Original Research Article

Machine Learning Prediction of Mohs Hardness from Atomic Descriptors: Evaluation on an Independent Synthetic-Crystal Dataset

Srinikash Venkatasalapathi

Lynbrook High School, 1280 Johnson Ave, San Jose, CA 95129, United States

ABSTRACT

Predicting how hard a material is just from its chemical makeup is difficult, because atomic-level properties don't translate simply into large-scale mechanical behavior. This study tested four machine learning models — Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting — to evaluate how well they could predict Mohs hardness using eleven averaged atomic descriptors. The models were trained on 622 natural mineral compositions (training set) and then tested on an independent set of 52 synthetic crystals (test set) they had never seen before, which allowed evaluation of how well they generalize from natural to synthetic materials. Because Random Forest and Gradient Boosting involve randomness, each one was trained 100 times with different random seeds, and results are reported as an average with standard deviation. Random Forest performed best (RMSE = 0.94 ± 0.02; test-set R^2 = 0.62, 95% CI 0.59–0.64), cutting error by about 35% compared to Linear Regression (RMSE = 1.43, R^2 = 0.10); a paired bootstrap comparison confirmed this improvement was unlikely to be due to chance (95% CI for the RMSE difference: 0.20–0.82, which does not include zero). Gradient Boosting performed comparably well (RMSE = 1.02 ± 0.01; R^2 = 0.54 ± 0.01). A feature-removal analysis using Random Forest, repeated 10 times per descriptor, found that the atomic-number-to-mass ratio (Z/A) mattered most, followed by average ionization energy. Covalent radius turned out to matter less than earlier estimates suggested, and its effect wasn't clearly different from normal run-to-run noise once repeated trials were used. Density-related descriptors barely mattered at all. Overall, these results point to a real link between electronic and bonding-related descriptors and Mohs hardness, but the model's test-set R^2 of about 0.62 means it only accounts for roughly 62% of the variation in the test set, and it struggles more with very hard materials — showing the limits of predicting hardness from composition alone.

Keywords: Mohs hardness; machine learning; Random Forest; Gradient Boosting; atomic descriptors; feature importance; materials informatics

Corresponding author: Srinikash Venkatasalapathi, E-mail: srinikash.venki@gmail.com.

Accepted July 16, 2026
https://doi.org/10.70251/HYJR2348.44408415

INTRODUCTION

Hardness is an important material property that reflects resistance to surface indentation, scratching, and wear (1). In structural engineering, for instance, hardness (along with strength) determines whether a building can hold up under heavy loads, wind, and earthquakes.

Hardness also matters in aerospace and automotive engineering, where it affects how long parts last, how well they resist wear, and how efficient they are. The problem is that hardness data exists for a lot of materials, but actually testing every candidate material by hand takes a long time and often isn't practical. That's why predictive models that estimate Mohs hardness from easier-to-get compositional information are useful.

Machine learning can pick up on complicated, non-linear relationships between composition and hardness that simple rule-of-thumb physics models tend to miss (2). Earlier research found that atomic-level properties like atomic radius, valence electron count, and electronegativity are correlated with hardness (3). That suggests a model trained on these properties could approximate the relationship between structure and hardness without needing detailed structural data.

This study tests four models — Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting — for predicting the Mohs hardness of minerals using eleven averaged atomic descriptors (Section 2.2). The models were trained only on natural mineral compositions, then tested on a separate set of synthetic crystals that were not used in training. That setup lets us see whether a model trained on natural minerals can generalize synthetic materials. We also ran a feature-removal analysis to find out which descriptors matter most for predictions. The goal is to identify which atomic properties are most closely tied to Mohs hardness, and to evaluate how well machine learning models trained only on compositional data actually perform in practice.

