OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

# Composition-Only Screening of Battery Materials: A Reproducible Baseline Using the Stability –Voltage–Capacity Score

Raghav Kapre
raghavkapre@gmail.com

## ABSTRACT

Identifying new host materials for lithium-ion batteries that offer high voltage, capacity, and long-term stability remains a central challenge in materials science, and most existing machine-learning screening workflows rely on crystal-structure descriptors or deep-learning models that are computationally expensive and difficult to reproduce. This study asks how much predictive power can be recovered from composition-level information alone. From 6,683 materials in the publicly available Materials Project battery dataset, a Stability–Voltage–Capacity (SVC) composite score is defined as the equal-weighted average of min-max-normalised stability, average voltage, and gravimetric capacity, and materials at or above the 90th percentile of SVC are labelled as top-decile candidates. Two composition-only feature sets — a 25-feature manual set of elemental statistics and the 132-feature MAGPIE preset from matminer — are used to train four binary classifiers (logistic regression and random forest crossed with each feature set) across 20 stratified 80/20 splits. The random-forest–MAGPIE pairing was the strongest configuration, with test-set AUROC of 0.878 (95% CI: 0.847–0.897); the other three combinations reached mean AUROC values between 0.805 and 0.866. A supplementary group-safe evaluation, in which every row sharing an identical composition-feature vector was kept together, gave a comparable random-forest–MAGPIE AUROC of 0.871 (95% CI: 0.809–0.914). Composition-only descriptors therefore provide a defensible, reproducible baseline for screening within the Materials-Project DFT-labelled corpus at laptop scale.

## INTRODUCTION

Throughout the world, there has been a transition towards renewable energy that has greatly increased demand for high-performance rechargeable batteries. Lithium-ion batteries, which use the insertion and extraction of lithium ions into a host material dominate the market for both portable electronics and battery electric vehicles.[1,2] The biggest challenge however, is identifying new viable host materials which offer high voltage, capacity and long term stability. The paperby Goodenough and park clearly outlines these requirements which have shaped research on new materials since the advent of Lithium-ion cells.[3]

Traditionally, materials discovery has relied on the costly and time consuming techniques of experimental synthesis and characterization. However, since the advent of large, open-source computational databases,

like the Materials Project, which 'provides a well-tested set of structure and thermodynamic analyses' for thousands of compounds,[4] it has become possible to screen potential compounds before experimentation begins.

One of the key measures of material stability is the Energy above the convex hull (E_hull). It is a geometric construction that defines the set of thermodynamically stable phases for a given chemical system.[5] If a material possesses E_hull = 0, it lies on the hull and is predicted to be stable, while a material which possesses an E_hull value greater than 0 may spontaneously decompose, with a higher value suggesting a higher instability. The Materials Project computes E_hull using Density Functional Theory (DFT),[6] corrected for known systematic errors in exchange-correlation functionals.[7]

While this data is widely available, most publicly available published machine learning screening workflows for battery batteries mainly use structural descriptors - these features are derived from the crystal structure of a material - and deep learning models, which require exorbitant amounts of computational resources.[8–10] While these resources are powerful, they are difficult to reproduce, especially in scenarios where there is an extreme resource constraint.

To close this gap, this study builds a screening pipeline that relies solely on compositional information, using the publicly available MP-API battery dataset,[11,12] with full reproducibility as a central design goal. Instead of pulling features from the crystal structure of a material, we derive a set of elemental statistics directly from its chemical formula — quantities like the mean electronegativity, atomic radius, and atomic mass, each averaged across the elements present in the compound. From these composition-level features, we construct a single composite metric, the Stability–Voltage–Capacity (SVC) score, calculated as an equal-weighted average of three normalised quantities: stability, average voltage, and gravimetric capacity. The screening problem is then posed as one of binary classification — using nothing but this composition-derived information, can a simple supervised classifier pick out the materials that land in the top decile of SVC score?

For classification, we evaluate two standard models — logistic regression and random forest — and use AUROC as our primary evaluation metric. Beyond raw predictive performance, we test how robust our results are to changes in the weighting scheme used to construct the SVC score, and we identify which composition-level descriptors carry the most predictive signal for high-performing candidates. The broader aim of this work is to show that open data, combined with laptop-scale computation, can yield a meaningful and interpretable baseline for battery material screening — one that complements, rather than competes with, more computationally intensive approaches.

## METHODS

### Dataset

The research made use of the battery dataset from the Materials Project,[4] obtained through a public Kaggle copy Battery Dataset From MP-API,[11] which included the chemical formulas and the measured electrochemical properties for a large number of potential battery materials. Each material had a separate record containing its composition as well as its thermodynamic stability, average voltage, and gravimetric capacity. Before being compiled, the dataset was cleaned by eliminating any entry that had a missing value for stability, voltage, or capacity, as these three properties were the foundation of the composite score mentioned below. As a result of this cleaning process, 6,683 materials were left and were then used in all the subsequent analysis.

### The Composite Score and the Binary Label

Since stability, voltage, and capacity are on different scales and therefore cannot be meaningfully combined in their original form, a single composite metric was created in order to rank the materials according to their overall desirability. To start, each of the three properties was rescaled to a common 0–1 range using maximum-minimum normalisation, a process which rescales any set of numbers so the smallest becomes 0 and the largest becomes 1, the intermediate values being scaled in proportion to this. Stability, in this case given as energy above the convex hull, was handled differently from the other two properties since a lower value means that the material is more stable; therefore, before averaging, the normalised stability score was subtracted from 1 to make sure that a higher value always indicated a more desirable material in all three aspects. The Stability–Voltage–Capacity (SVC) score was then defined as:

$$SVC = [(1 - \text{normalised stability}) + \text{normalised voltage} + \text{normalised capacity}] / 3$$

The same weight was given to each of the three terms, since there is no agreement in the literature on battery materials regarding which property should be given priority; a later section of the paper looks at how sensitive the results are to this assumption. Those materials which had an SVC score at or above the 90th percentile (that is, $SVC \geq 0.451$) were classified as top-decile candidates (label = 1), and all the other materials were given a label of 0, resulting in 669 positive and 6,014 negative examples.

