the right (OSG): evaluation and self-checkpointing of structures in the optimization process
- Both workflows continuously integrate new solutions into the lists of individuals, ensuring a constant workflow to keep the optimization queues full.

RESULTS AND DISCUTIONS





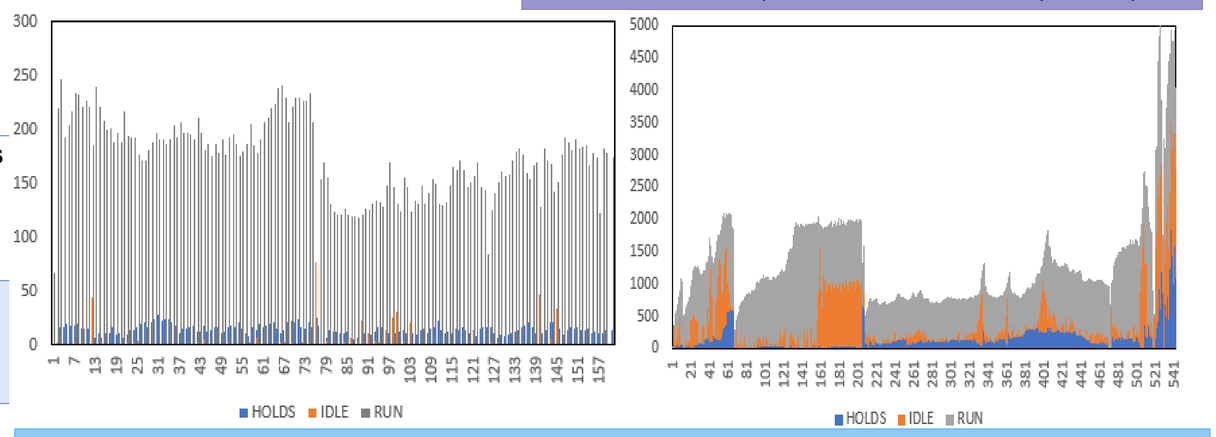


Tested molecules a) Methanol y b) Ethanol: Comparison of the lowest energy structure found by the MGAC-QE-OSG (in red) with the experimental structures[3] [4]. And the energy distribution of MGAC-QE-OSG crystalline structures as a function of their volume.
In both graph, the convergence of the MGAC to a global energy minimum is observed.

Methanol: structures were generated for 32 space groups with 4 or fewer symmetry operations represented in blue and the crystals structures corresponding to the interested group (P_{212121}) are in orange.
Ethanol: structures were generated for those space groups with 1 or 2 symmetry operations (blue marks), the orange marks corresponding to the group P_{1C1}.
In both cases. Black and red dots are MGAC-QE-OSG best structure and the experimental structure respectively

Table 1. Summary of computational data

Test Molecule	No. of optimized structures	Number of MGAC iterations	Structures per iteration	Total Core Hours	The average of core hours per structure	OSG Resourses
Methanol	8567	160	~70	27K	6.48	CPU: 1 per job
Ethanol	109561	558	~300	1.5M	13.48	RAM: 5-20 GB
						Disk: 1GB



Number of processes running, on hold, and idle during each MGAC-QE-OSG iteration of the optimization on left: methanol , and right ethanol

Table 2. Energy and crystallographic parameters for predicted molecules for the MGAC-QE-OSG and its comparison with the reference experimental structure

	Space group	Energy	Cell parameters a	b	c	α	β	γ	RMS	pxrd
					METHANOL					
MGAC-QE-OSG		-63075.01	4.3938	4.8473	9.0826	90	90	90		
Experimental	$P2_12_12_1$		4.6469	4.9285	9.0403	90	90	90	0.1432	0.8721
Exp-QEopt		-63074.68	4.3707	4.9176	9.0898	90	90	90	0.0533	0.9915
					ETHANOL					
MGAC-QE-OSG		-162334.81	5.3909	6.8493	8.0829	90	102.3630	90		
Experimental	P_1c_1		5.377	6.882	8.255	90	102.2000	90	0.0749	0.9806
Exp-QEopt		-162334.78	5.3649	6.8954	8.0416	90	102.5540	90	0.0359	0.9969

CONCLUSIONS

- The MGAC-QE-OSG has been updated to fully utilize the extensive resources provided by the OSG, which are crucial for enabling the use of DFT-D in CSP.
- The OSG platform allowed the elimination of the conventional generation concept within the classical genetic algorithm, enabling independent and asynchronous evaluation of individual crystal structures.
- Implementing procedures for managing computationally intensive jobs is essential, such as Self-Checkpoint applications, which allow periodic saving of information in case the work needs to be resumed on a different machine or at a later time.

REFERENCES

[1] K.N. Varela, G. I. Pagola, A. M. Lund, M. B. Ferraro, A. M. Orendt and J. C. Facelli. "An open science grid implementation of the steady state genetic algorithm for crystal structure prediction". Journal of computational science, vol 82, 2024.
[2] A. M. Lund, G. I. Pagola, A. M. Orendt and M. B. Ferraro and J. C. Facelli. "Crystal structure prediction from first principles: the crystal structures of glycine," Chemical Physics Letters, pp. 20-24, 2015.
[3] M. Kirchner, D. Das and R. Boese, "Cocrystallization with Acetylene: Molecular Complex with Methanol," Crystal Growth & Design, pp. 763-765, 2008 vol 8 No3
[4] P.-G. Jönsson, "Hydrogen bond studies. CXIII. The crystal structure of ethanol at 87 K," Acta Crystallogr., pp. 232-235, 1976.