METHODS AND MATERIALS

Data Sources

The training data came from 622 natural mineral compositions with measured Mohs hardness values (range 1.0–10.0; mean 4.61, median 5.5, SD 1.73). This dataset was originally put together by Garnett, using the Physical and Optical Properties of Minerals CRC Handbook and the American Mineralogist Crystal Structure Database (4). For testing, a separate set of 52 synthetic single crystals with their own measured hardness values (range 2.5–9.0; mean 5.45, median 5.5, SD 1.52) was used, covering seven crystal systems (monoclinic, n = 15; trigonal, n = 9; tetragonal, n = 7; hexagonal, n = 7; orthorhombic, n = 6; cubic, n = 4; rhombohedral, n = 3; unspecified, n = 1), also from the same source (4). These synthetic crystals were kept separate, they were never used for training or for picking hyperparameters, only for the final evaluation. The cleaned datasets and analysis code are posted in the author's public repository (5) (accessed on 2026-07-08).

The two datasets were harmonized before descriptor calculation. Chemical formulas from both sources were parsed and standardized to a common element-and-stoichiometry format (e.g., consistent handling of hydrate notation and site-occupancy fractions) so that the same parser could be applied without source-specific exceptions. The eleven descriptors in Table 1 were then computed with an identical script for both the training and test compositions, using the same reference table of elemental properties, so no dataset-specific processing was applied at any stage. Beyond the duplicate check described below, the only inclusion criterion was that a compound have both a complete, parseable chemical formula and a reported Mohs hardness value; no other filtering was applied to either dataset.

While double-checking the data for this revision, 23 of the 622 training rows (3.7%) were found to be exact duplicates of another row — same descriptor values and the same hardness label. Another 36 rows have identical descriptor values but different hardness labels. This is likely a byproduct of how the original dataset was built: minerals with the same name but different locality or minor composition differences appear to have been listed as separate rows (4). These rows were not removed, so results remain consistent with those reported previously, but this duplication is a real limitation (see Section 4.4) that future work should fix, either by removing duplicates or by explicitly modeling how hardness can vary within the same composition.

Atomic Descriptors

Both datasets describe each material with the same eleven compositional descriptors (Table 1). These are built from elemental properties, either averaged or summed across all the atoms in the formula. These descriptors were selected because they can be calculated directly from a chemical formula — no crystal-structure data needed — so they work even for materials that have not been structurally characterized yet. Density-related descriptors were included on the premise that denser materials resist deformation better (6), though Section 3.4 shows this hypothesis was not strongly supported by the data.

Data Preprocessing

Every model used all eleven descriptors from Table 1, and none were dropped before training. Features

Table 1. *Atomic descriptors used as input features for Mohs hardness prediction. Definitions and hypothesized relationships with hardness are provided for all 11 descriptors.*

Descriptor	Description	Hypothesized relationship to hardness
allelectrons_Total	Total number of electrons added up across all atoms in the composition	More electrons might mean a denser, heavier composition — and possibly more hardness
density_Total	Sum of elemental densities across all atoms in the composition	Denser compositions might resist deformation better
allelectrons_Average	Average number of electrons per atom	Higher average electron counts are linked to metallic bonding, which can make materials softer
val_e_Average	Average number of valence electrons per atom	More valence electrons could mean stronger covalent or ionic bonds
atomicweight_Average	Average atomic mass	Heavier atoms might mean a more stable lattice, though this link is expected to be weak
ionenergy_Average	Average first ionization energy	Higher ionization energy means electrons are held more tightly, which could mean stronger bonding and more hardness
el_neg_chi_Average	Average electronegativity	Higher electronegativity might mean stronger, more polar bonds and more hardness
R_vdw_element_Average	Average van der Waals radius (unbonded)	Bigger atoms might mean more space between atoms and less hardness
R_cov_element_Average	Average covalent radius (bonded)	Bigger bonded radii might mean longer, weaker bonds and less hardness
zaratio_Average	Average ratio of atomic number to atomic mass (Z/A)	A higher ratio means more positive charge per unit mass, which could mean stronger electrostatic attraction and more hardness
density_Average	Average elemental density per atom	Denser compositions might resist deformation better

were scaled with a min–max scaler that was fit only on the 622-mineral training set, then applied to the 52-crystal test set, so no test-set information leaked into preprocessing. No rows had missing values for these descriptors, so none needed to be dropped.

