УДК: 620.1:004.89 DOI 10.36910/775.24153966.2026.85.29

Н. Медюх, А. Красіков, О. Васільєв

Інститут проблем матеріалознавства імені І.М. Францевича НАН України

ПОШУК МАТЕРІАЛІВ, ПОДІБНИХ ДО КАРБІДУ БОРУ НА ОСНОВІ ВІДКРИТИХ ОБЧИСЛЮВАЛЬНИХ БАЗ ДАНИХ МАТЕРІАЛІВ

У цьому дослідженні представлено підхід для виявлення матеріалів із властивостями, подібними до цільових сполук, з використанням відкритих обчислювальних баз даних матеріалів. Як тестовий матеріал використовується карбід бору (B_4C) із певними співвідношеннями бор-вуглець (0,78–0,91), для якого розроблено послідовність зі збору, обробки та аналізу даних. Для групування матеріалів за структурними, енергетичними та механічними властивостями застосовуються шість алгоритмів кластеризації, що представляють різні методологічні підходи, а також кілька методів селекції властивостей. Запропонована методологія успішно виявляє кластери з високою концентрацією цільових складів карбіду бору та розкриває інші бороровмісні матеріали зі схожими властивостями. Цей підхід демонструє потенціал використання відкритих баз даних матеріалів для прискореного пошуку нових матеріалів і забезпечує підґрунтя, яке можна застосувати до різноманітних матеріальних систем поза межами карбіду бору.

Ключові слова: *обчислювальний скринінг матеріалів, інформатика матеріалів, карбід бору*

N. Mediukh, A. Krasikov, O. Vasiliev

DATA-DRIVEN DISCOVERY OF BORON CARBIDE-LIKE MATERIALS USING OPEN COMPUTATIONAL MATERIALS DATABASES

This study presents a data-driven approach for identifying materials with similar properties to target compounds using open computational materials databases. We use boron carbide (B_4C) with specific boron-to-carbon ratios (0.78-0.91) as our test case material and develop a robust pipeline for data collection, processing, and analysis. Six clustering algorithms representing different methodologies and several feature engineering techniques are employed to group materials based on structural, energetic, and mechanical properties. Our methodology successfully identifies clusters with high concentrations of target boron carbide compositions and reveals other boron-containing materials with similar properties. This approach demonstrates the potential of leveraging open materials databases for accelerated materials discovery and provides a framework applicable to various material systems beyond boron carbide.

Keywords: *computational materials discovery, materials informatics, boron carbide*

1. Introduction

In this context, computational materials discovery has advanced significantly through the development of high-throughput screening methodologies, primarily based on density functional theory (DFT) calculations. These efforts are substantially supported by the growth of open materials databases like the Materials Project [1] and NOMAD (Novel Materials Discovery) [2]. Such platforms facilitate property-based filtering and similarity searches across extensive datasets of calculated materials. However, a limitation of these approaches is their frequent reliance on pre-existing knowledge and defined chemical descriptors, which may restrict the identification of unanticipated compositions or structures with desired mechanical properties. To address these limitations, the field of materials informatics has emerged, with machine learning (ML) techniques becoming integral to materials research [3]. For example, supervised ML models have been successfully applied to predict various material properties [4]. Unsupervised learning methods, particularly clustering algorithms, are gaining attention for their capacity to identify inherent patterns and relationships within large materials datasets without requiring a priori labeling of material characteristics [5].

The application of clustering algorithms in materials discovery is predicated on their ability to perform unsupervised pattern recognition within complex, high-dimensional datasets. These methods can reveal intrinsic groupings of materials that share similarities in their features, which might not be discernible through conventional analysis [6]. The central hypothesis is that materials exhibiting analogous fundamental crystallographic, energetic, or mechanical properties will co-locate within a defined feature space. This offers a pathway to discover novel material analogues, such as those with mechanical properties similar to boron carbide, by identifying compounds that cluster with the target material, irrespective of exact stoichiometry or elemental composition. However, significant challenges in applying clustering to materials data include the appropriate selection and representation of material features (descriptors) relevant to mechanical behavior [5]. The choice of clustering algorithm and the subsequent validation of cluster quality both require careful consideration and often require domain-specific expertise. Furthermore, the development of quantitative evaluation metrics that are meaningful for materials science applications focusing on mechanical properties is essential for translating clustering results into actionable insights.

