Implementation of a methodology for crystal structure prediction using genetic algorithms integrated into the Python ASE library

Bohdan Y. SEMENIUK [1], *Oleh D. FEIA* [2 3]

[1)] *Kyiv Academic University, 36 Vernadsky Blvd., 03142 Kyiv, Ukraine*
E-mail: semenuk.b.20@gmail.com
ORCID: https://orcid.org/0009-0007-0294-5859

[2)] *Kyiv Academic University, 36 Vernadsky Blvd., 03142 Kyiv, Ukraine*
E-mail: o.feia@kau.edu.ua
ORCID: https://orcid.org/0000-0002-6749-950X

[3)] *G.V. Kurdyumov Institute for Metal Physics of NAS of Ukraine, 36 Academician Vernadsky Blvd., 03142 Kyiv, Ukraine*

Received June 6, 2025, in the final form June 25, 2025
https://doi.org/10.3842/confKAU.2025.phys.01

Abstract. This work is dedicated to the development and implementation of a methodology for crystal structure prediction using genetic algorithms integrated into the Python ASE library. Crystal structure prediction plays a critical role in materials science, chemistry, and nanotechnology, enabling the discovery of novel compounds with tailored properties. By combining the flexibility of ASE with the speed of classical relaxers and the accuracy of DFT-based methods, our approach significantly reduces computational costs while maintaining predictive reliability. The methodology was validated on polymorphs of silica (SiO_2), where our system successfully recovered both global and local minima of the energy landscape. We also explore the integration of neural network relaxers such as MACE and AIMNet2 to further accelerate the search process. This study lays the groundwork for efficient, scalable, and accurate predictive modeling of crystalline materials.

Keywords: crystal structure prediction; genetic algorithm; energy landscape; crystalline silica polymorphs; structure relaxation; ASE; GULP

1 Introduction: The general concept of materials design

Humanity is in constant need of new technologies, which in turn necessitates the development of novel materials to enable their implementation. However, this progress is often hindered by traditional iterative approaches to material discovery, which typically require substantial financial investment, time, and researcher endurance. The advent of personal computing has enabled the automation of this process by integrating new components, such as computer simulations and structure prediction [1, 2], into the conventional materials research workflow. As a result, the current approach to the synthesis of new materials can be represented as a cyclical and iterative process (Figure 1).

One of the most critical and still insufficiently developed elements of this cycle is the technology for structure prediction and synthesis. The key challenges in this area include not only the speed and accuracy of prediction algorithms, but also the practical parameters of the predicted results. Ideally, these outcomes should enable straightforward experimental reproduction of the predicted compounds in a laboratory setting [2].

Our current research focus lies in the field of materials prediction, particularly in improving the speed and precision of this fascinating and promising process.