### Feature Engineering

From each material's chemical formula[13] two distinct feature sets were created. The manually derived feature set included 25 features, which were obtained by calculating five summary statistics—namely the mean, standard deviation, minimum, maximum, and range—for the five elemental properties (atomic mass, electronegativity, periodic table row, periodic table group, and atomic radius) across the elements in

each compound. The other feature set was generated by using the MAGPIE preset from the matminer library's ElementProperty featurizer,[14,15] this producing 132 statistics covering 22 elemental properties per compound and making it a commonly used set of compositional descriptors in computational materials science.[16]

**Model Training and Evaluation**

Four binary classifiers were trained for each of the two different sets of features (the 25 manual features and the 132 MAGPIE features) and pairs of model families (logistic regression versus a random forest).[17] The data was stratified sampled to create an approximately 9:1 ratio between negative to positive examples and then split into an 80% training set and a 20% held-out test set. The random seed was kept constant at 42 along the pipeline to ensure that every split, initialisation and every downstream evaluation is exactly the same. Both logistic regression models used a scaler fitted on the training partition, and then applied to the test partition that would not contaminate the fitting process by using information from the held-out partition. All 200 decision trees and the scikit-learn defaults for depth and split criterion[18] were used in both models of the random forest.

Two ranking-based metrics were used to evaluate the classifiers. The Area Under the Receiver Operating Characteristic Curve (AUROC) is a number that measures how reliably a model can rank a randomly selected positive material higher than a randomly selected negative one, with 0.5 indicating a totally random ranker and 1.0 indicating a completely perfect one. The Area Under the Precision–Recall Curve (PR-AUC) is the same ranking quality but it is more informative when there is a class imbalance because it is based on the ratio of true positives to false positives in the part of the ranking that would be first screened by a screening pipeline. Both measures were done on the unseen test set.

To quantify uncertainty on the reported metrics, the stratified 80/20 train–test split was repeated for 20 different random seeds (integer seeds 42 through 61). For each seed, all four classifiers were retrained from scratch and re-scored on the corresponding held-out test partition; the mean, standard deviation, and 2.5th/97.5th percentile 95% confidence interval of AUROC and PR-AUC were then reported across the 20 runs. Because the manual and MAGPIE feature sets are elemental-statistic summaries of composition rather than of stoichiometry, many rows in the dataset produce identical feature vectors, and a purely row-level stratified split can place near-duplicates on both sides of the split. A complementary group-safe protocol was therefore also run, in which every row sharing an identical manual-feature fingerprint (rounded to six decimals) was assigned to a single partition using GroupShuffleSplit; the resulting positive-class rate in the test partition was chosen from 10 candidate group-splits per seed by minimising the deviation from the overall 10% positive rate. This group-safe evaluation removes the dominant source of composition-only leakage while retaining the same ~80/20 partition sizes.

### Duplicate-Composition Audit and Sensitivity Analysis

Two additional checks were run to characterise the study's scope. First, each row's host-composition string (the tail chemical formula produced by the pipeline's parser) and its manual-feature vector were compared across the standard train and test partitions to quantify the extent to which the composition-only setup permits near-duplicate leakage. Second, a sensitivity analysis was carried out over the SVC construction itself. Four weighting schemes — the equal-weight baseline (1/3, 1/3, 1/3) and three variants that up-weight stability, voltage, or capacity to 0.50 with the other two properties down-weighted to 0.25 — were combined with three percentile thresholds (85th, 90th, 95th) to produce twelve alternative labellings. For each, the RF + MAGPIE model was refit and evaluated at seed 42 under both the stratified and the group-safe splitting protocols.

### Feature Importance and Ablation

The built-in mean decrease in impurity (MDI) score was extracted from each tree ensemble to find the most important features used by each random forest. MDI was calculated as the mean of the values obtained across all splits and all trees for each model (RF + Manual and RF + MAGPIE), and measures the reduction in class impurity caused by a given feature.

To determine the most important features for the best-performing classifier an additional test-set permutation-importance analysis was conducted. The AUROC was calculated for each of the 132 MAGPIE features in turn, as the values of that column were shuffled across the set of test rows, with the decrease in AUROC being noted; a large drop suggests that the feature put up a significant amount of predictive signal. Based on this ranking, an ablation experiment was conducted by retraining the RF + MAGPIE model with only the top-k features, where k was chosen to be in the range 1 to 132. This procedure is used to determine the minimal number of features needed to get back the full-model AUROC.

## RESULTS

### SVC Score Distribution and Class Balance

The composite scores for SVC had a moderately right skewed distribution with a mode around 0.43; the data for the 6,683 materials used was spread fairly evenly from 0.35 to 0.50 (Figure 1). A short right tail is beyond 0.55, indicating the small number of materials that possess exceptionally good voltage, stability and capacity at the same time. When applying the 90th percentile at SVC = 0.451 a highly imbalanced binary label was created with 6,014 materials in the majority class, and 669 in a minority class, the top

10% (ratio of approximately 9:1) (Figure 2). The main reason for choosing AUROC and PR-AUC as primary evaluation metrics is because both of these remain interpretable under this degree of imbalance, while threshold-based metrics like raw accuracy would be influenced by the majority class.