Boron carbide (B_4C) is a ceramic material characterized by a combination of desirable properties, including ultra-high hardness and low density [7]. These attributes underpin its utilization in demanding applications such as ballistic armor and abrasive cutting tools [8]. Nevertheless, challenges associated with the synthesis and processing of boron carbide, including high sintering temperatures and difficulties in achieving full densification, motivate the search for alternative materials with comparable or superior mechanical performance characteristics [9]. The discovery of such novel materials traditionally relies on experimental methodologies that are often resource-intensive and time-consuming. The combinatorial explosion of possible chemical compositions and structural configurations renders exhaustive experimental exploration impractical. Consequently, computational approaches are increasingly relevant for guiding and accelerating experimental efforts in materials discovery [10].

This work introduces a systematic, data-driven framework to identify novel materials with mechanical properties analogous to boron carbide. By applying unsupervised clustering to computational materials databases, we address the core challenges of feature engineering and algorithm selection to establish a robust pipeline for materials discovery. Our approach evaluates diverse feature sets and clustering models, using domain-specific metrics to ensure physical relevance. Ultimately, this methodology identifies specific candidates for experimental validation and provides a transferable architecture for targeting custom mechanical attributes across diverse material systems.

2. Materials and Methods

2.1 Data Collection and Database Construction

Materials data acquisition was evaluated through two primary approaches: the OPTIMADE [11] interface and direct database APIs. OPTIMADE is a standardized API specification that provides unified access to multiple materials databases, including AFLOW [12], Materials Project [1], and NOMAD [2]. However, practical implementation revealed significant limitations in the OPTIMADE approach. Individual database providers implement the OPTIMADE specification inconsistently, leading to incompatible data structures and response formats across providers. Additionally, the federated queries frequently encountered random API errors and timeout issues, making large-scale data harvesting unreliable. The theoretical advantage of accessing multiple databases through a single interface was undermined by these implementation inconsistencies and reliability issues.

The NOMAD database was selected as the primary data source due to its native API stability and comprehensive data coverage. Unlike the OPTIMADE interface, NOMAD's native API provides direct, stable access to its complete repository of computational materials data without the intermediary standardization layer that introduces inconsistencies. The NOMAD API demonstrated superior reliability for large-scale data harvesting operations and provided access to detailed computational metadata necessary for this analysis.

B_xZ_y, where Z - any element, materials data were systematically harvested from the NOMAD database using its native API (https://nomad-lab.eu/prod/v1/api/v1/entries). The data collection focused exclusively on Density Functional Theory (DFT) calculations to ensure computational consistency across the dataset. Filtering criteria included restricting the computational method to DFT calculations, requiring complete system geometry information, and requiring the availability of both total energy and stress tensor data. All available chemical systems were included without initial composition-based filtering to maintain dataset diversity.

The raw NOMAD data were stored in a MongoDB database to facilitate efficient querying and processing of large-scale materials data. This intermediate storage approach enabled robust error handling during data acquisition and flexible downstream analysis. The current analysis was executed on all 70k materials records available in the NOMAD database by the B* search criteria.

2.2 Dataset Preparation and Feature Engineering

From each DFT calculation record, 13 quantitative features were systematically extracted to characterize the materials. Structural features included crystal space group number, unit cell volume per atom, lattice parameters (a, b, c lengths), and lattice angles (α, β, γ). Chemical features comprised boron atomic fraction (B_ratio), atomic number of the second most abundant element, and periodic table group of the second element. Physical features included the DFT total energy per atom and the von Mises equivalent stress calculated from the stress tensor components.

Statistical outliers were identified and removed using a 3-sigma criterion applied specifically to the energy and stress features (energy_per_atom, equivalent_stress). This targeted approach preserved natural diversity in structural parameters while removing potentially erroneous computational results. All features

were standardized using z-score normalization, and categorical variables (space groups) were handled with one-hot encoding when necessary.