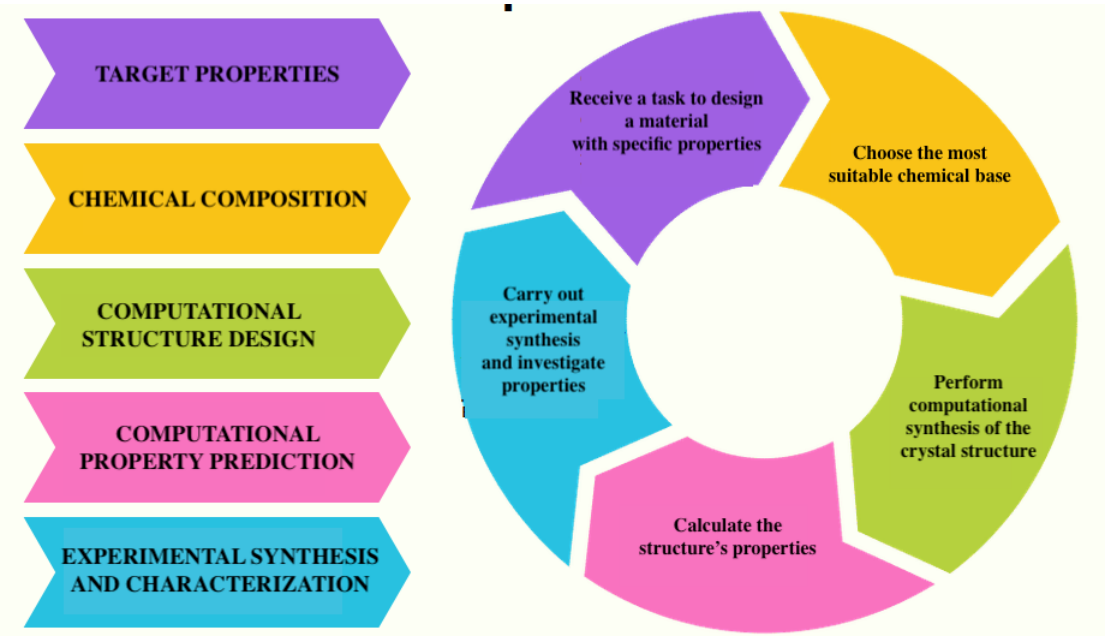

Figure 1. Diagram of the modern materials development process.

2 The problem of computational crystal structure synthesis

As with any scientific problem, it is essential to first establish a foundation of abstract principles. In the case of computational crystal structure synthesis, these principles include the concept of a multidimensional energy landscape (Figure 2) and the presence of global and local minima in that landscape [3].

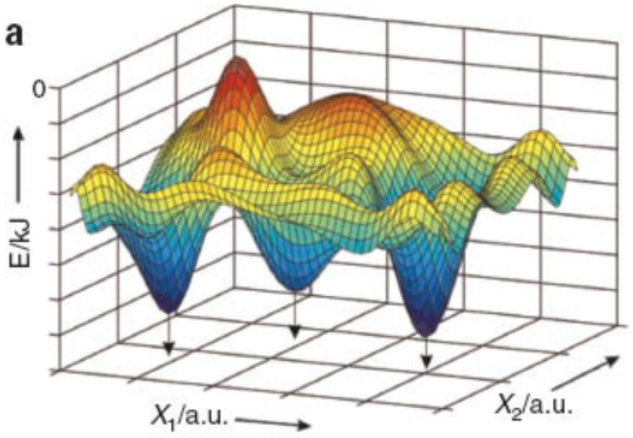

Figure 2. Multidimensional energy landscapes [3].

The energy landscape can be envisioned as an infinite set of points that form a continuous surface, where each point corresponds to a specific energy value and thus represents a distinct atomic configuration. The vertical axis denotes energy, while the remaining axes represent structural parameters subject to variation. In our case, these parameters include changes in lattice vectors and the atomic coordinates within the cell.

In the context of structure prediction, the primary objective is to locate the energy minima, both global and local. The global minimum corresponds to the most stable structure, which is

most likely to be experimentally synthesized. In contrast, local minima represent metastable structures that may be realizable under specific experimental conditions.

To identify these minima, a variety of search algorithms are employed, such as basin hopping [4], minima hopping [5], metadynamics [6]. In our work, we use a genetic algorithm (Figure 3) [7], which enables efficient exploration and analysis of large portions of the energy landscape with reasonable precision. The core principles of this algorithm will be described in detail later, but for now, it is important to note that its successful implementation requires the use of computational "calculators". These calculators must be capable not only of evaluating the total energy of a given structure, but also of performing structural relaxation, that is, optimizing the atomic configuration to find the nearest local energy minimum.

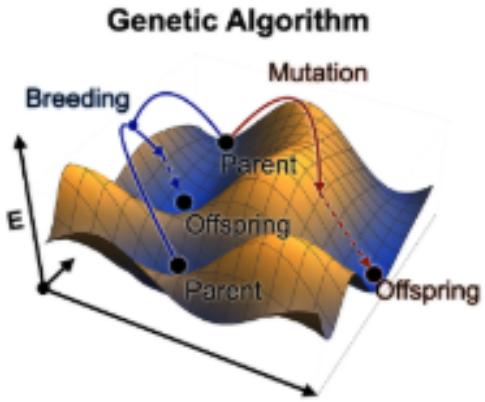

Figure 3. Genetic algorithm search scheme [7].

3 AGE methodology

Our approach relies on two core components: a genetic algorithm and a relaxation calculator. In our implementation, the genetic algorithm is provided by the Python-based ASE (Atomic Simulation Environment) [8] library, while the relaxation procedures are performed using two calculators: GULP [9] and Quantum ESPRESSO [10] (Figure 4).

Figure 4. AGE methodology [8–10].

