ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

A Comparative Study of Structural Representations for 2D Materials: Insights from Dynamic Collision Fingerprint and Matminer Library

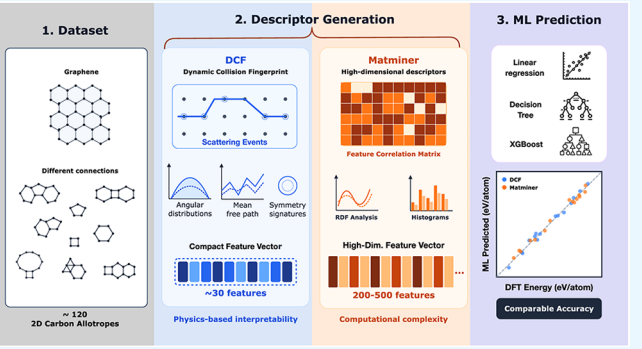
Raphael M. Tromer, Isaac M. Felix, Rafael Besse, Marcelo L. Pereira Junior,* and Marcos G. E. da Luz

Cite This: *ACS Omega* 2026, 11, 42389−42396

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In materials science, the choice of structural descriptors for machine learning protocols strongly influences both predictive performance and model interpretability. High-dimensional descriptors can improve numerical accuracy, but often introduce substantial computational overhead and reduce transparency. To address this, the Dynamic Collision Fingerprint (DCF) framework generates concise descriptors via the dynamical probing of atomic structures. In this work, we benchmark DCF against the widely used Matminer library using a data set of 120 two-dimensional (2D) carbon allotropes. We evaluate performance across three regression algorithms, linear regression, decision trees, and XGBoost, utilizing train-test partitions from 10% to 90%. Our results demonstrate that DCF matches the predictive accuracy of Matminer across all algorithms. Although Matminer can be faster in terms of execution time, DCF accomplishes its predictive performance using descriptors that are significantly lower-dimensional, pointing to manageable computing costs in feature space. Moreover, compared to the rather technical Matminer descriptions, DCF exhibits considerably clearer physical interpretability. These findings suggest that DCF serves as a viable alternative to high-dimensional descriptor libraries for structural representation, since it remains both computationally flexible and physically grounded.



INTRODUCTION

Atomic level structural characterization constitutes a central element of materials science and computational chemistry. It provides the basis for establishing structure−property relationships that guide both experimental interpretation and the rational design of novel materials.[1−9] Over the past decade, the integration of atomistic simulations, structural descriptors, and machine learning models has significantly transformed high throughput materials discovery.[8−13] In this context, the definition of descriptors that are simultaneously efficient, interpretable, and transferable has emerged as a critical challenge, since the predictive capability and robustness of machine learning models depend strongly on the quality and representational adequacy of the selected structural features.[14−16]

A large variety of structural descriptors has been proposed and consolidated in the literature.[17−24] These approaches include representations based on radial and angular distribution functions, orientational order parameters, spectral formulations such as the Smooth Overlap of Atomic Positions (SOAP) method,[17] and graph based or topology driven embeddings derived from message passing frameworks.[24,25] Many of these descriptors are available through dedicated software libraries. Among them Matminer[26] has become widely adopted in computational materials modeling due to its extensive collection of structural, electronic, and chemical features and its support for reproducible workflows across diverse material classes.

Despite such versatility, the employment of generic high-dimensional descriptor libraries often introduces practical limitations when applied to structurally complex systems, the case of two-dimensional (2D) materials. A first limitation concerns the physical interpretability of many features, as they cannot be directly associated with intuitive structural characteristics. A second limitation concerns sensitivity to disorder, defects, and aperiodicity, which are frequently present in 2D systems, where local distortions, vacancies, and topological irregularities are common.[27−30] Beyond idealized atomically thin crystals, the broader class of complex 2D materials also includes confined and interfacially stabilized systems, such as the graphitic-like gallium nitride obtained at the graphene/SiC interface[31] and the 2D indium oxide proposed at the same interface,[32] whose structural diversity further requires the

Received: March 23, 2026
Revised: June 17, 2026
Accepted: June 24, 2026
Published: July 2, 2026

development of robust and physically meaningful descriptors for data-driven analysis.

Motivated by these challenges, a recently proposed descriptor scheme, the Dynamic Collision Fingerprint (DCF), was introduced aiming to be a more physically based alternative.[33] Guided by well-established concepts from classical statistical mechanics,[34,35] DCF departs from static coordinate-based representations. Instead, it seeks to dynamically probe atomic structures through trajectories of idealized particles undergoing elastic collisions with the lattice. In this way, statistical analyses of free paths, collision angles, recurrence events, and angular symmetries obtained through Fourier analysis and Shannon entropy allow the descriptor to encode structural signatures associated with symmetry, porosity, and disorder.[33] This construction yields descriptors that are directly interpretable in physical terms, independent of electronic or energetic models, and computationally adjustable according to the assumed sampling parameters.