Model Training and Hyperparameters

Four regression models were implemented in Python 3 using scikit-learn 1.8.0, pandas 3.0.2, and NumPy 2.4.4. Linear Regression was fitted using ordinary least squares with the default scikit-learn settings. Ridge Regression was implemented with an L2 penalty and ; this value was selected without cross-validation. Random Forest was fitted using the default configuration of 100 trees, no maximum depth, and a minimum of two samples per split. Gradient Boosting was also implemented with default settings, including 100 boosting stages, a learning rate of 0.1, and a maximum tree depth of 3. No formal hyperparameter tuning was performed for either ensemble model. No separate validation set or cross-validation procedure was used before evaluation on the independent synthetic-crystal test set. Therefore, all reported performance metrics were derived from a single external test set, and the model and hyperparameter choices were not independently validated. This limitation is discussed further in Section 4.4.

Linear Regression and Ridge Regression give the same result every time, since they don't involve any randomness. Random Forest and Gradient Boosting do involve randomness, because they randomly sample features and data points while building trees. To capture how much that randomness matters, each model was retrained 100 times with seeds 0–99, and the average and standard deviation across those runs are reported.

Evaluation Procedure

Performance was measured with root-mean-square error (RMSE) and R^2 on the 52-crystal test set. For

Random Forest and Gradient Boosting, RMSE and R^2 were calculated for each of the 100 runs, and the mean, standard deviation, and 95% interval across those runs are reported. To directly compare Random Forest with Linear Regression, a paired bootstrap was used: the 52 test points were resampled (with replacement) 10,000 times and RMSE was recalculated for both models each time — for Random Forest, using the average prediction across its 100 runs. The resulting spread of RMSE differences was used to build a 95% confidence interval and assess whether Random Forest's edge over Linear Regression held up consistently.

Feature Importance (Ablation) Analysis

To identify which descriptors mattered most, an ablation test was run on the Random Forest model: for each of the eleven descriptors, it was removed, the model was retrained on the remaining ten, and the test RMSE was recorded. This was repeated 10 times per descriptor with different random seeds, and Section 3.4 reports the mean and standard deviation of that RMSE, along with how much it changed compared to the full-feature-set baseline (also averaged over 10 runs).

RESULTS

Feature Correlations

Figure 1 shows how the eleven descriptors correlate with each other and with Mohs hardness in the training set. Some pairs are strongly correlated — especially atomicweight_Average with density_Average and zaratio_Average, since all three depend on atomic mass, and allelectrons_Average with several ionization- and bonding-related descriptors. But no single descriptor correlates strongly with hardness on its own (all $|r| \leq 0.28$). That fits with hardness depending on a mix of descriptors together, something nonlinear models (Section 3.2) can capture better than any one variable alone.

Model Performance Comparison

Table 2 shows how all four models did on the synthetic-crystal test set. Linear Regression got an RMSE of 1.43 and R^2 of 0.098 — a purely linear combination of the eleven descriptors only explains a small part of what drives hardness. Ridge Regression did slightly worse (RMSE = 1.47, R^2 = 0.05), which suggests the L2 penalty didn't help generalization, and that overfitting wasn't really the linear model's main problem.

Random Forest had the best average performance across its 100 runs (RMSE = 0.94 ± 0.02; R^2 = 0.616, 95% CI 0.592–0.637), clearly beating both linear models. A paired bootstrap comparison backed this up: the 95%