GULP offers relatively lower accuracy due to its reliance on classical Coulomb-based potentials [9], but it provides significantly faster calculations. In contrast, Quantum ESPRESSO is based on density functional theory (DFT) [11, 12], offering much higher accuracy at the cost of considerably longer computation times.

Throughout the execution of the algorithm, we primarily rely on GULP to perform rapid evaluations over a number of generations (populations). The algorithm is terminated based on a predefined convergence criterion – in our case, the repetition of structures in the energy-fitness diagram, which will be discussed in detail later.

As a result, we obtain a pool of candidate structures that exhibit a symmetry and geometry resembling that of the target (reference) phase. These structures are then subjected to a final

relaxation using Quantum ESPRESSO. It is worth noting that at this stage, only a few dozen calculations are typically required, as opposed to the hundreds performed during the initial GULP-based exploration, significantly reducing the overall computational cost.

4 Flowchart of the genetic algorithm

The genetic algorithm is an iterative process [8] (Figure 5). Each iteration is referred to as a generation (or population). The initial population consists of randomly generated structures; when the energy of these structures is calculated, it yields a random distribution of points throughout the energy landscape.

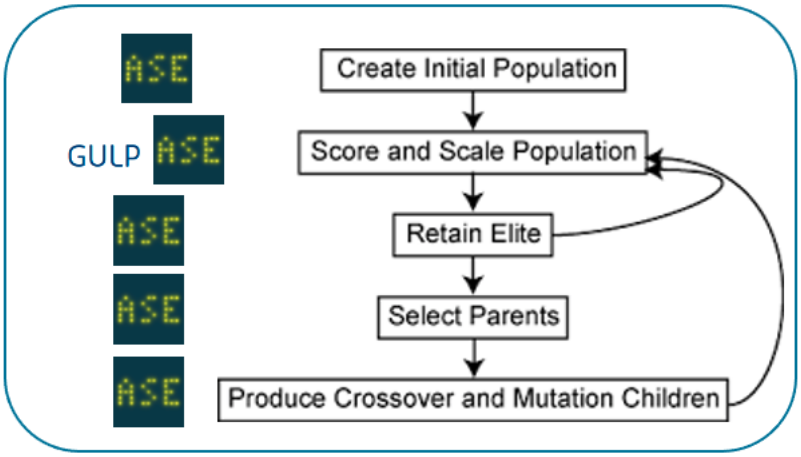

Figure 5. Iterative genetic algorithm [7, 8].

The next step involves evaluating and ranking these structures based on their energies. To do this, each structure undergoes relaxation (as discussed earlier), and its final relaxed energy is calculated. The structures are then sorted according to energy, with the lowest energy corresponding to the most stable configuration.

The subsequent population is formed by combining a certain percentage of the best-performing structures from the previous generation, a portion of mutated structures (i.e., previously selected structures modified via mutation operators), and a fraction of newly generated random structures (analogous to those in the first population). The relaxation-evaluation-selection cycle is then repeated.

Mutation operators play an important role in this process by enabling exploration of local minima near the parent structures. They introduce controlled perturbations that help the algorithm escape local traps and further sample the surrounding energy landscape.

5 Results for crystalline silica

To validate and confirm the effectiveness of our methodology, we selected a material that clearly demonstrates the presence of global and plenty of local minima in the energy landscape. Silicon dioxide (SiO_2) [13] was chosen for this purpose due to its rich polymorphism (Figure 6). In particular, the stable phase under ambient conditions – α-quartz – serves as the global minimum, while metastable phases such as β-quartz, α- and β-cristobalite, tridymite, keatite, coesite, and stishovite represent local minima.

The results of the population generations are conveniently visualized as diagrams (Figure 7), where the energy of each structure is plotted against its population number and position within

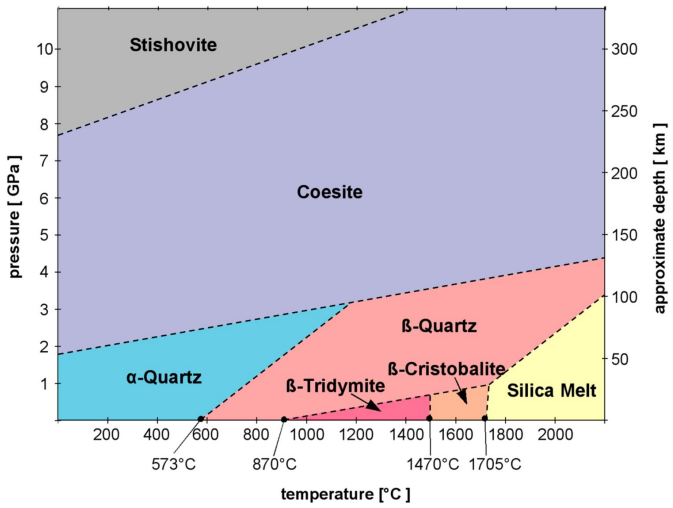

Figure 6. Phase diagram of silica [13].