Notwithstanding the promising performance of DCF displayed in the original work,[33] a thorough evaluation against popular high-dimensional descriptor libraries and under controlled and statistically consistent benchmarking settings is still missing. In this contribution we fulfill this gap by systematically contrasting the DCF with the Matminer descriptor library. Using a data set composed of 120 distinct 2D carbon allotropes, descriptors generated by both approaches are considered within three representative machine learning models, namely, linear regression,[36] decision trees,[37] and XGBoost.[38] Also, a wide variety of train and test partitions are examined in the comparisons.

The comprehensive analyses carried out in the present work allow for the evaluation of DCF's relative performance in terms of physical interpretability, descriptor compactness, computational cost-performance trade-off, prediction accuracy, and resilience. Indeed, our results demonstrate that DCF can serve as a complementary and, in certain settings, competitive descriptor framework for materials informatics workflows. Specifically, by framing structural characterization as a dynamical response problem rather than a static geometric structure, we demonstrate that predictors can remain remarkably compact while retaining the critical information required to accurately model 2D materials, even though the baseline descriptor generation cost may depend on the sampling settings.

METHODOLOGY

The methodology followed here comprises data set preparation, descriptor construction, machine learning modeling, and statistical evaluation, all performed under consistently specified conditions to ensure direct comparability between descriptor approaches. A schematic overview of the complete workflow adopted in this study is presented in Figure 1, summarizing the main stages from data set preparation to descriptor generation, model training, and statistical validation.

The benchmark data set employed in this study is entirely computational, comprising 120 distinct 2D carbon allotropes whose structures and formation energies were obtained from density functional theory (DFT) calculations reported in the literature.[39] Although materials databases can in general be assembled from either first-principles calculations or experimental measurements, the benchmark adopted in this work is specifically based on DFT-derived data. All coordinates and lattice information for the addressed structures are provided in the Supporting Information (SI). The systems were standardized using the Pymatgen library.[40] Each structure was reduced to its primitive cell, assuming a symmetry tolerance of 10^{-3} Å, and atomic coordinates were scaled to ensure uniform density normalization. The chosen target property is the

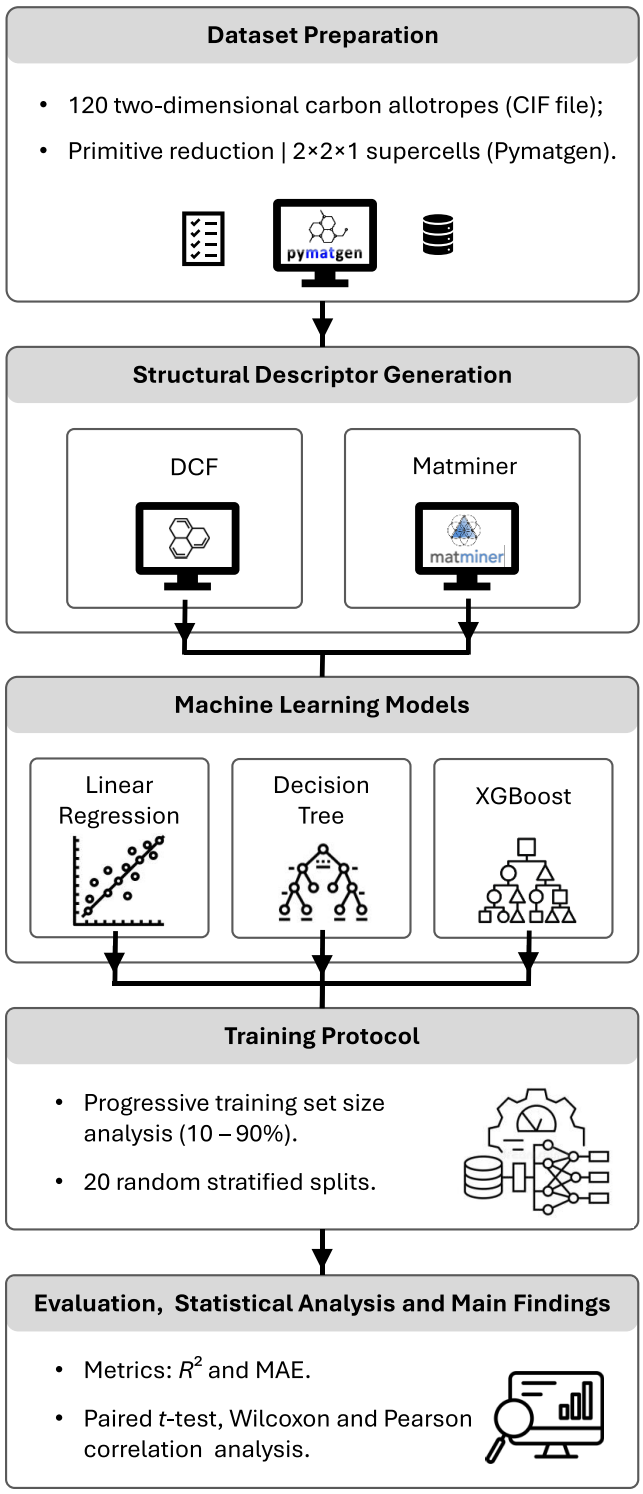


Figure 1. Schematic workflow of the computational pipeline employed in this study. After data set preparation, structural descriptors were generated using DCF and Matminer, and the models were trained using linear regression, decision trees, and XGBoost. Performance was assessed through MAE and R^2, complemented by paired statistical tests and correlation analyses.