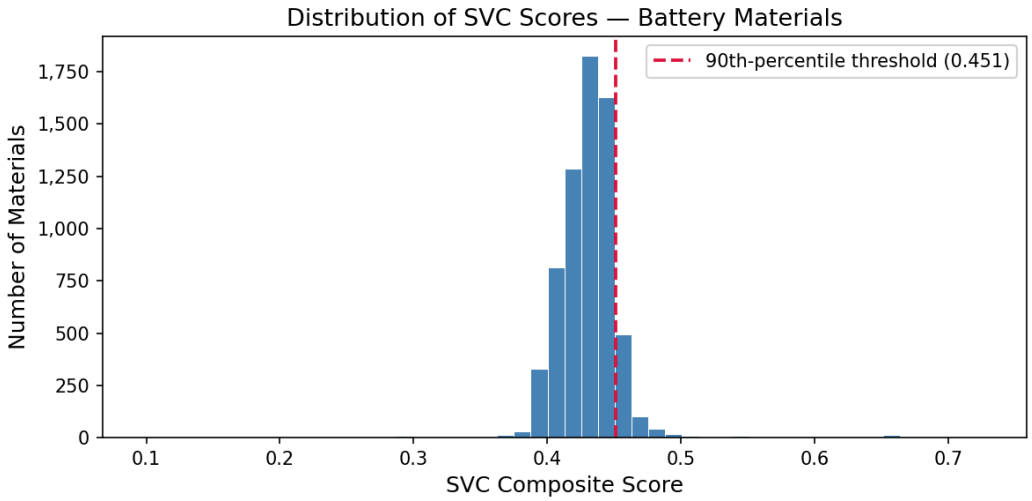



*Figure 1. Distribution of the Stability–Voltage–Capacity (SVC) composite score across 6,683 Materials Project battery entries. The dashed line marks the 90th-percentile threshold (SVC = 0.451) used to define the top-decile label.*

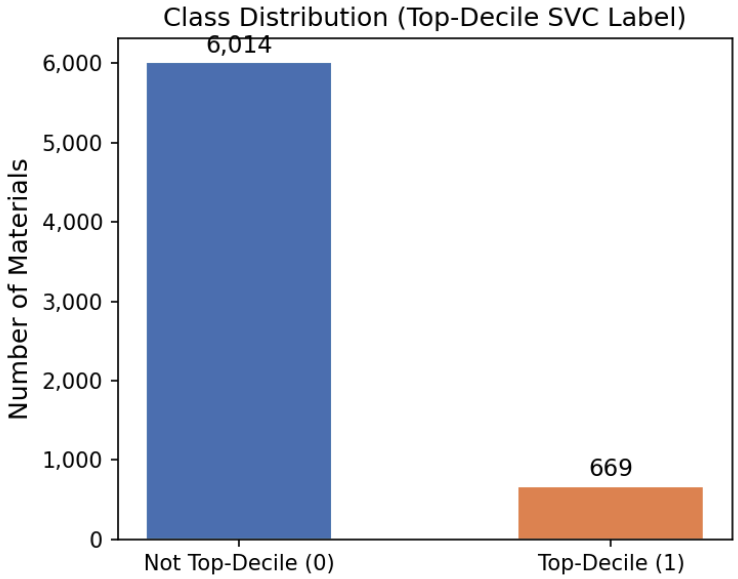


*Figure 2. Class balance of the top-decile SVC binary label: 6,014 negative (label = 0) versus 669 positive (label = 1), a ratio of roughly 9:1.*

**Feature Correlation with SVC Score**

To provide a descriptive screen of the signal available in the two composition feature sets, the absolute Spearman rank correlation was calculated between each composition feature and the SVC score prior to any model training. Periodic-table-row statistics (max_row, mean_row, range_row and max_atomic_mass, mean_atomic_mass, range_atomic_mass) had the largest values for |r| (0.41 to 0.47) of the 25 manual features, followed closely by the atomic-mass statistics (std_atomic_mass, max_atomic_mass and range_atomic_mass and mean_atomic_mass) (Figure 3). The same qualitative trend can be seen for the MAGPIE set, with the mean of the covalent radius, MagpieData, and the mean of the covalent radius, CovalentRadius, in the rankings at |r| ≈ 0.53, and the average deviation of atomic weight and atomic number trailing slightly (Figure 4). The two sets of features shared a common underlying signal: the atomic size/position in the periodic table.

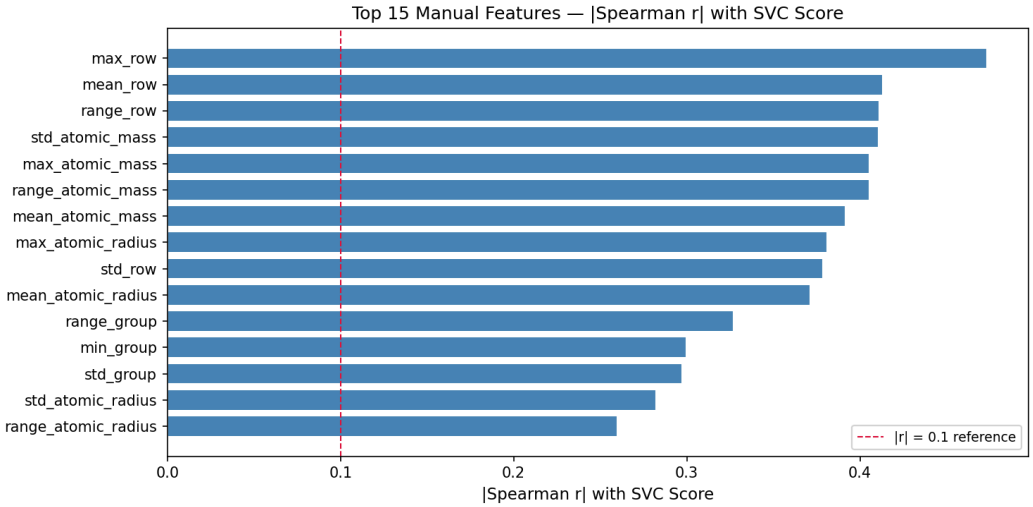



*Figure 3. Top 15 manual composition features ranked by absolute Spearman correlation with the SVC score. The dashed line indicates |r| = 0.1 as a reference.*