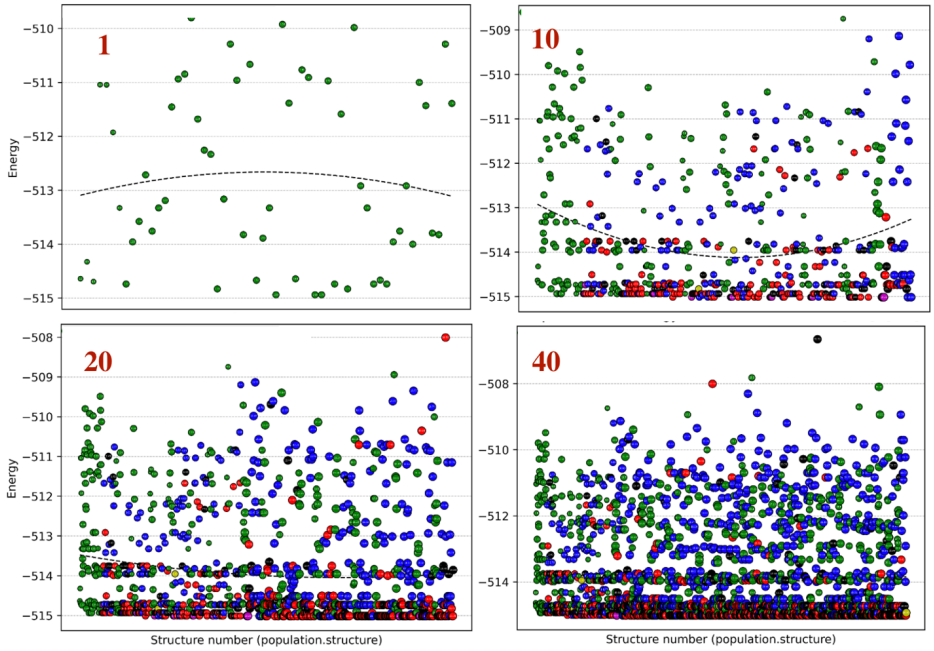

Figure 7. Results of silica population generation. Green circles represent randomly created structures; pink – best structures from previous population; blue, yellow and red – produced by pairing, softmutation and strain mutation variational operators; black – structures, removed from search because of inadequate parameters.

that population. Each structure is represented as a colored dot. The initial randomly generated population is shown in green, the best-performing structures are shown in purple, blue, yellow, and red (those generated through mutation operators), and black dots represent structures that were discarded due to an excessively large unit cell volume.

When comparing all four plots, a consistent pattern emerges: distinct horizontal lines formed by structures from different generations, each located at a specific energy level. These lines are visual representations of the local and global minima in the energy landscape. In this example, the lowest line corresponds to α-quartz, while the higher ones represent local minima corresponding to β-quartz, α- and β-cristobalite, tridymite, keatite, coesite, and stishovite. All structures in

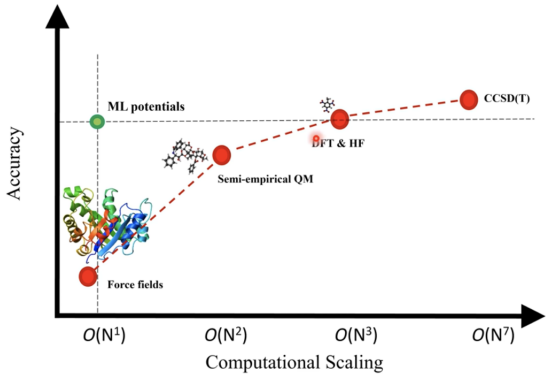

Figure 8. Dependence of calculation quality on resource consumption [17].

this phase diagram were successfully identified except coesite, which was not recovered because of its large unit cell size (we did not provide such massive calculations for this system).