https://doi.org/10.1021/acsomega.6c03154
ACS Omega 2026, 11, 42389−42396

formation energy reported in the literature.[39] Then, supercells of at least 2 × 2 × 1 are required for appropriate simulations, and this was consistently adopted throughout the study.[33]

Two descriptor frameworks were considered, DCF and the Matminer feature library. The DCF model is explained in detail in,[33] but its essential points can be summarized as follows. One considers the propagation of classical point particles undergoing elastic collisions within an atomic lattice supercell under periodic boundary conditions. Each trajectory propagates during a specific number of steps N_S. A single "step" corresponds to a straight line segment between successive boundary crossings. For proper averages, a total of N_L trajectories are launched. Along propagation, quantities such as traveled distances and angular deflections are recorded. The final descriptor vector is constructed from statistical analyses of angular distributions using Shannon entropy as well as Fourier decomposition of recurrence frequencies (associated with one-to-9-fold lattice symmetries), mean free paths, and recurrence times, with explicit expressions provided in.[33]

Unless otherwise explicitly stated, the standard parametrization was $N_S = 10^4$ and $N_L = 200$, yielding descriptor vectors from 25 to 30 dimensions. Larger N_S and N_L were considered for sensitivity analysis, with the fast (extended) configuration corresponding to $N_S = 10^3$ and $N_L = 100$ ($N_S = 10^4$ and $N_L = 300$). The implementation relies on NumPy, SciPy, and pandas and is available in the repository indicated in this work.

Matminer descriptors were generated using the Matminer library,[26] combining three featurizers, namely density features, the radial distribution function, and bond fractions. The structures were read from the corresponding CIF files using Pymatgen. The density descriptor was used in its default configuration, yielding the mass density, the volume per atom, and the packing fraction. The radial distribution function was discretized with a cutoff of 20.0 Å and a bin size of 0.1 Å, generating 200 binned components. Bond fractions were computed using CrystalNN as the neighbor-detection algorithm, with exact rather than approximate bond fractions and no prespecified restriction on allowed bond types. Because the bond-fraction descriptor requires the set of allowed bond types to be defined prior to feature extraction, it was fitted to the full structure set before descriptor computation. Depending on the bond-fraction contribution, which is determined by the chemistry of the data set, the resulting feature vectors contain approximately 200−500 components, thus substantially larger than those of DCF. All descriptors were normalized using z-score scaling prior to training. For a full Matminer featurized specification see Section S1 of the SI.

Three representative regression algorithms were addressed. Linear regression was implemented using ordinary least-squares from scikit learn. The decision tree model was trained with a maximum depth of 8 and a minimum sample count per leaf of 2. The XGBoost model was implemented with the hyperparameters n_estimators, max_depth, learning_rate, and subsample set to 500, 8, 0.05, and 0.8, respectively. Prior to training, all features were standardized to a mean of 0 and a variance of 1. No feature selection or dimensionality reduction was applied in order to preserve direct comparability between descriptor schemes. Fixed hyperparameters were adopted for all models, as the objective here is to compare descriptor families under identical modeling conditions rather than to optimize individual performance. But to verify that this descriptor-level comparison is not biased by the common fixed-hyperparameter setup, an independent hyperparameter optimization based on Optuna was additionally carried out for the XGBoost regressor and for both descriptor families. The full protocol, the search space, and the resulting performance comparison are reported in Section S2 of the SI.

Predictive robustness was assessed through repeated random train and test partitions, where the test fraction X_T denotes the portion of the data set assigned to the test set. We considered X_T values ranging from 0.1 to 0.9 in increments of 0.1. Each partition was repeated 20 times with random seeds to ensure statistical reliability, and reported quantities correspond to averages across all seeds. To make the statistical variability across the 20 random train and test partitions explicit, the XGBoost results are additionally reported with error bars in Section S3 of the SI, for both DCF and Matminer and for both the baseline and the Optuna optimized configurations. These extra results confirm that the discussed trends remain within the corresponding uncertainty ranges, while also making transparent the larger dispersion observed at high test fractions, particularly for R^2, as expected for a data set of limited size.

Predictive performance was quantified using the coefficient of determination R^2 and the mean absolute error (MAE), where MAE corresponds to the average absolute difference between predicted and reference values. To evaluate whether DCF and Matminer exhibit statistically distinguishable performance and to quantify their concordance across test fractions, paired t tests, Wilcoxon signed rank tests, and Pearson correlation analyses were performed on the resulting metric distributions. Statistical analyses were carried out using SciPy, adopting a significance threshold of $p < 0.05$.

RESULTS AND DISCUSSION

The predictive performance of DCF descriptors was first evaluated with respect to the simulation parameters controlling trajectory sampling. Figure 2 summarizes the dependence of the MAE on X_T for different combinations of N_S and N_L and for the three machine learning models considered in this work.