Four distinct feature sets were designed to investigate the relative importance of different material properties: (1) all_features - complete 13-feature set including structural, chemical, and physical properties; (2) structural - crystal structure and lattice parameters only; (3) energy_stress - DFT energy and mechanical stress only; (4) structural_energy_stress - combined structural and energy/stress features excluding chemical composition.

Principal Component Analysis (PCA) was systematically applied to each feature set with varying numbers of components (2, 3, 4, 5, 6, and no PCA) to investigate the effect of dimensionality reduction on clustering performance.

2.3 Clustering Methodology

Six distinct clustering algorithms were implemented to capture different clustering paradigms: K-Means (centroid-based), Gaussian Mixture Model (probabilistic), Agglomerative Clustering (hierarchical with Ward linkage), Birch (hierarchical optimized for large datasets), DBSCAN (density-based with noise detection), and HDBSCAN (hierarchical density-based with varying density handling).

Each algorithm underwent systematic parameter optimization. K-Means, Agglomerative, Birch, and Gaussian Mixture Model were tested with cluster numbers from 2 to 20. DBSCAN explored epsilon values from 0.1 to 2.0 with minimum samples fixed at 5. HDBSCAN tested 36 parameter combinations varying minimum samples (5-30) and minimum cluster size (5-30).

The complete experimental matrix consisted of 4 feature sets × 6 PCA configurations × 6 algorithms with their parameter grids, resulting in 2,317 unique clustering experiments. All results were stored in MongoDB with complete parameter tracking for reproducibility and systematic analysis.

2.4 Discovery-Focused Evaluation Framework

The evaluation framework was designed as a general methodology for identifying clusters enriched with target materials, demonstrated here using boron carbide (BC) compounds as an example case. Target materials were defined as compounds containing both boron and carbon atoms with a boron-to-carbon atomic ratio within $0.78 \leq B/(B+C) \leq 0.91$, representing established BC stoichiometries. This approach can be adapted for any target material class by modifying the identification criteria.

A comprehensive scoring system was developed, combining multiple evaluation criteria:

- Silhouette Score: Traditional clustering quality metric ensuring geometric coherence, though weighted lower than domain-specific criteria.

- Cluster Count Score: Favors clustering solutions producing 10-25 total clusters, balancing granularity with interpretability. Solutions with fewer than 5 clusters receive minimal scores, while those exceeding 40 clusters are penalized for over-fragmentation.

- Target Concentration Score: Evaluates how effectively target compounds (BC materials) are concentrated within a few clusters. Optimal scores are achieved when 70%+ of target compounds reside in a single cluster, or 80%+ in two clusters. Solutions with target compounds dispersed across many clusters receive lower scores.

- Diversity Score: Assesses the composition diversity within target-rich clusters to ensure discovery potential. Clusters with 20-70% target compound ratios receive maximum scores, representing an optimal balance for identifying similar materials. Pure target clusters receive low scores because they offer limited discovery opportunities.

Each score range is from 0 to 1. Those scores are combined as a linear combination with the same weights to get the final score. The final score shows not only whether this clustering attempt is good enough according to the classical silhouette score, but also whether we can discover materials similar to boron carbide.

3. Results and Discussion

3.1 Data Collection Performance

Our data collection pipeline successfully harvested materials data from open computational material databases using Optimade API (via Python client library and via direct API calls) and Nomad API. Table 1 summarizes the data collection performance across different APIs and providers.

The NOMAD native API provided the most comprehensive dataset, but required significantly more time for complete harvesting. For targeted studies focusing on specific element combinations (e.g., boron carbide), filtered queries substantially reduced collection time while maintaining a high yield of relevant data.

