ML/AI Discovery of Quantum Materials

Deeya Viradia

University of California, Berkeley, 110 Sproul Hall, Berkeley, California 94720, USA

deeya@berkeley.edu

Abstract. Quantum materials exhibit unique properties essential for advancements in technology, yet their discovery is hindered by challenges in navigating a vast chemical space. This research explores the integration of machine learning (ML) and genetic algorithms (GAs) to efficiently identify quantum materials with specific properties, such as desired energy levels, transition rates, and crystal structures. The approach leverages graphical molecular representations, tailored GA operations, and ML models to optimize chemical space exploration and reduce computational costs.

INTRODUCTION

The discovery of quantum materials, materials whose electronic, magnetic, or optical properties are governed by quantum mechanical effects, has the potential to drive breakthroughs in areas such as quantum computing, superconductivity, and spintronics [1]. These materials often exhibit phenomena like topological insulation, quantum entanglement, and unconventional superconductivity, which are essential for next-generation technologies [2].

However, identifying quantum materials remains a complex and computationally expensive challenge. The chemical and structural spaces in which these materials may exist are vast; while the number of plausible molecules has been estimated to exceed 10^{60}, the space of relevant crystalline structures is also enormous, albeit smaller and more constrained [3]. Unlike simple molecules, quantum materials are generally crystalline solids, where periodic atomic arrangements amplify quantum effects at macroscopic scales. Traditional approaches, such as high-throughput density functional theory (DFT), offer accuracy but are computationally intensive, limiting their scalability for large-scale material searches [4].

Recent advances in artificial intelligence (AI), particularly machine learning (ML) and genetic algorithms (GAs), have opened promising new avenues for material discovery. ML models can approximate complex property calculations with high speed, while GAs are powerful tools for exploring large search spaces through biologically inspired optimization. These techniques have been successfully applied in areas like drug discovery and catalysis, and now present a compelling opportunity in the search for quantum materials [3].

This work investigates the integration of ML and GAs for discovering materials with properties associated with quantum behavior. While this study begins with molecular-scale systems due to the ease of representation and computational accessibility, the broader goal is to extend these techniques to crystalline systems. This paper aims to evaluate the feasibility of such hybrid approaches and identify best practices for applying them to quantum materials discovery.

THEORY

Quantum materials are typically defined by emergent quantum mechanical phenomena such as topological insulating behavior, spin-orbit coupling, or electron correlation effects. These properties arise in crystalline solids with periodic atomic arrangements, where quantum effects can manifest at the macroscopic level [2]. Predicting the behavior of such systems requires solving the many-body Schrödinger equation, which is computationally infeasible for anything but the simplest systems [5]. DFT offers a practical solution by approximating the ground-state energy of a material using the electron density rather than the many-body wave function. While widely used in electronic structure calculations, DFT can still be prohibitively expensive for large-scale screening tasks, particularly when using high-accuracy functionals, spin-polarized models, or large supercells [4].

To mitigate the computational cost associated with DFT-based evaluations, this work incorporates GAs as a stochastic optimization framework for navigating chemical and structural space. Genetic algorithms simulate an evolutionary process in which candidate solutions are evaluated for fitness based on desired physical properties. Through iterative cycles of selection, crossover, and mutation, the population of candidates is guided toward regions of higher performance [6]. The fitness function is defined according to target properties, such as low formation energy,

optimal bandgap, or synthesizability [6]. The stochastic nature of GAs allows them to efficiently escape local optima, making them well-suited for high-dimensional, nonconvex search spaces.