Across all models, the MAE remains stable with respect to variations in N_S and N_L. As observed in Figure 2(a), the only noticeable deviation corresponds to a single outlier near $X_T \approx 0.8$ in the linear regression case. In contrast, Figure 2(b),2(c) show tightly grouped results across the full range of X_T, with differences comparable to the expected statistical fluctuations from random train-test partitions. This behavior indicates that the DCF descriptors are robust with respect to moderate variations in trajectory length and sampling density, and that convergence is achieved without the need for extensive parameter tuning.

This stability aligns with the central concept of the DCF formulation:[33] structural information should be better captured through aggregated dynamical statistics rather than through static high-dimensional signatures. Once the trajectory sampling explores the relevant structural motifs, the resulting descriptors become statistically stable. In the present data set, increasing N_L beyond 100 does not produce systematic gains in predictive accuracy, indicating that relatively inexpensive parametrizations are already sufficient to capture the dominant structural information required for machine learning predictions.

For the following analyses, we set $N_S = 10^4$ and $N_L = 200$, representing a balanced trade-off between statistical stability and computational cost. Figure 3 compares the predictive performance obtained with DCF and Matminer descriptors

https://doi.org/10.1021/acsomega.6c03154
ACS Omega 2026, 11, 42389−42396

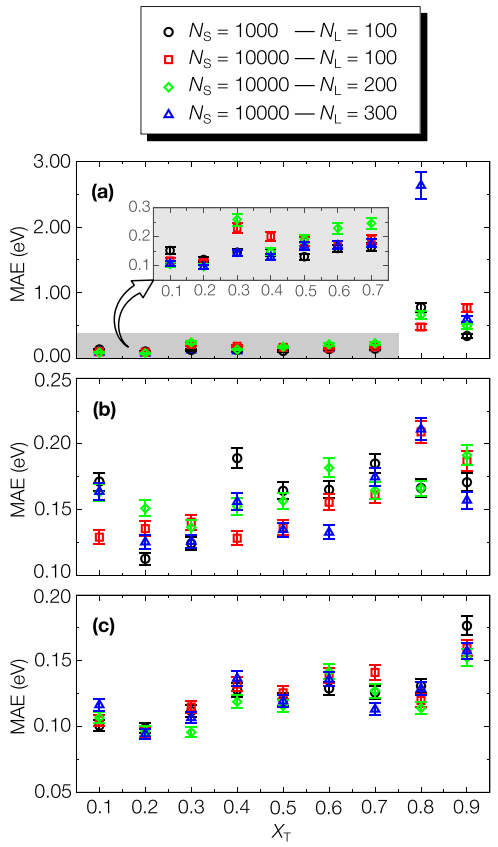


Figure 2. MAE as a function of X_T for different DCF parametrizations defined by N_S and N_L for (a) linear regression, (b) decision tree, and (c) XGBoost.

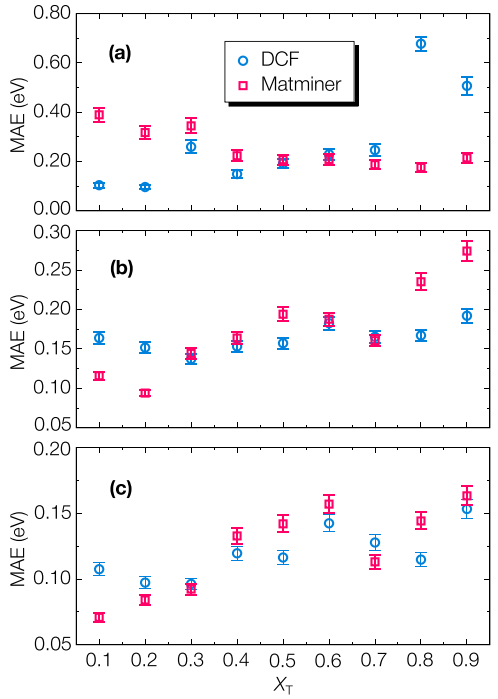


Figure 3. MAE as a function of X_T comparing DCF and Matminer descriptors for (a) linear regression, (b) decision tree, and (c) XGBoost.

through the MAE as a function of X_T for our three machine learning models. The results illustrate that DCF systematically reaches predictive errors comparable to those obtained with Matminer, despite its substantially lower dimensionality and simpler construction.

For linear regression, depicted in Figure 3(a), DCF frequently produces similar or slightly smaller MAE values

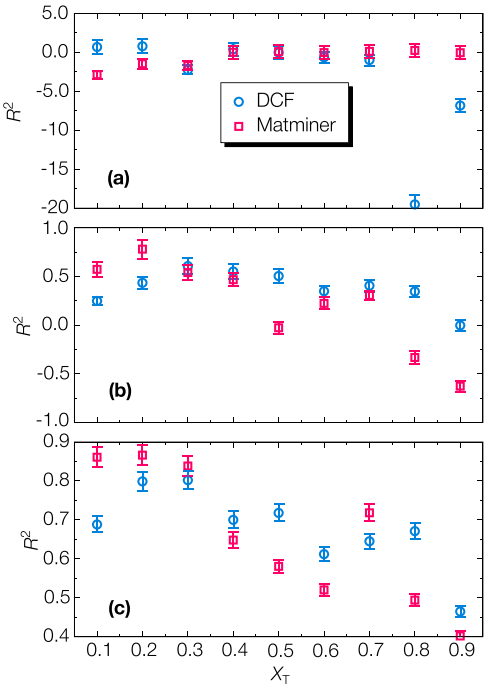


Figure 4. R^2 as a function of X_T comparing DCF and Matminer descriptors for (a) linear regression, (b) decision tree, and (c) XGBoost.