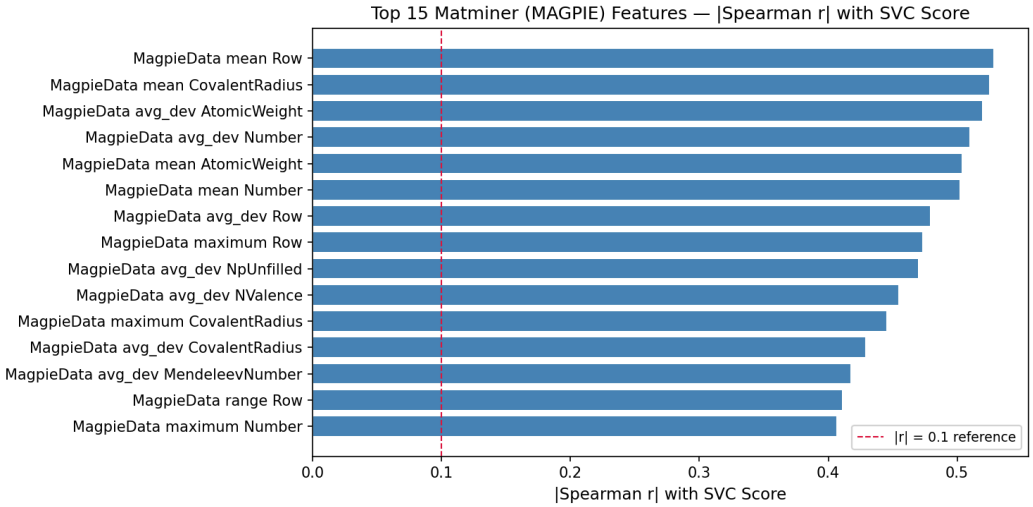



*Figure 4. Top 15 MAGPIE composition features ranked by absolute Spearman correlation with the SVC score.*

**Model Performance**

The four classifiers achieved significantly better performance than the random baseline (AUROC = 0.5) on the test set that was not used for training. The four models were LR + Manual (0.817), RF + Manual (0.854), LR + MAGPIE (0.863) and RF + MAGPIE (0.872), ranked in order of ascending AUROC, which is shown as the ROC curve in Figure 5. Table 1 provides a complete set of metrics.

| **Model** | **Features** | **Mean AUROC (95% CI)** | **Mean PR-AUC (95% CI)** |
|---|---|---|---|
| Logistic Regression | Manual (25) | 0.805 (0.781–0.829) | 0.401 (0.353–0.458) |
| Logistic Regression | MAGPIE (132) | 0.865 (0.846–0.887) | 0.471 (0.416–0.534) |
| Random Forest | Manual (25) | 0.843 (0.817–0.868) | 0.426 (0.387–0.492) |
| Random Forest | MAGPIE (132) | 0.878 (0.847–0.897) | 0.494 (0.435–0.543) |

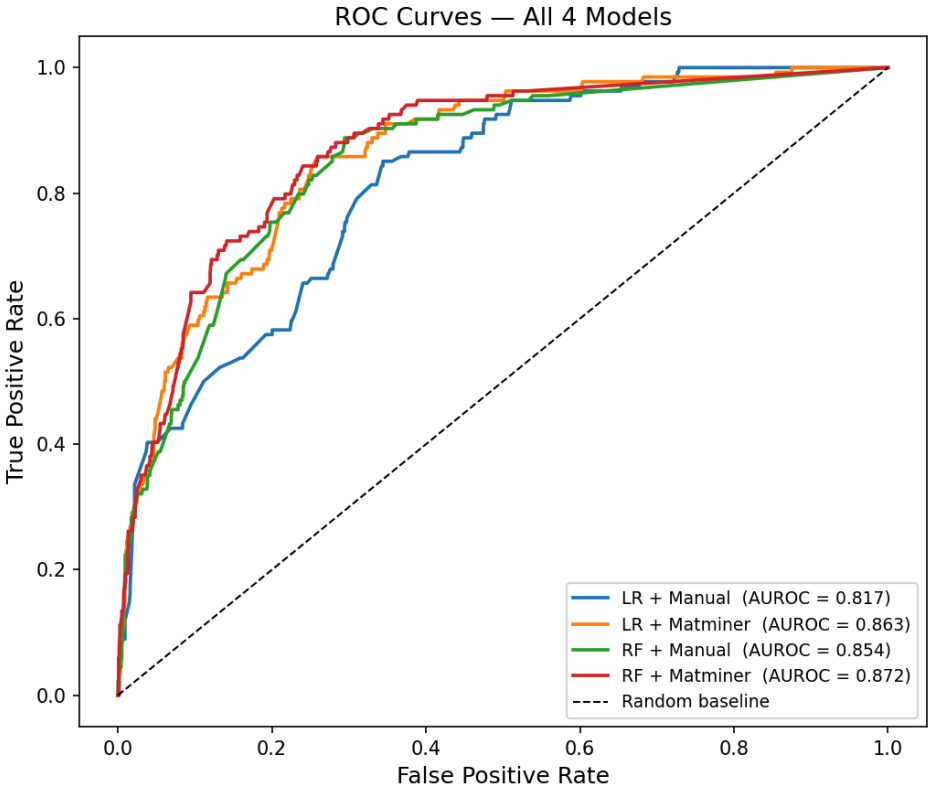



*Figure 5. ROC curves for all four composition-only classifiers on the held-out test set (n = 1,337). AUROC values appear in the legend, with a diagonal random baseline for reference.*

*Table 1. Multi-seed test-set performance of the four composition-only classifiers (n = 20 stratified 80/20 splits, seeds 42–61). Values are means with 2.5–97.5 percentile 95% confidence intervals.*

Table 1 shows that there are two trends. First, the AUROC improvement for the MAGPIE feature set over the manual set (25 features) was modest but significant for both of the model families, ranging from 0.02 to 0.05. Secondly, the random forest was better than the logistic regression for the fixed feature set, suggesting that the class SVC is not linearly separable in either feature space. The model of interest for the feature-importance and ablation analyses described below is RF + MAGPIE, which was the best overall model.