To facilitate comparison, we compiled the results into Table 1, matching our predicted structures with the reference structures obtained from the Materials Project database [14]. Each structure in the table was carefully relaxed with Quantum ESPRESSO with the same input parameters to be sure our comparison is correct. We quantitatively evaluated structural similarity using the cosine distance between the fingerprint vectors of each structure (a cosine value of zero indicates identical structures). On the basis of this analysis, we conclude that the structures generated by our method are identical to the known reference phases.

Name	$a\ b\ c$ (Å)	$\alpha\ \beta\ \gamma$	Space group	Energy/atom (Ry)	Cosine distance
α-quartz	4.961 4.961 5.452	90° 90° 120°	152	−42.767192	0.0
	4.96 4.96 5.452	90° 90° 120°	152	−42.767615	
β-quartz	5.081 5.081 5.562	90° 90° 120°	180	−42.76719	0.0039
	5.081 5.081 5.562	90° 90° 120°	180	−42.767187	
β-tridymite	5.254 5.254 8.575	90° 90° 120°	194	−42.766982	0.0
	5.254 5.254 8.575	90° 90° 120°	194	−42.766976	
β-cristobalite	7.430 7.430 7.430	90° 90° 90°	227	−42.766908	0.0355
	7.430 7.430 7.430	90° 90° 90°	227	−42.766929	
stishovite	4.192 4.192 2.6805	90° 90° 90°	136	−42.762639	0.0069
	4.193 4.193 2.680	90° 90° 90°	136	−42.762690	
fluorite	4.543 4.543 4.543	90° 90° 90°	225	−42.684103	0.02
	4.545 4.545 4.545	90° 90° 90°	225	−42.684113	
hydrophilite	4.082 5.039 4.497	90° 90° 120°	60	−42.581702	0.0025
	4.082 5.041 4.497	90° 90° 120°	60	−42.758837	

Table 1. Results of evolutionary search (in pairs of structures – top row) and relaxation of structures from Materials Project (bottom).

6 Prospects and future improvements

The presence of two relaxers significantly reduced computation time and proved to be an effective approach for solving problems related to finding global and local minima. However, this introduced other challenges, particularly the need to create input files for the GULP relaxer, which greatly limits its usability.

If we plot the performance of the calculator against its resource consumption (Figure 8), our two relaxers occupy opposite corners. GULP is fast, but the least accurate, while Quantum ESPRESSO, as a more accurate DFT method, requires substantial resources. However, neural networks have managed to reduce resource consumption. This happens through training on datasets computed by DFT methods, which essentially means that the accuracy is limited to that of the initial training set, but the speed of machine learning is gained.

Therefore, the decision was made to apply a machine learning-based approach. As it turned out, such tools already exist, and there is even some variety available. We have chosen MACE [15] and AIMNet2 [16] pseudopotentials and integrated them in place of GULP within the genetic algorithm, and are currently in the process of tuning these programs.

7 Conclusions

We have implemented a methodology for predicting crystal structures using Python along with the GULP and Quantum Espresso relaxers. The developed approach was successfully validated on a range of crystalline systems. In particular, the study of silica demonstrated excellent results, as we were able to reproduce most of the structures presented in the phase diagram.

In future work, we plan to integrate MACE and AIMNet2, neural networks that will be used for structural relaxation and energy calculations. This implementations are expected to improve the accuracy of energy computations compared to GULP, while maintaining high processing.

Acknowledgements

We are grateful to all our colleagues for fruitful discussions on KAU seminar meetings – Dr. Prof. Oleksandr Kordyuk, Dr. Prof. Evhen Len, Dr. Volodymyr Bezguba, and to our collaborator from Carnegie Mellon University Dr. Prof. Olexandr Isayev. OF thanks Ulrike Nitzsche for technical assistance with supercomputer cluster of IFW Dresden.

Funding

We acknowledge financial support from BMBF through the GU-QuMat project (01DK24008).

References

[1] Pettifor D.G., Computer-aided materials design: bridging the gaps between physics, chemistry and engineering, *Phys. Educ.* **32** (1997), 164–168.

[2] Louie S.G., Chan Y.-H., da Jornada F.H., Li Z., Qiu D.Y., Discovering and understanding materials through computation, *Nature Mater.* 20 (2021), 728–735.