across several test fractions, with more pronounced differences emerging at larger X_T, where the reduced amount of training data amplifies sensitivity to descriptor quality. In this regime, both approaches occasionally exhibit larger fluctuations, which is expected given the limited expressive capacity of linear models for complex structure−property relationships. For decision trees, Figure 3(b) shows that the two descriptor frameworks lead to closely overlapping MAE values across most test fractions. Also, both representations provide similar levels of structural information for moderate nonlinear learning.

The comparison becomes even more consistent for XGBoost, as seen from Figure 3(c), which reveals nearly indistinguishable results between DCF and Matminer over the entire range of X_T. This indicates that the essential predictive information captured by the higher-dimensional Matminer representation is effectively retained within the compact DCF when coupled to a strong nonlinear learner.

An additional perspective on the predictive capability discussed above is obtained by analyzing the coefficient of determination, whose evolution X_T is shown in Figure 4 for DCF and Matminer descriptors. We recall that while MAE quantifies the absolute prediction error, it R^2 provides additional insight into how effectively each descriptor framework captures the variance of the target property across different train and test partitions.

In the linear regression case, Figure 4(a), both descriptor sets lead to low predictive power, with R^2 values fluctuating around zero and occasionally becoming strongly negative at

https://doi.org/10.1021/acsomega.6c03154
ACS Omega 2026, 11, 42389−42396

Table 1. Comparison between DCF and Matminer in Terms of Feature Dimensionality, Interpretability, Descriptor-Generation Cost Per Structure, and Overall Predictive Behavior[a]

descriptor set	#features	interpretability	time per structure	performance summary
DCF (Standard)	∼25−30	High (physics based)	∼4 min	Predictive accuracy comparable to Matminer for both MAE and R^2, with consistent performance for decision tree and XGBoost models.
DCF (Fast)	∼25−30	High (physics based)	∼30 s	Results remain very close to the standard configuration, with differences within the statistical variations observed in Figure 2.
Matminer	∼200−500	Moderate to low (bin based)	∼10 s	Comparable MAE and R^2 behavior relative to DCF, showing similar trends for decision tree and XGBoost models.

[a]DCF standard corresponds to $\mathbf{N}_S = 10^4$ and $\mathbf{N}_L = 200$, while DCF fast corresponds to $\mathbf{N}_S = 10^3$ and $\mathbf{N}_L = 100$. Performance summaries are consistent with the trends discussed from Figures 3 and 4. For the Matminer descriptor set, the dimensionality range depends on the selected presets, including RDF binning range and resolution, as well as optional feature subsets.

larger X_T. This trait is consistent with the limited flexibility of linear models when applied to complex, intrinsically nonlinear structure−property relationships, reinforcing the interpretation already suggested by the MAE analysis.The strongly negative R^2 values observed in the most unfavorable regimes, particularly for the linear model at large X_T, do not indicate a flaw in the evaluation procedure but rather reflect the well-known sensitivity of R^2 under small training sample conditions, where the prediction error can exceed the variance of the corresponding test sample. We should note that the interpretability advantage attributed here to DCF is intended at the descriptor-definition level and is further illustrated in Section S4 of the SI for graphene as a reference structure. Whereas the Matminer representation is dominated by a large set of distance-binned components together with a few auxiliary scalar quantities, the DCF descriptor is composed of compact variables with direct physical meaning, including free-path statistics, relaxation-time-related quantities, diffusivity, angular entropy, and symmetry-resolved intensities. The 6-fold component provides an immediate descriptor-level link to the hexagonal symmetry of graphene. Moreover, the decision tree model, Figure 4(b), indicates a clear gain in predictive power, with predominantly positive R^2 values across the tested splits. The agreement between DCF and Matminer remains strong, for the small differences observed in specific regions do X_T not alter the overall picture of comparable performance.

The aforementioned agreement becomes even stronger for XGBoost, Figure 4(c). Indeed, the plots show consistently high R^2 values for both descriptor frameworks. The close overlap between the results confirms that DCF captures essentially the same predictive information content as Matminer when combined with a sufficiently expressive nonlinear learner. Variations between individual points remain within the range expected from statistical fluctuations across different random train-test partitions.

Beyond just the crude face value of predictive metrics, it is also relevant to examine how the two descriptor protocols differ in terms of dimensionality, interpretability, and computational cost, as these aspects directly affect their practical applicability in large-scale materials informatics workflows. A central distinction concerns feature dimensionality and physical transparency. The DCF representation remains compact, typically containing approximately 25 to 30 descriptors, each one directly associated with a physical or geometric characteristic, as the already mentioned mean free path, recurrence time, angular entropy, and rotational symmetry intensities from 1 to 9-fold. In contrast, the Matminer representation contains several hundred features, largely derived from discretizing the radial distribution function across multiple distance bins (up to approximately 20 Å), complemented by density and packing statistics. While such a high-dimensional description is versatile and broadly applicable, individual components are less intuitive from a physical perspective, noting that isolated RDF bins rarely carry a clear standalone interpretation.