**Duplicate-Composition Audit and Group-Safe Evaluation**

The composition-only setup carries a substantial amount of within-corpus redundancy. Of the 6,683 rows retained after cleaning, only 3,123 host formulas were distinct, and 4,649 rows (69.6%) shared a host

formula with at least one other row; the ten most common host formulas alone (MnO2, TiO2, VO2, NiO2, CrO2, FeO2, CoO2, MnBO3, MnPO4, FePO4) accounted for 489 rows. In the baseline stratified 80/20 split at seed 42, 890 of 1,337 test rows (66.6%) had a host formula that also appeared in the training partition, and 1,134 rows (84.8%) had a manual-feature vector that exactly matched at least one training row. A further 178 host formulas mapped to both label values across the corpus, involving 1,553 rows. This means the row-level stratified split evaluates the classifier in part on rows whose composition-based features it has already seen during training. The group-safe protocol described in Methods eliminates the identical-feature-vector variety of this leakage by design. Under the group-safe protocol, the RF + MAGPIE mean AUROC was 0.871 (95% CI: 0.809–0.914), compared with 0.878 (95% CI: 0.847–0.897) under the stratified protocol; the mean AUROC ordering across the four classifiers was preserved, but the confidence intervals under the group-safe protocol were substantially wider (standard deviations of approximately 0.03 versus approximately 0.015), reflecting the extra variance introduced when nearly-identical rows are no longer implicitly shared.

**SVC Weight and Threshold Sensitivity**

The RF + MAGPIE model was refit at seed 42 for every combination of the four weighting schemes and the three percentile thresholds described in Methods. Under the stratified split, AUROC ranged from 0.840 to 0.910 across the twelve variants; under the group-safe split, from 0.832 to 0.952. Every variant remained above the 0.75 target, and no ordering was reversed under either protocol. The full sensitivity results are reported in Table 2. The variant that most changed the AUROC was the 0.25/0.50/0.25 (voltage-weighted) scheme at the 95th percentile, which drove the stratified AUROC down to 0.840 but left the group-safe AUROC at 0.908; the capacity-weighted scheme at the same threshold showed the opposite pattern (0.898 vs. 0.952). The headline claim — that composition-only classifiers reach test-set AUROC of roughly 0.85–0.90 for top-decile SVC identification in the Materials-Project battery corpus — is therefore robust to reasonable perturbations of both the weighting scheme and the threshold percentile.

| **Weights (S/V/C)** | **Threshold pctile** | **SVC threshold** | **n positives** | **AUROC — stratified** | **AUROC — group-safe** |
|---|---|---|---|---|---|
| 1/3, 1/3, 1/3 (baseline) | 85 | 0.448 | 1003 | 0.883 | 0.878 |
| 1/3, 1/3, 1/3 (baseline) | 90 | 0.451 | 669 | 0.872 | 0.907 |
| 1/3, 1/3, 1/3 (baseline) | 95 | 0.458 | 335 | 0.896 | 0.921 |
| 0.50 S / 0.25 V / 0.25 C | 85 | 0.583 | 1003 | 0.878 | 0.926 |

| | | | | | |
|---|---|---|---|---|---|
| 0.50 S / 0.25 V / 0.25 C | 90 | 0.585 | 669 | 0.882 | 0.915 |
| 0.50 S / 0.25 V / 0.25 C | 95 | 0.591 | 335 | 0.904 | 0.927 |
| 0.25 S / 0.50 V / 0.25 C | 85 | 0.408 | 1003 | 0.903 | 0.892 |
| 0.25 S / 0.50 V / 0.25 C | 90 | 0.412 | 669 | 0.910 | 0.901 |
| 0.25 S / 0.50 V / 0.25 C | 95 | 0.417 | 335 | 0.840 | 0.908 |
| 0.25 S / 0.25 V / 0.50 C | 85 | 0.354 | 1003 | 0.879 | 0.847 |
| 0.25 S / 0.25 V / 0.50 C | 90 | 0.360 | 669 | 0.891 | 0.832 |
| 0.25 S / 0.25 V / 0.50 C | 95 | 0.372 | 335 | 0.898 | 0.952 |

*Table 2. SVC weight and threshold sensitivity for the RF + MAGPIE model at seed 42. Weighting schemes give the coefficients on min-max-normalised stability (S), voltage (V) and capacity (C); threshold-percentile columns give the percentile cut on the SVC distribution used to define the top-decile label.*

**Feature Importance**

The most influential predictor for the 25 manual features trained random forest was std_electronegativity, with an MDI score of ~0.115, followed by range_atomic_mass and std_atomic_mass. In general, the top ranked are not based on the magnitude of an individual element, but on the diversity of the elements of the compound (electronegativity and atomic mass) (Figure 6).

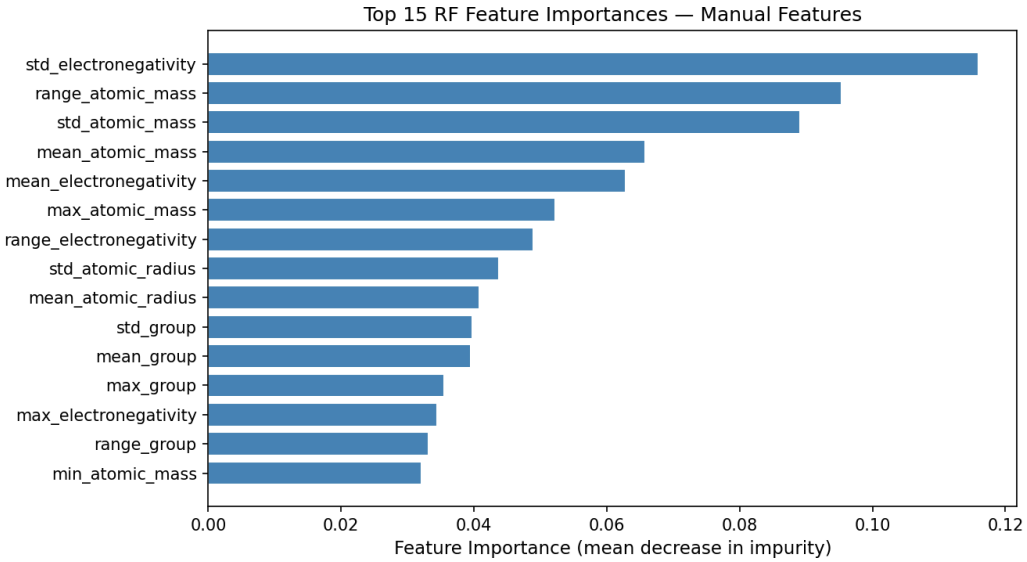



*Figure 6. Top 15 features by mean-decrease-in-impurity (MDI) importance from the random forest trained on the 25 manual composition features.*