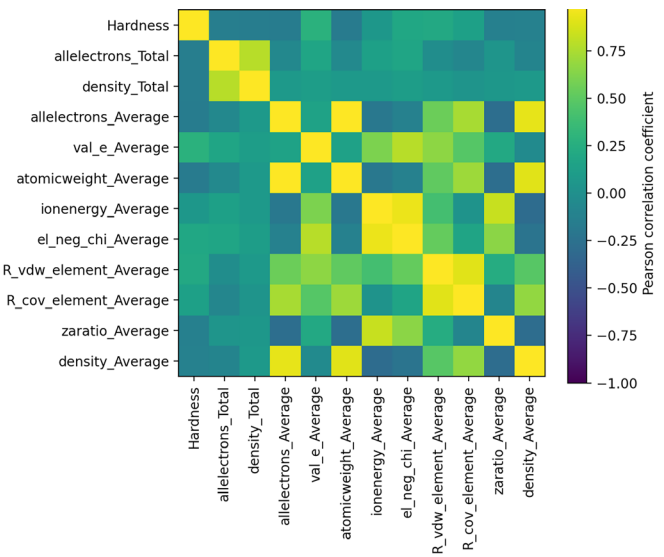

Figure 1. *Pearson correlation matrix for the eleven atomic descriptors and Mohs hardness, training set (N = 622).*

Table 2. Performance of four regression models on the independent synthetic-crystal test set (N = 52). *Lower RMSE and higher indicate better performance. Linear Regression and Ridge Regression are deterministic and are reported as single values. Random Forest and Gradient Boosting results are reported as mean ± SD across 100 training runs using random seeds 0–99.*

Model	RMSE (mean ± SD)	R^2 (mean ± SD)	95% CI for R^2
Linear Regression	1.43[a]	0.098[a]	—
Ridge Regression (α = 0.01)	1.47[a]	0.050[a]	—
Random Forest	0.94 ± 0.02[b]	0.616 ± 0.013[b]	0.592–0.637
Gradient Boosting	1.02 ± 0.01[b]	0.539 ± 0.005[b]	0.531–0.549

confidence interval for the RMSE difference between Linear Regression and Random Forest was (0.20, 0.82), which doesn't include zero, and not one of the 10,000 bootstrap resamples favored Linear Regression. Gradient Boosting also beat the linear models (RMSE = 1.02 ± 0.01; $R^2 = 0.539 \pm 0.005$), though it did not reach Random Forest's level.

Overall, this shows the relationship between these atomic descriptors and Mohs hardness is quite nonlinear, and tree-based ensemble methods handled it better than linear approaches for this descriptor set and dataset size.

Predicted versus Actual Hardness

Figure 2 shows how Mohs hardness is distributed in the training and test sets. The training set is right-skewed, with a clear peak around Mohs 5.5–6 and not many examples above Mohs 7. The test set is spread more evenly, but it still only has eight materials above Mohs 6.5.

Figure 3 plots Random Forest's predicted hardness (averaged across its 100 runs) against the actual hardness for the test set, with a reference line where predictions would exactly match reality. For materials in the middle of the hardness range (roughly Mohs 4–6.5), predictions cluster close to that line, with a mean absolute error of 0.55–0.66. But for the hardest materials in the test set (Mohs > 6.5, n = 8), the model tends to underpredict, with a mean absolute error of 1.36. This is probably because the training set has so few high-hardness minerals (Figure 2a), not because the modeling approach itself has some fundamental flaw — though with only eight high-hardness test examples, this remains uncertain.

Feature Importance

Table 3 and Figure 4 show how much Random Forest's test RMSE changed when each descriptor was removed, compared to the full-feature-set baseline of 0.938 ± 0.017 (mean ± SD across 10 seeded runs). Removing the atomic-number-to-mass ratio (zaratio_Average) caused the biggest jump in RMSE (+0.084, about five baseline standard deviations), making it the single most important descriptor here. Average ionization energy came next