A second relevant aspect concerns computational cost. In the current workflow, Matminer requires approximately 10 s per structure, whereas DCF in its standard configuration ($N_S = 10^4$, $N_L = 200$) requires about 4 min per structure. Therefore, the practical utility of DCF does not stem from a uniformly lower raw generation cost compared to Matminer in standard settings. Rather, its primary advantage lies in combining descriptor compactness, physical interpretability, and competitive predictive performance, alongside a tunable cost-accuracy trade-off. In this context, the fast configuration ($N_S = 10^3$, $N_L = 100$) becomes particularly relevant. It yields MAE values remarkably close to those of extensive sampling while reducing the average wall time to roughly 30 s per structure, bringing the computational footprint to a comparable order of magnitude. Accordingly, the practical relevance of DCF in this benchmark stems less from a lower raw generation time than from its compact, physically interpretable representation and the ability to reduce computational overhead through relaxed sampling parameters without altering the descriptor-level conclusions.

Combined with the results presented in Figures 3 and 4, these observations highlight a clear trade-off between descriptor compactness, interpretability, and computational cost. Linear models remain limited across both descriptor frameworks, whereas decision tree and XGBoost models consistently achieve high predictive performance, demonstrating that DCF preserves the essential structural information despite its reduced dimensionality. A consolidated comparison summarizing dimensionality, interpretability, computational cost, and predictive behavior is presented in Table 1. To confirm that this descriptor-level comparison is not an artifact of the common fixed-hyperparameter setup, an independent Optuna-based hyperparameter optimization was performed for XGBoost on both DCF and Matminer, with the complete protocol reported in Section S2 of the SI. The optimized results introduce only minor quantitative variations and do not modify the overall conclusion that the two descriptor families display closely comparable predictive performance. The interpretability advantage attributed here to DCF is intended at the descriptor-definition level and is further illustrated in Section S4 of the SI for graphene as a reference structure. Whereas the Matminer representation is dominated by a large set of distance-binned components together with a few

https://doi.org/10.1021/acsomega.6c03154
ACS Omega 2026, 11, 42389−42396

auxiliary scalar quantities, the DCF descriptor is composed of compact variables with direct physical meaning, including free-path statistics, relaxation-time-related quantities, diffusivity, angular entropy, and symmetry-resolved intensities. The 6-fold component provides an immediate descriptor-level link to the hexagonal symmetry of graphene.

To further complement the performance trends discussed above, additional statistical analyses were conducted to evaluate whether DCF and Matminer exhibit significant differences across the full ensemble of repeated random splits. Paired statistical tests were applied to the metric distributions obtained from 20 random seeds for each X_T value, along with correlation analyses across test fractions. This provides a quantitative assessment of agreement between descriptor frameworks beyond direct visual comparison.

For linear regression, no statistically significant differences were detected between DCF and Matminer for either R^2 or MAE. Both paired *t*-tests and Wilcoxon signed-rank tests yielded nonsignificant outcomes ($p > 0.05$), indicating that the small variations observed between descriptors are compatible with statistical fluctuations rather than systematic performance differences. Correlation analyses across test fractions revealed weaker alignment than in nonlinear models, consistent with the higher variability observed in this regime, particularly at large X_T's, when both descriptor sets can produce unstable, even negative, R^2 values. For the decision tree model, paired statistical tests again indicated no relevant distinctions between DCF and Matminer for either MAE or R^2 ($p > 0.05$). Also, performance trends across test fractions remain positively correlated, pointing to similar responses of both descriptor frameworks to variations in training-set size. Regimes where performance improves or degrades, therefore, tend to coincide for DCF and Matminer, corroborating the traits observed in Figures 3(b) and 4(b). For XGBoost, paired *t*-tests and Wilcoxon tests show that DCF and Matminer are statistically indistinguishable across all evaluated metrics ($p > 0.05$), while correlation analyses across test fractions reveal strongly positive alignment for both MAE and R^2.

CONCLUSIONS

This work presented a systematic comparison of the DCF and Matminer descriptor frameworks, evaluating their predictive performance across multiple machine learning models and train and test data splits. The analysis combined direct performance curves, statistical tests, and an explicit comparison of descriptor dimensionality, interpretability, and computational cost.

Across the full set of simulations, both descriptors exhibited similar predictive limits under linear regression, in which model simplicity dominates the learning behavior and leads to unstable performance at large test fractions. In contrast, nonlinear learners consistently displayed higher accuracy, as it should be expected given the data set inherently nonlinear structure−property relationships. Notably, for decision tree and XGBoost models, DCF and Matminer converged to closely aligned performance profiles. Statistical tests confirmed the absence of significant differences between the two descriptor sets, while correlation analyses across test fractions revealed consistent predictive trends.