This was the case again, though in greater detail, for the random forest trained on the MAGPIE features. First of all, MagpieData avg_dev was ranked first, followed by MagpieData avg_dev GSvolume_pa and MagpieData mean CovalentRadius (Figure 7). The dominant MDI signals were based on the compositional diversity, in terms of the variations of unfilled p-orbital electrons and also the variations of the ground state atomic volume of the constituent elements, and a mean-covalent-radius term that reflects the overall size of the atoms.

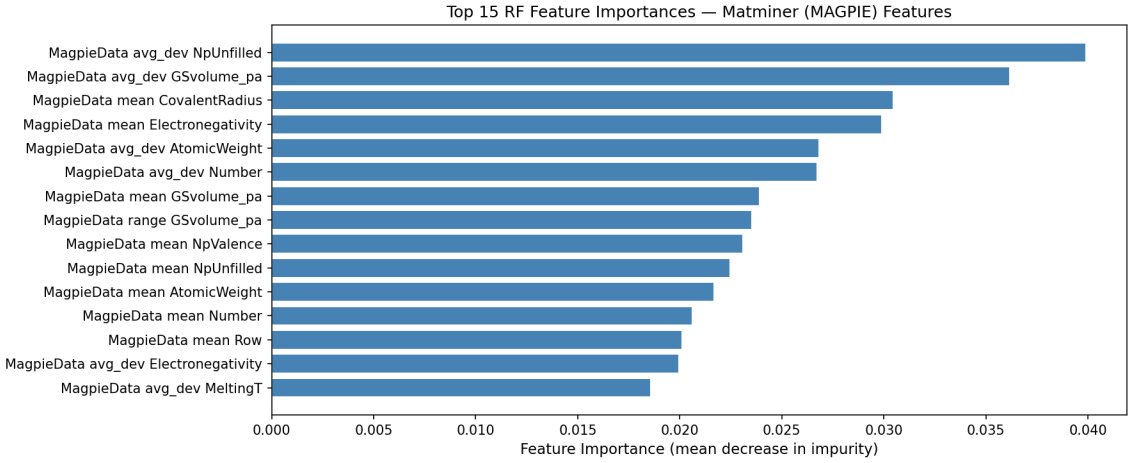


*Figure 7. Top 15 features by MDI importance from the random forest trained on the 132 MAGPIE composition features.*

**Ablation**

This ablation experiment confirmed that the RF + MAGPIE model was able to regain almost all its predictive power from only a handful of the 132 features that it was provided (Figure 8). AUROC with k = 1 to 3 features was around 0.70, significantly higher than chance but not as good as the full model. An elbow was seen at k = 5, with AUROC increasing to 0.864 (which is 0.01 lower than the value of AUROC in the full model of 0.872). After 7 features, there was a negligible improvement in AUROC as it increased gradually with each added feature up to 132, ranging from 0.868 to 0.885 inclusive. The single feature most significantly impacted by the random shuffling of its value was MagpieData maximum Number, which is essentially the atomic number of the heaviest element in the compound, whose shuffling led to an average decrease of just about 0.009 in test-set AUROC (Figure 9), significantly larger than the decreases produced by shuffling any of the other individual features. The results as a whole indicate that the signal in the MAGPIE feature set is contained in just a few period-row elements and heavy elements, not evenly distributed across the 132-dimensional space. The ablation reported in this subsection is exploratory rather than confirmatory: the same held-out test set is used both to rank the features by permutation importance and to evaluate the reduced models at each k. The k = 5 elbow and the reported AUROC values therefore describe the ranking–evaluation loop on this single partition, not the expected AUROC of a five-feature model on a fresh held-out sample.

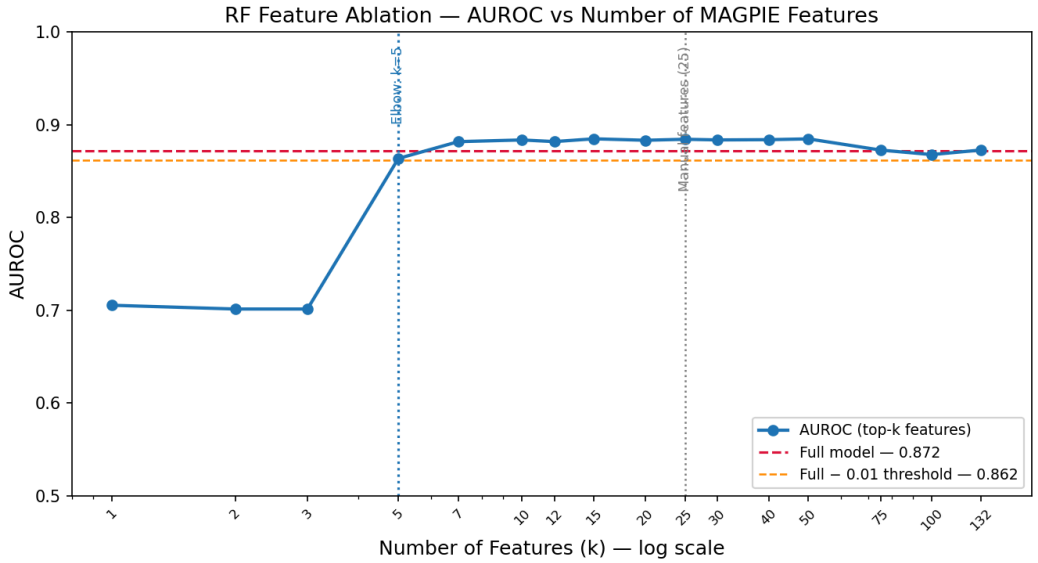


*Figure 8. Ablation experiment: test-set AUROC of the RF + MAGPIE classifier as a function of the number of top-ranked features (k) retained, from k = 1 to k = 132.*