Table 1

Data collection performance comparison

Method	Provider	Structures Retrieved	Time (min)	Success Rate (%)
OPTIMADE Client	All	20,957	~10	95.3
OPTIMADE Client (filtered)	All	4,971	~5	94.7
OPTIMADE API	NOMAD	9,920	~15	99.6
OPTIMADE API	AFLOW	100,000	~120	98.2
NOMAD Native API	NOMAD	13,000,000	~1,800	99.9

3.2 Dataset Characteristics and Clustering Performance

Our systematic clustering analysis processed 29,378 materials records from the NOMAD database after outlier removal, representing a comprehensive survey of DFT-calculated crystalline materials. The dataset encompassed diverse chemical compositions, with particular focus on boron-containing compounds as our target case study. A total of 2,317 distinct clustering configurations were evaluated across six algorithms, four feature sets, and multiple PCA dimensionalities, demonstrating the robustness of our systematic approach.

3.2.1 Algorithm Performance Comparison

The comprehensive evaluation revealed significant performance differences across clustering algorithms and feature sets (Table 2). Birch clustering consistently achieved the highest performance with an average score of 0.552 across all configurations, followed by DBSCAN (0.511) and K-means (0.486). The superior performance of Birch can be attributed to its hierarchical approach and optimization for large datasets, making it particularly well-suited for materials discovery applications.

Table 2.

Algorithm Performance Summary

Algorithm	Average Score	Best Configuration	Optimal Features
Birch	0.552	0.801	energy_mechanics
DBSCAN	0.511	0.797	energy_mechanics
K-means	0.486	0.691	energy_mechanics
Agglomerative	0.472	0.700	energy_mechanics
GMM	0.454	0.608	all_features
HDBSCAN	0.454	0.586	structural (PCA=3)

3.2.2 Feature Set Effectiveness

The energy_mechanics feature set demonstrated superior performance (average score 0.567) compared to other feature combinations, validating our hypothesis that energetic and mechanical properties are primary determinants of materials similarity. This finding suggests that materials with similar thermodynamic stability and mechanical response tend to exhibit analogous behaviors regardless of their exact chemical composition.

The structural feature set showed moderate performance (0.481), while the complete all_features set achieved comparable results (0.480). Interestingly, the structural_energy_mechanics combination (0.468) performed slightly lower than energy_mechanics alone, suggesting that the inclusion of structural parameters may introduce noise when energetic properties are already well-represented.

3.3 Target Material Identification

3.3.1 Optimal Clustering Configuration

The best-performing configuration employed Birch clustering with energy_mechanics features and no PCA dimensionality reduction, achieving a combined score of 0.801. This configuration successfully concentrated 65 boron carbide compounds with target B/(B+C) ratios (0.78-0.91) into 5 clusters out of 18 total clusters (Fig. 1).