More generally, the present contribution comprehensively illustrates a key trade-off between descriptor compactness, interpretability, and computational cost. On one hand, Matminer offers a broad and computationally inexpensive representation but relies on a high-dimensional feature space that is less directly interpretable. On the other hand, the stability of DCF under reduced sampling parameters enables substantial reductions in computational overhead without significant loss of accuracy, reinforcing its practical relevance as a tunable framework rather than a strictly cheaper alternative. These features support the use of DCF as a physically grounded, compact approach for data-driven investigations of structure−property relationships in complex materials, particularly when model interpretability (as concretely illustrated in the Section S4 of the SI) and flexible sampling are of interest.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.6c03154.

Detailed Matminer featurizer specification (Section S1); Optuna-based hyperparameter optimization protocol, search space, and comparison against the fixed-hyperparameter baseline for XGBoost on both descriptor families (Section S2, Figures S1 and S2); statistical uncertainty across the 20 random train and test partitions for XGBoost, with and without Optuna optimization (Section S3 and Figure S3); descriptor-level interpretability example for graphene comparing Matminer and DCF representations (Section S4 and Figure S4); coordinates and lattice information for the 120 two-dimensional carbon allotropes used in the benchmark (PDF)

AUTHOR INFORMATION

Corresponding Author

Marcelo L. Pereira Junior − *Department of Electrical Engineering, College of Technology, University of Brasília, 70910-900 Brasília, Brazil;* orcid.org/0000-0001-9058-510X; Email: marcelo.lopes@unb.br

Authors

Raphael M. Tromer − *Institute of Physics, University of Brasília, 70910-900 Brasília, Brazil;* orcid.org/0000-0002-6180-5641

Isaac M. Felix − *Center for Agri-food Science and Technology, Federal University of Campina Grande, 58840-000 Pombal, Paraíba, Brazil;* orcid.org/0000-0003-0062-2195

Rafael Besse − *International Center of Physics, Institute of Physics, University of Brasília, 70910-900 Brasília, Brazil*

Marcos G. E. da Luz − *Department of Physics and Multidisciplinary Laboratory for Modeling and Analysis of Data in Complex Systems (MADComplex), Center for Scientific Modeling and Computing, Federal University of Paraná, 81531-980 Curitiba, Paraná, Brazil;* orcid.org/0000-0003-3865-2621

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.6c03154

Funding

The Article Processing Charge for the publication of this research was funded by the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES), Brazil (ROR identifier: 00x0ma614).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Council for Scientific and Technological Development (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES), and the Federal District Research Support Foundation (FAPDF). R.M.T. acknowledges CNPq for financial support (grant no. 307371/2025-5). M.L.P.J. acknowledges financial support from FAPDF (grant no. 00193-00001807/2023-16), CNPq (grant nos. 444921/2024-9 and 308222/2025-3), and CAPES (grant no. 88887.005164/2024-00). M.G.E.L. acknowledges financial support from CNPq (grant nos. 307512/2023-1 and 04577/2021-0) and CAPES (grant no. 88881.311780/2018-00).