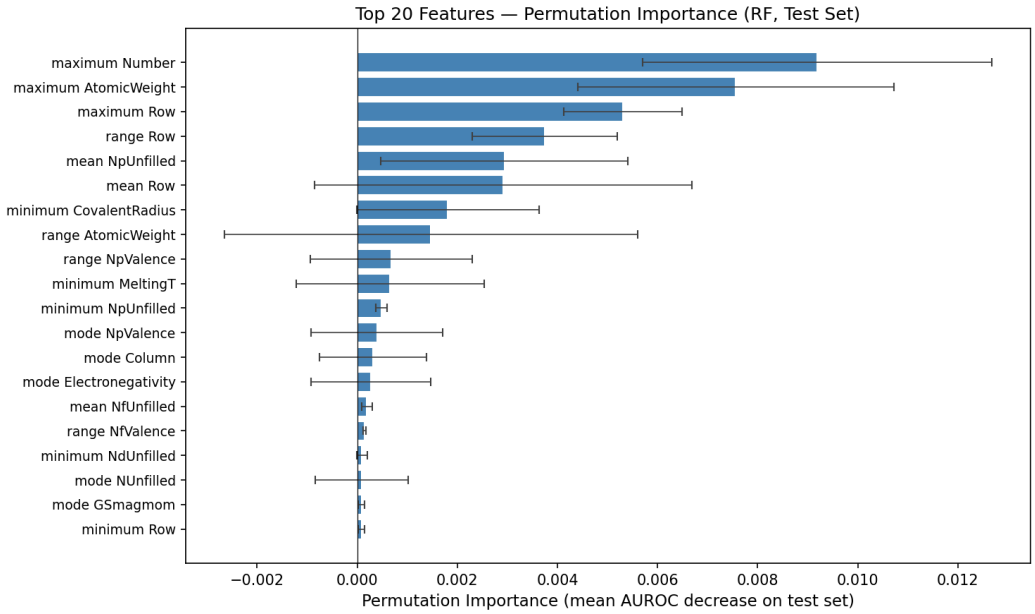



*Figure 9. Top 20 MAGPIE features by test-set permutation importance for the RF + MAGPIE classifier, measured as the mean AUROC decrease when each feature's values are randomly shuffled.*

## DISCUSSION

The findings above suggest three results that will be explored individually in light of the framing provided in the Introduction.

The first is that the composition of a battery material in the Materials Project dataset alone contains a significant fraction of the signal required to identify high-SVC materials. None of the four classifiers was ever trained with a crystal structure, a lattice parameter, or a DFT calculated band structure, but all had an AUROC value between 0.82 and 0.87. This aligns with the larger literature on composition-based-learning, where models constructed from elemental-statistic descriptors have been found to retain a meaningful amount of the predictive power that is attainable with structure-aware representations for various thermodynamic and electronic properties.[15,16,19] For the purpose of battery screening, this implies that a composition-only filter may be a first-pass screen filter to narrow down the

amount of candidates by about an order of magnitude or more before using more costly structure-based methods.

The second is about the additional value of a more extensive feature set. The transition from the 25 feature manual descriptor to the MAGPIE descriptor with 132 features didn't result in a significant AUROC gain, ranging from 0.02 to 0.05 across model families. The ablation experiment reinforces this finding: RF + MAGPIE is only using 5 out of 132 features to achieve within 0.01 of its full AUROC, and only adds approximately 10 more features beyond that. Two implications follow. First, the composition-level signal for this specific target is reflected mostly by a few dozen descriptors that are already reasonably well captured in the manual feature set, and are related to variables such as periodic-table position, atomic mass, and atomic size. Second, the fact that the gains of increasing the number of features were diminishing indicates that richer target definitions or a few structure-derived features may yield additional gains on this problem, rather than more features at the composition level.

The third observation is that the correlation screen, the MDI's ranking and the permutation importance results are in agreement with a consistent chemical story. All of the following measures scale, at least in some manner, with atomic size and with the position of an element in the periodic table: periodic-table row, covalent radius, atomic weight, unfilled p-orbital count, and ground-state volume. This stability over three independent methods, a rank correlation on the entire database, an impurity-based importance during training and a permutation-based importance on the test set, is reassuring and is in line with the intuition that the stability, voltage and capacity of insertion-type battery materials are affected by the size and electronic characteristics of the host atoms.

It is important to highlight some of the constraints before the results are extrapolated. The SVC score is not a community metric, and the equal weights across stability, voltage and capacity was not optimal but rather chosen for interpretability; using a sensitivity analysis with varying weights across these three categories would be a natural extension of this study. The 90th percentile threshold is also a modelling decision and both metrics (AUROC and PR-AUC) are threshold independent, but the threshold is not. The Spearman correlation reported in the Feature Correlation section was computed on all data, not just on the training partition; the results should therefore be interpreted as a descriptive process, and not as a formal process of feature-dropping, as no feature was actually dropped based on these correlations and no information from the test partition was fed into the trained models. The uncertainty for the reported AUROC values was not reported in this iteration of the pipeline, but adding a bootstrapping of the test set would allow 95% confidence intervals to be attached to each of the four metrics and is recommended for future work. Finally, all the data comes from the DFT-computed properties and contains any systematic errors in the exchange-correlation functionals used and in the energy corrections used for the calculations in the Materials Project7,20; this performance is not indicative of the performance for experimental measurements, but how well composition-only classifiers reproduce the DFT computational label.