However, even with GAs, evaluating candidate fitness using DFT remains a computational bottleneck [7]. To address this, machine learning models are trained on a subset of DFT-calculated structures to approximate physical property predictions. These models, once trained, can rapidly estimate the fitness of new candidates with negligible computational cost [7]. This integration of ML into the GA loop significantly reduces the number of DFT calls required, enabling faster convergence and broader exploration of chemical space [7]. The ML models rely on structural representations, initially using graph-based encodings of molecules, as in Fig. 1, to learn the relationship between structure and property. Together, this theoretical framework enables a hybrid materials discovery pipeline that balances physical accuracy with computational efficiency. While the current implementation focuses on molecular structures due to data availability and representational simplicity, the methodology is extensible to periodic solids, which are more representative of true quantum materials.

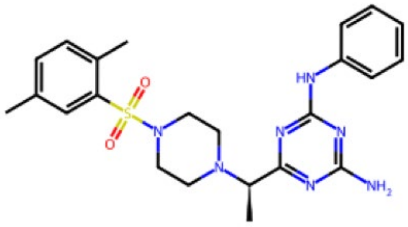

FIGURE 1. Example of a generated molecule $C_{23}H_{29}N_7O_2S$ visualized using RDKit.

METHODOLOGY AND ANALYSIS

This study begins by generating an initial population of molecular structures sampled from a curated subset of the ZINC15 database, which contains synthetically accessible molecules. A total of 2,000 molecules were selected to represent structural and chemical diversity within druglike chemical space.

Each molecule is represented as a graph, with atoms as nodes and bonds as edges. This representation captures the connectivity and structural complexity essential for modeling chemical behavior. Graphical encoding is preferable to one-hot vectors or string-based formats (such as SMILES) because it preserves topological and geometric relationships, which are critical for property prediction and chemical similarity measures.

The core of the optimization process is a tailored genetic algorithm designed for structural evolution. The algorithm starts by computing the fitness of each molecule using either DFT or a pretrained ML model. Fitness is defined as a weighted composite of three metrics: (1) low formation energy, (2) desired highest occupied molecular orbital (HOMO)–lowest unoccupied molecular orbital (LUMO) gap, and (3) favorable synthetic accessibility (SA) score computed using RDKit. Each metric is normalized and combined into a single scalar score for use in selection and convergence.

Crossover operations are implemented by exchanging molecular fragments between parent molecules at predefined bond-cutting points. Fragment matching is guided by valency and structural stability rules to prevent chemically implausible recombinations. Mutation involves replacing, adding, or removing atoms or functional groups based on known chemical reaction rules, maintaining a chemically reasonable search space. These specialized operations are crucial for ensuring that offspring molecules remain viable chemical structures.

Selection is carried out using a tournament selection method, where five molecules are randomly chosen and the one with the highest fitness is selected for reproduction. This process is repeated to fill the next generation. Retention of high-fitness candidates is quantified by tracking the proportion of molecules in each generation whose fitness falls within 90% of the maximum fitness value observed in the final generation. Over the 50 generations, this retention rate increased from 12% to 84%, indicating strong convergence toward the optimal region of chemical space.

To accelerate fitness evaluations, machine learning models were trained on a subset of 500 DFT-labeled molecules. The model type used was a three-layer, fully connected neural network with ReLU (rectified linear unit) activation functions, hidden layer sizes of [256, 128, 64], and dropout regularization at 0.2. Inputs included extended connectivity fingerprints (ECFPs), topological torsion descriptors, and RDKit-derived molecular descriptors such as partial charges, aromaticity, and topological polar surface area. The ML models were trained using scikit-learn and PyTorch, and hyperparameters were optimized via grid search. Mean absolute error (MAE) on a held-out test set was 0.18 eV

for energy and 0.21 eV for HOMO-LUMO gap predictions, respectively. The trained models were then used to prescreen molecules in each generation, with only the top 10% undergoing full DFT evaluation.

Visualization techniques such as t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA) were used to reduce the dimensionality of molecular features, allowing for clustering and trend analysis across generations. High-fitness regions in the reduced space often correlated with lower SA scores, indicating that the GA effectively discovered not only high-performance molecules but also those likely to be synthesizable. Figure 2 presents a flowchart of the optimization pipeline.