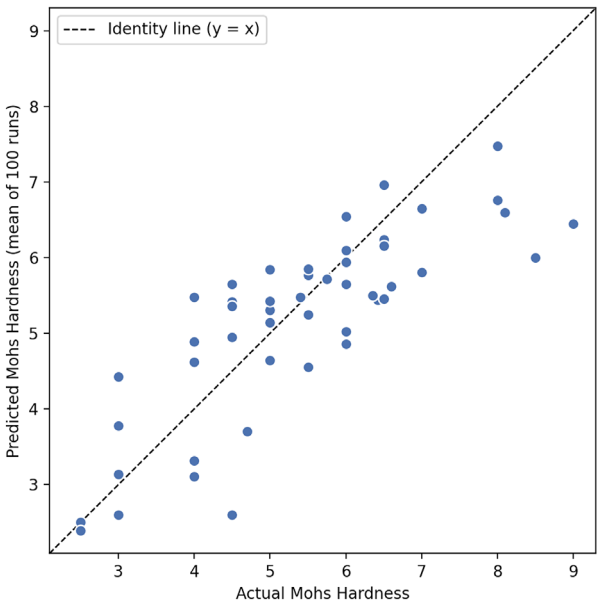

Figure 3. *Predicted vs. actual Mohs hardness for Random Forest on the synthetic-crystal test set (N = 52), with the identity line (y = x) shown for reference.*

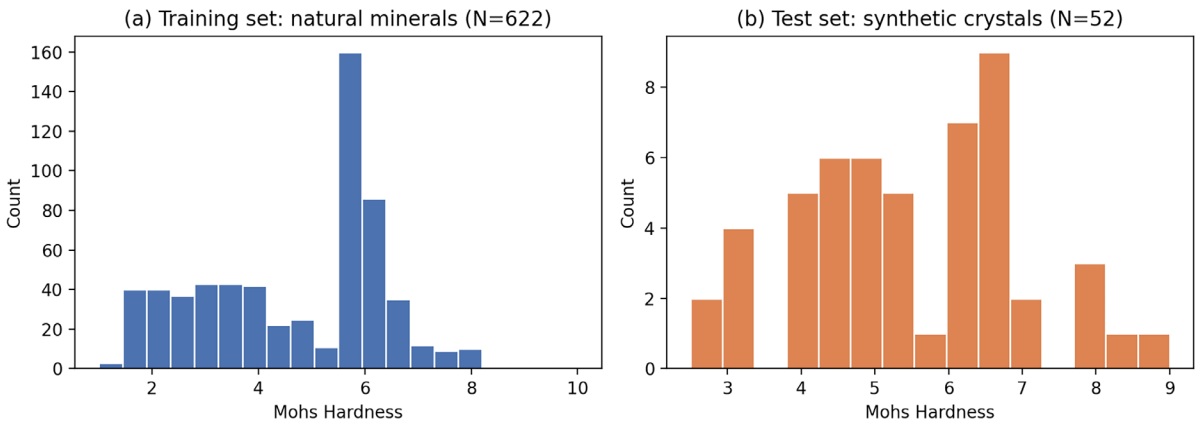

Figure 2. *Distribution of Mohs hardness in (a) the training set of natural minerals (N = 622) and (b) the test set of synthetic crystals (N = 52).*

(+0.061, about 3.6 baseline standard deviations). Average valence electron count had a smaller effect that's closer to the noise floor (+0.041, about 2.4 baseline standard deviations).

Table 3. Random Forest ablation results after removing each descriptor individually. *Test RMSE values are reported as mean ± SD across 10 seeded runs. ΔRMSE is the difference between the mean RMSE after descriptor removal and the full-feature baseline RMSE of 0.938. Positive ΔRMSE values indicate reduced model performance. Shaded rows represent changes exceeding approximately three baseline standard deviations (0.051).*

Descriptor removed	Test RMSE after removal (mean ± SD)[c]	Δ RMSE from full-feature baseline[d]
allelectrons_Total	0.943 ± 0.015	+0.005
density_Total	0.935 ± 0.014	−0.003
allelectrons_Average	0.932 ± 0.015	−0.006
val_e_Average	0.979 ± 0.023	+0.041
atomicweight_Average	0.933 ± 0.019	−0.005
ionenergy_Average	1.000 ± 0.016	+0.061
el_neg_chi_Average	0.946 ± 0.021	+0.008
R_vdw_element_Average	0.923 ± 0.013	−0.016
R_cov_element_Average	0.961 ± 0.009	+0.022
zaratio_Average	1.022 ± 0.015	+0.084
density_Average	0.930 ± 0.010	−0.008