Three further constraints, prompted by the review, should be foregrounded before the results are generalised. First, the two composition-only feature sets are elemental-statistic summaries that ignore stoichiometric ratios: 84.8% of test rows in the baseline stratified split had a manual-feature vector already present in the training partition, and 69.6% of all rows shared a host formula with at least one other row. This corpus-level redundancy means that the row-level stratified AUROC values reflect performance on materials whose composition signature the classifier has largely seen during training, not on genuinely unseen compositions; the accompanying group-safe evaluation was run precisely to remove this leakage, and although the mean AUROC values held up (RF + MAGPIE 0.871 vs. 0.878 under the stratified protocol), the confidence intervals under the group-safe split were roughly twice as wide. Second, the ablation experiment reported above is exploratory: because the permutation importance and the reduced-model AUROC were both measured on the same test set, the $k = 5$ elbow describes the current partition rather than the AUROC to expect from a five-feature model on new data. A confirmatory analysis would require an independent held-out partition dedicated to reduced-model scoring. Third, the reported metrics therefore describe a within-corpus, DFT-labelled screening baseline and should not be extrapolated to unseen materials without further validation on chemically distinct compositions.

Under these constraints, the findings are consistent with a bounded and reproducible claim: composition-only descriptors, when used in conjunction with well-established supervised classifiers, offer a defensible basis for battery-material screening at laptop scale, and the signal lies in a few, chemically interpretable subset of descriptors.

## CONCLUSION

This study established that composition-only descriptors, extracted from the chemical formulas of 6,683 Materials Project battery entries, can identify top-decile Stability–Voltage–Capacity candidates with a mean test-set AUROC of 0.878 (95% CI: 0.847–0.897) under a stratified 80/20 split and 0.871 (95% CI: 0.809–0.914) under a group-safe split that removes identical-feature-vector duplicates, using a standard random-forest classifier on laptop-scale hardware. Across four weighting schemes and three percentile thresholds for the SVC construction, the classifier's AUROC remained in the 0.83–0.95 range under both splitting protocols, indicating that the headline result is robust to reasonable perturbations of the label definition. Taken together, the findings support a bounded claim: a composition-only classifier is a defensible first-pass sieve for the DFT-labelled Materials-Project corpus and can reduce the candidate pool by roughly an order of magnitude before more expensive structure-based methods are applied. The results should not be extrapolated to unseen materials without additional validation; the exploratory ablation described above should likewise be confirmed on an independent held-out partition. Future work will address these gaps and test whether adding a small number of structure-derived features closes the remaining performance gap to structure-aware benchmarks.

## REFERENCES


1. 1. Nitta, N., Wu, F., Lee, J. T., & Yushin, G. (2015). Li-ion battery materials: Present and future. *Materials Today*, 18(5), 252–264. https://doi.org/10.1016/j.mattod.2014.10.040
2. 2. Manthiram, A. (2020). A reflection on lithium-ion battery cathode chemistry. *Nature Communications*, 11, 1550. https://doi.org/10.1038/s41467-020-15355-0
3. 3. Goodenough, J. B., & Park, K.-S. (2013). The Li-ion rechargeable battery: A perspective. *Journal of the American Chemical Society*, 135(4), 1167–1176. https://doi.org/10.1021/ja3091438
4. 4. Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K. A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), 011002. https://doi.org/10.1063/1.4812323
5. 5. Bartel, C. J. (2022). Review of computational approaches to predict the thermodynamic stability of inorganic solids. *Journal of Materials Science*, 57, 10475–10498. https://doi.org/10.1007/s10853-022-06915-4
6. 6. Kohn, W., & Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A), A1133–A1138. https://doi.org/10.1103/PhysRev.140.A1133
7. 7. Wang, A., Kingsbury, R., McDermott, M., Horton, M., Jain, A., Ong, S. P., Dwaraknath, S., & Persson, K. A. (2021). A framework for quantifying uncertainty in DFT energy corrections. *Scientific Reports*, 11, 15496. https://doi.org/10.1038/s41598-021-94550-5
8. 8. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715), 547–555. https://doi.org/10.1038/s41586-018-0337-2
9. 9. Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301
10. 10. Sendek, A. D., Cubuk, E. D., Antoniuk, E. R., Cheon, G., Cui, Y., & Reed, E. J. (2019). Machine learning-assisted discovery of solid Li-ion conducting materials. *Chemistry of Materials*, 31(2), 342–352. https://doi.org/10.1021/acs.chemmater.8b03272
11. 11. Hasandafa. (2024). Battery Dataset from MP-API [Data set]. *Kaggle*. https://www.kaggle.com/datasets/hasandafa1201/battery-dataset-from-mp-api
12. 12. Materials Project. (n.d.). Battery Explorer documentation. Retrieved from https://next-gen.materialsproject.org/batteries
13. 13. Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., & Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68, 314–319. https://doi.org/10.1016/j.commatsci.2012.10.028

14. 14. Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E. R., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., Chard, K., Asta, M., Persson, K. A., Snyder, G. J., Foster, I., & Jain, A. (2018). Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152, 60–69. https://doi.org/10.1016/j.commatsci.2018.05.018
15. 15. Ward, L., Agrawal, A., Choudhary, A., & Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2, 16028. https://doi.org/10.1038/npjcompumats.2016.28
16. 16. Meredig, B., Agrawal, A., Kirklin, S., Saal, J. E., Doak, J. W., Thompson, A., Zhang, K., Choudhary, A., & Wolverton, C. (2014). Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*, 89(9), 094104. https://doi.org/10.1103/PhysRevB.89.094104
17. 17. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
18. 18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
19. 19. Zhuo, Y., Mansouri Tehrani, A., & Brgoch, J. (2018). Predicting the band gaps of inorganic solids by machine learning. *Journal of Physical Chemistry Letters*, 9(7), 1668–1673. https://doi.org/10.1021/acs.jpclett.8b00124
20. 20. Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Rühl, S., & Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1, 15010. https://doi.org/10.1038/npjcompumats.2015.10