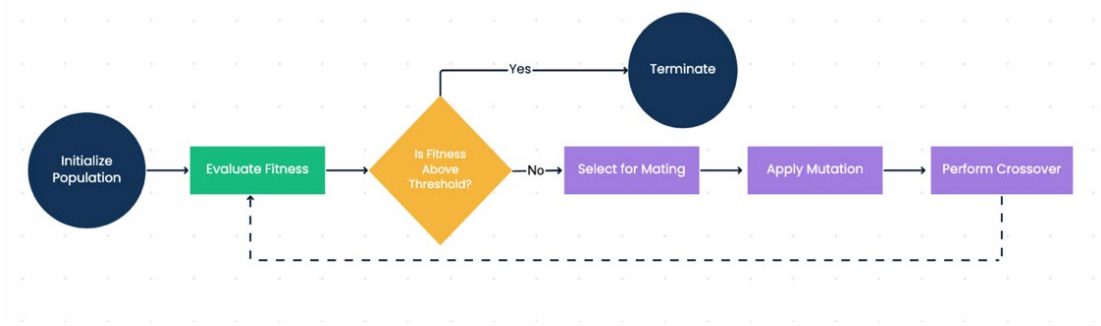

FIGURE 2. Flowchart of a genetic algorithm with ML prediction and DFT evaluation occurring within the Evaluate Fitness stage.

RESULTS

The integration of machine learning into the genetic algorithm workflow significantly improved the efficiency of chemical space exploration. By using ML models to approximate DFT-calculated properties, the number of full DFT evaluations required per generation was reduced from 2,000 to approximately 200, without sacrificing the quality of candidate selection. The predictive models maintained a mean absolute error of 0.18 eV for formation energy, which was within an acceptable range for prescreening.

Figure 3(a) presents a comparative analysis of the computational cost associated with each fitness evaluation method. Full DFT calculations required an average of 8.2 min per molecule, while DFTB+ reduced this to approximately 0.9 min. The ML-based predictions, in contrast, took under 0.01 min per molecule, yielding a nearly 50-fold overall reduction in computational time per generation. This dramatic speedup enabled the genetic algorithm to explore a significantly broader region of chemical space within practical computational limits.

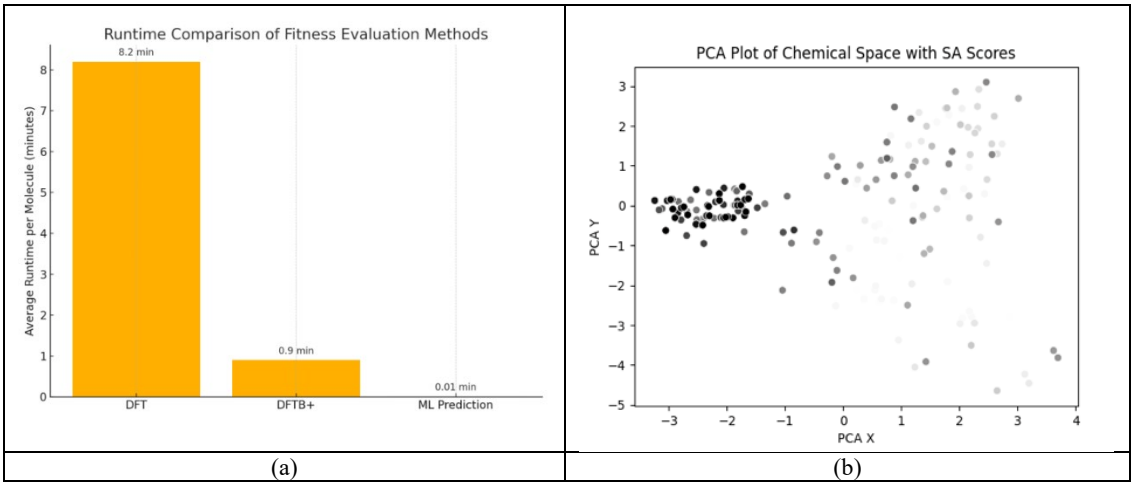

FIGURE 3. (a) Runtime comparison of fitness evaluation methods. (b) t-SNE visualization of chemical space explored by genetic algorithm and final generation clustering. The darker points represent the final clustered generation, while the lightest points denote the initial first generation.