[3] Jansen M., The energy landscape concept and its implications for synthesis planning, *Pure Appl. Chem.* **86** (2014), 883–898.

[4] Wales D.J., Doye J.P.K., Global optimization by basin-hopping and the lowest energy structures of Lennard–Jones clusters containing up to 110 atoms, *J. Phys. Chem. A* **101** (1997), 5111–5116.

[5] Schaefer B., Mohr S., Amsler M., Goedecker S., Minima hopping guided path search: An efficient method for finding complex chemical reaction pathways, *J. Chem. Phys.* **140** (2014), 214102.

[6] Barducci A., Bonomi M., Parrinello M., Metadynamics, *Wiley Interdiscipl. Rev. Comput. Mol. Sci.* **1** (2011), 826–843.

[7] Falls Z., Avery P., Wang X., Hilleke K.P., Zurek E., The XtalOpt evolutionary algorithm for crystal structure prediction, *J. Phys. Chem. C* **125** (2021), 1601–1620.

[8] Larsen A.H., Mortensen J.J., Blomqvist J. et al., The atomic simulation environment – a Python library for working with atoms, *J. Phys. Cond. Mat.* **29** (2017), 273002.

[9] Gale J.D., Rohl A.L., The general utility lattice program (GULP), *Mol. Simul.* **29** (2003), 291–341.

[10] Giannozzi P., Baroni S., Bonini N. et al., QUANTUM ESPRESSO: a modular and open-source software project for quantumsimulations of materials, *J. Phys. Cond. Mat.* **21** (2009), 395502.

[11] Hohenberg P., Kohn W., Inhomogeneous electron gas, *Phys. Rev. B* **136** (1964), 864–871.

[12] Kohn W., Sham L.J., Self-consistent equations including exchange and correlation effects, *Phys. Rev. A* **140** (1965), 1133–1138.

[13] Royce K., Baars C., Viles H., Defining damage and susceptibility, with implications for mineral specimens and objects: Introducing the mineral susceptibility database, *Stud. Conserv.* **68** (2023), 298–317.

[14] Jain A., Ong S.P., Hautier G. et al., Commentary: The materials project: A materials genome approach to accelerating materials innovation, *APL Mater.* **1** (2013), 011002.

[15] Batatia I., Kovacs D.P., Simm G.N.C., Ortner C., Csányi G., MACE: Higher order equivariant message passing neural networks for fast and accurate force fields, *Adv. Neural Inf. Process. Syst.* **35** (2022), 7138.

[16] Zubatyuk R., Smith J.S., Leszczynski J., Isayev O., Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network, *Sci. Adv.* **5** (2019), eaav6490.

[17] Smith J.S., Isayev O., Roitberg A.E., ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost, *Chem. Sci.* **8** (2017), 3192–3203.

Реалізація методології передбачення кристалічних структур за допомогою генетичних алгоритмів імплементованих в бібліотеку Python ASE

Богдан Я. СЕМЕНЮК [1], *Олег Д. ФЕЯ* [2 3]

[1)] *Київський академічний університет, бул. Академіка Вернадського 36, Київ, Україна*
E-mail: semenuk.b.20@gmail.com
ORCID: https://orcid.org/0009-0007-0294-5859

[2)] *Київський академічний університет, бул. Академіка Вернадського 36, Київ, Україна*
E-mail: feyaolkodmi@gmail.com
ORCID: https://orcid.org/0000-0002-6749-950X

[3)] *Інститут металофізики ім. Г.В. Курдюмова НАН України, бул. Академіка Вернадського 36, Київ, Україна*

Отримано 6 червня 2025, у фінальній формі 25 червня 2025
https://doi.org/10.3842/confKAU.2025.phys.01

Анотація. Дана робота присвячена розробці та реалізації методології передбачення кристалічних структур за допомогою генетичних алгоритмів, які імплементовані в бібліотеку Python ASE. Застосування генетичних алгоритмів для передбачення кристалічних структур має великий потенціал у різних галузях, зокрема у матеріалознавстві та нанотехнологіях. Основною метою є розробка ефективного інструменту, який дозволить автоматизувати та прискорити процес пошуку оптимальних кристалічних структур для конкретних застосувань.

Ключові слова: передбачення кристалічних структур; генетичний алгоритм; енергетичний ландшафт; кристалічні поліморфи кремнезему; релаксація структури; ASE; GULP