■ REFERENCES

(1) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **2015**, *114*, No. 105503.
(2) Ludwig, A. Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. *npj Comput. Mater.* **2019**, *5*, No. 70.
(3) Maffettone, P. M.; Banko, L.; Cui, P.; Lysogorskiy, Y.; Little, M. A.; Olds, D.; Ludwig, A.; Cooper, A. I. Crystallography companion agent for high-throughput materials discovery. *Nat. Comput. Sci.* **2021**, *1*, 290−297.
(4) Ma, J.; Cao, B.; Dong, S.; Tian, Y.; Wang, M.; Xiong, J.; Sun, S. MLMD: a programming-free AI platform to predict and design materials. *npj Comput. Mater.* **2024**, *10*, No. 59.
(5) Shahzad, K.; Mardare, A. I.; Hassel, A. W. Accelerating materials discovery: combinatorial synthesis, high-throughput characterization, and computational advances. *Sci. technol. Adv. Mater.: Methods* **2024**, *4*, No. 2292486.
(6) Song, K.; Tanvir, A. N. M.; Bappy, M. O.; Zhang, Y. New Directions for Thermoelectrics: A Roadmap from High-Throughput Materials Discovery to Advanced Device Manufacturing. *Small Sci.* **2025**, *5*, No. 2300359.
(7) Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Wang, Z.; Shysheya, A.; Crabbé, J.; Ueda, S.; et al. A generative model for inorganic materials design. *Nature* **2025**, *639*, 624−632.
(8) Allen, A. E. A.; Tkatchenko, A. Machine learning of material properties: Predictive and interpretable multilinear models. *Sci. Adv.* **2022**, *8*, No. eabm7185.
(9) Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80−85.
(10) Jain, A.; Shin, Y.; Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **2016**, *1*, No. 15004.
(11) Kim, K.; Ward, L.; He, J.; Krishna, A.; Agrawal, A.; Wolverton, C. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds. *Phys. Rev. Mater.* **2018**, *2*, No. 123801.
(12) Shen, L.; Zhou, J.; Yang, T.; Yang, M.; Feng, Y. P. High-throughput computational discovery and intelligent design of two-dimensional functional materials for various applications. *Acc. Mater. Res.* **2022**, *3*, 572−583.
(13) Abroshan, H.; Kwak, H. S.; Chandrasekaran, A.; Chew, A. K.; Fonari, A.; Halls, M. D. High-throughput screening of hole transport materials for quantum dot light-emitting diodes. *Chem. Mater.* **2023**, *35*, 5059−5070.
(14) Mobarak, M. H.; Mimona, M. A.; Islam, M. A.; Hossain, N.; Zohura, F. T.; Imtiaz, I.; Rimon, M. I. H. Scope of machine learning in materials researchA review. *Appl. Surf. Sci. Adv.* **2023**, *18*, No. 100523.
(15) Li, K.; DeCost, B.; Choudhary, K.; Greenwood, M.; Hattrick-Simpers, J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *npj Comput. Mater.* **2023**, *9*, No. 55.
(16) Wang, H.; Li, K.; Ramsay, S.; Fehlis, Y.; Kim, E.; Hattrick-Simpers, J. Evaluating the performance and robustness of LLMs in materials science Q&A and property predictions. *Digital Discovery* **2025**, *4*, 1612−1624.
(17) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, No. 184115.
(18) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, No. 074106.
(19) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, No. 058301.
(20) Faber, F.; Lindmaa, A.; Von Lilienfeld, O. A.; Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094−1101.
(21) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Muller, K.-R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(22) Chadha, A.; Andreopoulos, Y. Voronoi-based compact image descriptors: Efficient region-of-interest retrieval with VLAD and deep-learning-based descriptors. *IEEE Trans. Multimedia* **2017**, *19*, 1596−1608.
(23) Huo, H.; Rupp, M. Unified representation of molecules and crystals for machine learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, No. 045017.
(24) Wang, A.; Sosso, G. C. Graph-based descriptors for condensed matter. *Phys. Rev. E* **2025**, *111*, No. 064302.
(25) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *Neural Message Passing for Quantum Chemistry*, Proceedings of the 34th International Conference on Machine Learning; PMLR, 2017; pp 1263−1272.
(26) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.
(27) Pereira, V. M.; dos Santos, J. L.; Neto, A. C. Modeling disorder in graphene. *Phys. Rev. B* **2008**, *77*, No. 115109.
(28) Hong, J.; Hu, Z.; Probert, M.; Li, K.; Lv, D.; Yang, X.; Gu, L.; Mao, N.; Feng, Q.; Xie, L.; et al. Exploring atomic defects in molybdenum disulphide monolayers. *Nat. Commun.* **2015**, *6*, No. 6293.
(29) Thomas, D. M.; Asiri, Y.; Drummond, N. Point defect formation energies in graphene from diffusion quantum Monte Carlo and density functional theory. *Phys. Rev. B* **2022**, *105*, No. 184114.
(30) Kirchhoff, A.; Deilmann, T.; Krüger, P.; Rohlfing, M. Electronic and optical properties of a hexagonal boron nitride monolayer in its pristine form and with point defects from first principles. *Phys. Rev. B* **2022**, *106*, No. 045118.
(31) Sfuncia, G.; Nicotra, G.; Giannazzo, F.; Pécz, B.; Gueorguiev, G. K.; Kakanakova-Georgieva, A. 2D graphitic-like gallium nitride and other structural selectivity in confinement at the graphene/SiC interface. *CrystEngComm* **2023**, *25*, 5810−5817.
(32) Kakanakova-Georgieva, A.; Giannazzo, F.; Nicotra, G.; Cora, I.; Gueorguiev, G. K.; Persson, P. O.; Pécz, B. Material proposal for 2D indium oxide. *Appl. Surf. Sci.* **2021**, *548*, No. 149275.
(33) Tromer, R. M. Dynamic Collision Fingerprints (DCF): Introducing a New Descriptor Linking Lattice Interactions to 2D Structural Data Signatures. *J. Chem. Theory Comput.* **2025**, *21*, 8106−8118.
(34) Drude, P. Zur elektronentheorie der metalle. *Ann. Phys.* **1902**, *312*, 687−692.

https://doi.org/10.1021/acsomega.6c03154
ACS Omega 2026, 11, 42389−42396

(35) Schneider, T. In *Statics and Dynamics of Nonlinear Systems*, Proceedings of a Workshop at the Ettore Majorana Centre; Ettore Majorana Centre: Italy, 1983; pp 212−241.
(36) Hope, T. M. *Machine learning*; Mechelli, A.; Vieira, S., Eds.; Elsevier, 2020; pp 67−81.
(37) Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81−106.
(38) Chen, T.; Guestrin, C. In *XGBoost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM Sigkdd International Conference On Knowledge Discovery And Data Mining; ACM: New York, NY, USA, 2016; pp 785−794.
(39) Shi, X.; Li, S.; Li, J.; Ouyang, T.; Zhang, C.; Tang, C.; He, C.; Zhong, J. High-throughput screening of two-dimensional planar sp^2 carbon space associated with a labeled quotient graph. *J. Phys. Chem. Lett.* **2021**, *12*, 11511−11519.
(40) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.

https://doi.org/10.1021/acsomega.6c03154
ACS Omega 2026, 11, 42389−42396