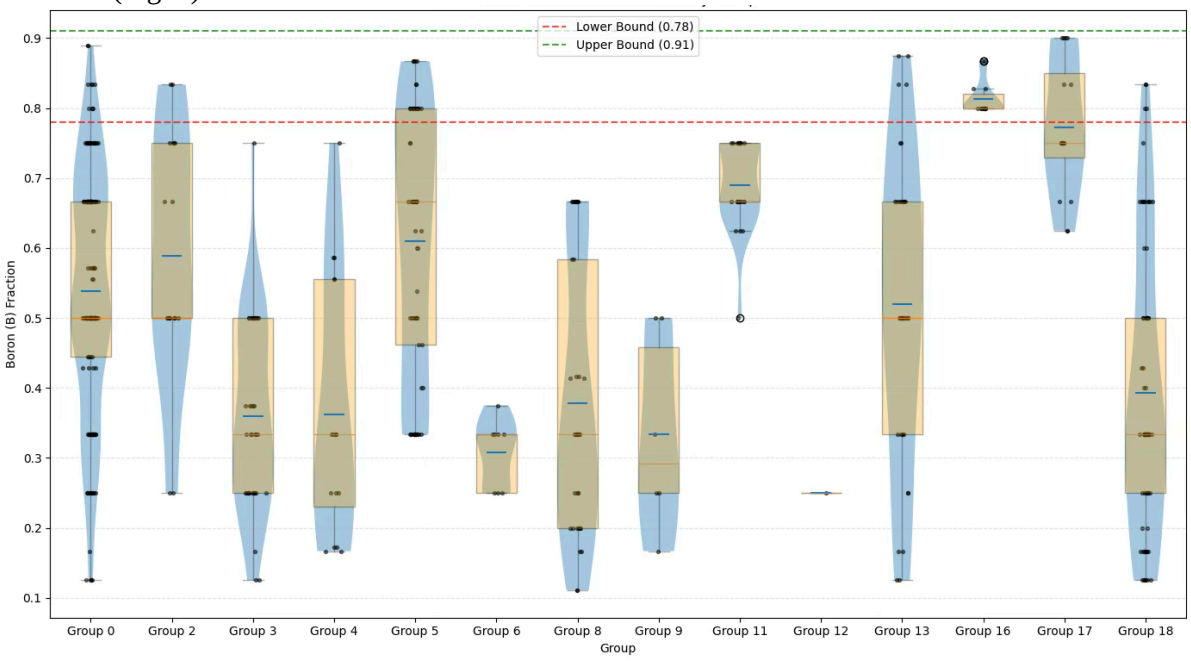

Fig. 1. **Distribution of Boron content per cluster. Only clusters that have at least one boron carbide calculation are represented**

3.3.2 Discovery Potential Analysis

The clustering analysis identified two distinct discovery clusters with contrasting compositional profiles, each offering unique insights for materials discovery.

Cluster 16 contained 148 compounds, dominated by alkali and alkaline-earth metal borides, exhibiting remarkable compositional consistency. The boron ratios ranged from 0.75 to 0.923, with most compounds falling in the 0.857-0.923 range—closely matching the target B_4C stoichiometry. The cluster was characterized by extensive representation of alkaline earth metal borides, including B-Ba (17 compounds), B-Sr (15 compounds), B-Ca (11 compounds), and B-Mg (11 compounds), alongside alkali metal borides such as B-Li (16 compounds), B-K (13 compounds), B-Na (10 compounds), and B-Cs (10 compounds). Rare earth borides were also present, notably B-Y (18 compounds) and B-La (13 compounds), while pure borocarbides (C-B, 14 compounds) exhibited boron ratios between 0.80 and 0.87. This cluster's chemical coherence suggests shared electronic structure and bonding characteristics relevant to B_4C-like materials.

Cluster 17 comprised 81 compounds exhibiting greater chemical diversity, capturing materials with varied bonding character. Boron oxides (B-O, 25 compounds, boron ratio 0.89) dominated this cluster. Metal borides were represented by B-Na (15 compounds, boron ratio 0.87 - 0.94) and B-Mg (13 compounds, boron ratio 0.80 - 0.88). C-B borocarbides (16 compounds) spanned a wide compositional range from 0.625 to 0.90. The cluster also contained B-H borohydrides (12 compounds, boron ratio 0.29 - 0.38), representing alternative boron coordination environments distinct from the target B_4C structure.

The separation into electropositive metal borides and mixed-bonding structures provides actionable chemical insights for materials discovery. The first cluster suggests that electropositive metals may stabilize boron-rich structures through ionic bonding mechanisms similar to those in B_4C, making these compounds promising candidates for experimental synthesis. The second cluster, with its diverse bonding environments including covalent B-O and coordination-based B-H interactions, represents alternative pathways to achieve similar characteristics through fundamentally different chemical strategies. This focused clustering demonstrates that machine learning-driven materials discovery can simultaneously achieve high selectivity and chemical interpretability when appropriate feature engineering and scoring metrics are employed.

4. Conclusions

This study presents a systematic data-driven methodology for identifying materials with similar properties to target compounds using open computational material databases. Our comprehensive analysis of 70k DFT-calculated materials from the NOMAD database, evaluated through 2,317 distinct clustering configurations, establishes several key contributions to computational materials discovery.

The systematic harvesting and processing of large-scale materials data from NOMAD demonstrates the feasibility of leveraging open databases for materials discovery. Our multi-tier feature engineering approach, encompassing structural, energetic, mechanical, and chemical descriptors, combined with rigorous outlier removal and standardization protocols, provides a robust foundation for unsupervised analysis. The comprehensive evaluation framework, testing six clustering algorithms across four feature sets with systematic PCA dimensionality reduction, ensures methodological rigor and reproducibility.