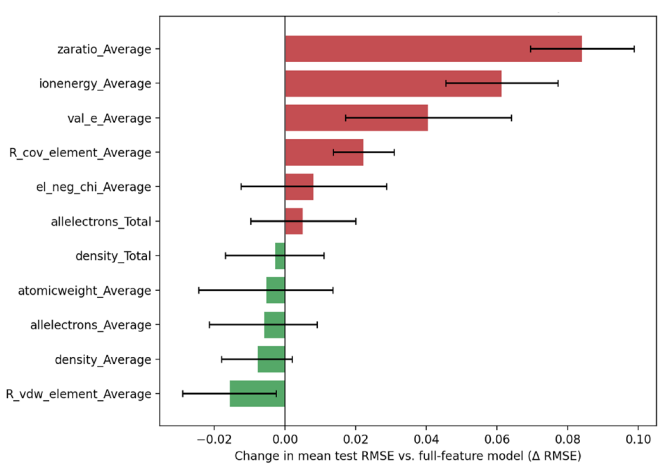

Figure 4. *Change in Random Forest test RMSE upon removing each descriptor, relative to the full-feature-set baseline (mean ± SD, 10 repeats per descriptor).*

Covalent radius (R_cov_element_Average) had the second-largest RMSE increase (+0.06) in the original single-run analysis, but once averaged over repeated trials, its effect shrank and became harder to distinguish from noise (+0.022, about 1.3 baseline standard deviations). So it shouldn't be treated as a solidly top-ranked descriptor. The density-related descriptors (density_Total, density_Average) and van der Waals radius barely changed RMSE at all, or even made it slightly better when removed — meaning they weren't adding much unique information once the other descriptors were already in the model.

DISCUSSION

Interpretation

The results show that Random Forest and Gradient Boosting, both nonlinear ensemble methods, clearly beat the linear models at predicting Mohs hardness from averaged atomic descriptors — and that advantage held up when moving from natural minerals to a separately measured set of synthetic crystals. The ablation analysis suggests electronic and bonding-related descriptors, especially the atomic-number-to-mass ratio and average ionization energy, carried more useful information about hardness than the macroscopic density descriptors did in this dataset.

It's worth being clear about what these findings actually mean: they describe a statistical link between descriptors and predicted hardness inside the trained model, not a proven explanation of what causes hardness at the atomic level. The descriptors here are just compositional averages — they don't capture crystal structure, bond topology, coordination number, defects, phase, or anisotropy, all of which are known to affect real hardness. Two materials could have identical average atomic descriptors but different measured hardness, just because their atoms are arranged differently. Our own training data backs this up: as noted in Section 2.1, dozens of rows share identical descriptor values but have different hardness labels. So claims about which physical mechanism drives hardness should be read as hypotheses coming from the model's feature-importance ranking, not as settled conclusions.

Model Performance and Its Limits

Random Forest's R^2 of about 0.62 is a big step up from the linear baselines, but it still leaves roughly 38% of the test-set variance unexplained. Performance also wasn't even across the hardness range: error was about

twice as large for the eight hardest synthetic crystals (Mohs > 6.5) as for materials in the middle of the range. This is probably because there just aren't many high-hardness examples in the training data (Figure 2a), rather than some basic flaw in the modeling approach — but it does mean the model shouldn't be trusted for high-hardness materials until it's tested on a bigger sample.

Domain Shift Between Natural Minerals and Synthetic Crystals

The training and test sets are systematically different: natural minerals can have impurities, structural defects, and compositional variation that synthetic crystals grown under controlled conditions don't have, and the two sets also have somewhat different hardness distributions (training mean 4.61 vs. test mean 5.45). Since our descriptors just average over composition and miss these structural differences, some of the unexplained variance in the test set might come from this mismatch between natural and synthetic materials, not just from the descriptors or model themselves.