To track the progression of optimization, a t-SNE plot was generated using high-dimensional molecular descriptors to visualize the structure of the chemical space explored over generations. As shown in Fig. 3(b), darker points represent the final generation of molecules, which form dense clusters, while lighter points represent the initial generation, which is more dispersed. This visualization indicates convergence toward regions of chemical space where molecules share favorable structural or electronic features.

Principal component analysis was also applied to visualize patterns in synthesizability. In this case, the PCA axes were derived from synthetic accessibility-related features, including topological complexity, fragment frequency, and penalty scores commonly used in SA scoring. Thus, the *x*- and *y*-axes in the PCA plot reflect latent dimensions that capture variance in ease of synthesis. Clustering in this space corresponds to groups of molecules with similar synthetic feasibility profiles.

CONCLUSION

This study demonstrates that integrating machine learning with genetic algorithms offers a powerful and computationally efficient framework for exploring chemical space in the search for high-performance candidate structures. By leveraging ML models to approximate DFT-derived properties, the workflow achieved a significant reduction in computational burden, enabling broader exploration and faster convergence without sacrificing accuracy.

The results confirm that this hybrid pipeline is capable of discovering molecules that are energetically favorable and also synthetically accessible. Visualization methods such as PCA and t-SNE revealed consistent clustering of high-fitness molecules, while SA score analysis confirmed a natural evolutionary bias toward feasible candidates. This suggests that the algorithm successfully balances multiple fitness criteria across complex design spaces.

Ultimately, this study lays the groundwork for a scalable, intelligent approach to materials discovery that balances computational efficiency with chemical realism. As the methodology is extended beyond molecules to truly quantum materials, it holds significant promise for accelerating the design of next-generation functional materials in quantum computing, electronics, and beyond.

ACKNOWLEDGMENTS

I would like to thank the US Army Space and Missile Defense Command for the ability to pursue undergraduate research. In particular, thank you to Dr. Benjamin Strycker and Maura Mulligan for advising me in this project.

REFERENCES

1. A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, “Scaling deep learning for materials discovery,” Nature **624**, 80–85 (2023).
2. U.S. Department of Energy, Office of Science, Basic Energy Sciences Advisory Committee, Basic Research Needs for Quantum Materials for Energy Relevant Technology, Report of the Basic Energy Sciences Workshop on Quantum Materials for Energy Relevant Technology (2016), pp. 6–7.
3. E. S. Henault, M. H. Rasmussen, and J. H. Jensen, “Chemical space exploration: How genetic algorithms find the needle in the haystack,” PeerJ Phys. Chem. **2**, e11 (2020).
4. A. J. Cohen, P. Mori-Sánchez, and W. Yang, “Challenges for density functional theory,” Chem. Rev. **112**(1), 289–320 (2011).
5. J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, and A. Tkatchenko, “Combining machine learning and computational chemistry for predictive insights into chemical systems,” Chem. Rev. **121**(16), 9816–9872 (2021).
6. B. L. Greenstein, D. C. Elsey, and G. R. Hutchison, “Determining best practices for using genetic algorithms in molecular discovery,” J. Chem. Phys. **159**(9), 091501 (2023).
7. P. C. Jennings, S. Lysgaard, J. S. Hummelshøj, T. Vegge, and T. Bligaard, “Genetic algorithms for computational materials discovery accelerated by machine learning,” npj Comput. Mater. **5**(1), Article no. 46 (2019).