Our approach offers several distinct advantages over traditional materials discovery methods. Unsupervised clustering eliminates composition-based biases, enabling the discovery of non-obvious material analogues across diverse chemical spaces. The methodology scales naturally with database growth, becoming more powerful as computational repositories expand. Emphasis on energetic and mechanical descriptors captures fundamental principles of similarity that transcend elemental composition.

The framework's generalizability extends far beyond boron carbide systems. The methodology can be readily adapted to any target material class by modifying identification criteria and adjusting scoring weights according to desired properties. The convergence of energy-mechanics features across multiple algorithms suggests universal applicability of thermodynamic and mechanical property descriptors for materials similarity assessment.

References

1. A. Jain *et al.*, “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation,” *APL Materials*, vol. 1, no. 1, p. 011002, Jul. 2013, doi: https://doi.org/10.1063/1.4812323.

2. M. Scheidgen *et al.*, “NOMAD: A distributed web-based platform for managing materials science research data,” *The Journal of Open Source Software*, vol. 8, no. 90, pp. 5388–5388, Oct. 2023, doi: https://doi.org/10.21105/joss.05388.

3. R. Batra, L. Song, and R. Ramprasad, “Emerging materials intelligence ecosystems propelled by machine learning,” *Nature Reviews Materials*, vol. 6, no. 8, Nov. 2020, doi: https://doi.org/10.1038/s41578-020-00255-y.

4. A. S. Rosen *et al.*, “Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery,” *Matter*, vol. 4, no. 5, pp. 1578–1597, May 2021, doi: https://doi.org/10.1016/j.matt.2021.02.015.

5. N. Bhat, N. Birbilis, and A. S. Barnard, “Unsupervised learning and pattern recognition in alloy design,” *Digital Discovery*, vol. 3, no. 12, pp. 2396–2416, 2024, doi: https://doi.org/10.1039/d4dd00282b.

6. J. Singh and D. Singh, “A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects,” *Advanced Engineering Informatics*, vol. 62, p. 102799, Sep. 2024, doi: https://doi.org/10.1016/j.aei.2024.102799.

7. V. Domnich, S. Reynaud, R. A. Haber, and M. Chhowalla, “Boron Carbide: Structure, Properties, and Stability under Stress,” *Journal of the American Ceramic Society*, vol. 94, no. 11, pp. 3605–3628, Oct. 2011, doi: https://doi.org/10.1111/j.1551-2916.2011.04865.x.

8. S. G. Savio, K. Ramanjaneyulu, Vemuri Madhu, and T. Balakrishna Bhat, “An experimental study on ballistic performance of boron carbide tiles,” *International Journal of Impact Engineering*, vol. 38, no. 7, pp. 535–541, Jul. 2011, doi: https://doi.org/10.1016/j.ijimpeng.2011.01.006.

9. A. B. Dresch, J. Venturini, S. Arcaro, O. R. K. Montedo, and C. P. Bergmann, “Ballistic ceramics and analysis of their mechanical properties for armour applications: A review,” *Ceramics International*, vol. 47, no. 7, pp. 8743–8761, Apr. 2021, doi: https://doi.org/10.1016/j.ceramint.2020.12.095.

10. M. Cheng *et al.*, “Artificial intelligence-driven approaches for materials design and discovery,” *Nature Materials*, vol. 25, no. 2, pp. 174–190, Jan. 2026, doi: https://doi.org/10.1038/s41563-025-02403-7.

11. M. L. Evans *et al.*, “Developments and applications of the OPTIMADE API for materials discovery, design, and data exchange,” *Digital discovery*, vol. 3, no. 8, Jan. 2024, doi: https://doi.org/10.1039/d4dd00039k.

12. *M. J. Mehl et al., “The AFLOW Library of Crystallographic Prototypes,” arXiv.org, 2016. https://*arxiv.org/abs/1607.02532 (accessed Apr. 01, 2026).