Limitations

This study has several limitations. First, we didn't use cross-validation or a held-out validation set during training — all the metrics here come from one external test set of 52 crystals, so our hyperparameter and model choices weren't checked independently of the final results. Second, we left Random Forest and Gradient Boosting at their scikit-learn default settings instead of doing a formal hyperparameter search, so the numbers reported here might not reflect each algorithm's best possible performance on this task. Third, the training data contains descriptor-level duplicates (Section 2.1), and we didn't explicitly correct for how that might affect training. Fourth, our descriptors are purely compositional and leave out structural, crystallographic, and defect information that's known to affect hardness. Fifth, the test set is small (N = 52) and skewed toward mid-range hardness values, which limits how precise our performance estimates are, especially for high-hardness materials. Finally, we only measured feature importance through single-descriptor ablation with Random Forest. Since some descriptors are correlated with each other (Figure 1), this approach can undersell how much any one descriptor in a correlated group really matters — so the ranking in Section 3.4 should be treated as a rough guide rather than the final word.

CONCLUSION

This study tested four machine learning models for predicting the Mohs hardness of minerals from atomic descriptors, training on 622 natural mineral compositions and testing on a separate set of 52 synthetic crystals. Random Forest performed best (RMSE = 0.94 ± 0.02, R^2 = 0.62), beating Linear Regression and Ridge Regression by a margin that a paired bootstrap comparison showed was unlikely to be chance, and Gradient Boosting performed comparably well. The feature-ablation analysis pointed to the atomic-number-to-mass ratio and average ionization energy as the descriptors most reliably tied to predictive performance, while density-related descriptors barely mattered — suggesting that, at least in this dataset, hardness is more closely tied to electronic and bonding-related features than to macroscopic density.

These results show that machine learning can predict Mohs hardness reasonably well from easy-to-calculate compositional descriptors, but the limits are real: the model only explains about 62% of test-set variance, it's less reliable for high-hardness materials, and it relies on descriptors that can't tell apart materials with the same composition but different structures.

Future work should focus on: (1) bigger, more balanced training datasets, especially for high-hardness materials (Mohs > 7); (2) adding structural and crystallographic descriptors (like space group, coordination number, and bond topology) alongside the compositional ones; (3) using nested cross-validation for hyperparameter selection, so model selection and final evaluation don't get mixed up; (4) formally comparing results across several independent test sets instead of just one 52-crystal set; and (5) trying complementary, structure-sensitive approaches — like convolutional neural networks applied to mineral or crystal images — for cases where compositional data alone isn't enough.

CONFLICT OF INTEREST

The author declares no conflicts of interest related to this work.

REFERENCES

1. National Park Service. Mohs hardness scale (Internet). 2023. Available from: https://www.nps.gov/articles/mohs-hardness-scale.htm (accessed on 2026-07-08).
2. Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TYJ. Explainable machine learning in

materials science. npj Comput Mater. 2022 Sep 22;8(1):204. doi:10.1038/s41524-022-00884-7

3. Dong X, *et al*. Electronegativity and chemical hardness of elements under pressure. *Proc Natl Acad Sci U S A*. 2022 Mar; 119 (10): e2117416119. doi:10.1073/pnas.2117416119
4. Garnett JC. Prediction of Mohs hardness with machine learning methods using compositional features. In: Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions. ACS Symposium Series, vol. 1326. *Washington (DC): American Chemical Society*; 2019. doi:10.1021/bk-2019-1326.ch002
5. Venkatasalapathi S. Predicting-Mohs-Hardness-of-Minerals-Using-Machine-Learning-on-Atomic-Descriptors (Internet). GitHub. Available from: https://github.com/HashtagOfficialAccount/Predicting-Mohs-Hardness-of-Minerals-Using-Machine-Learning-on-Atomic-Descriptors (accessed on 2026-07-08).
6. Algeos. Density measurement charts for materials (Internet). 2025. Available from: https://www.algeos.com/density-measurement (accessed on 2025-12-23).
7. Scikit-learn developers. sklearn.linear_model.Ridge — scikit-learn documentation (Internet). Date unknown. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html (accessed on 2026-07-08).