Supporting information

Data collection

A dataset has been constructed to predict the specific energy of Mg-air batteries, utilizing sample data derived from experimental studies published in relevant literature since 2010. These publications are sourced from reputable academic journals in the fields of materials science and electrochemistry. The majority of these resources are accessible via subscription-based platforms, such as Elsevier, American Chemical Society, Springer, and Royal Society of Chemistry, offering a wide array of high-quality research articles. The data collection focuses on capturing information on alloy composition, heat treatment processes (e.g., homogenization, hot extrusion, hot rolling), cathode materials and battery performance. This effort results in the compilation of 1,235 data points on Mg-air batteries.

Subsequently, data cleansing is performed. Initially, 103 entries with missing values are eliminated. In addition, while the diversity of alloy types and the intricate composition of alloying elements contribute to an expanded information space for sample analysis, this complexity also complicates the optimization of alloy compositions and the overall improvement of their properties. To simplify the dataset, an additional 92 samples of complex five-element and six-element alloys are excluded.

Addressing bias and noise in the dataset

Variability inherent to diverse data sources and experimental parameters poses challenges, potentially introducing biases and noise that can compromise predictive model performance. To systematically address these issues, the following strategies were rigorously implemented:

Rigorous data source selection: A stringent data source selection protocol was implemented, exclusively incorporating data points derived from peer-reviewed experimental studies published post-2010 in high-impact academic journals. This criterion established a fundamental level of data reliability and adherence to established methodological standards. Further prioritization was accorded to studies characterized by comprehensive experimental descriptions, complete datasets, and a demonstrably higher citation frequency, indicative of their broad influence within the scientific community.

Feature engineering of key experimental conditions: Key experimental parameters, such as alloy composition, heat treatment protocols, and cathode material classifications, were systematically extracted and engineered into distinctive features for the machine learning (ML) model. This approach enabled the model to explicitly learn the influence of these variables on battery performance, thereby effectively accounting for and reducing inherent biases stemming from varied experimental methodologies.

Data standardization: To minimize inter-study variability and mitigate inherent biases, experimental conditions and reported parameters were rigorously standardized. This involved normalization based on widely accepted metrics, including metal composition ratios, current density processing methods, and categorical classifications of cathode catalysts, ensuring comparability across disparate datasets.

Cross-validation: K-fold cross-validation was systematically employed to rigorously assess the generalization capability and robustness of the ML model, concurrently identifying potential latent biases or noise within the dataset. This involved systematically partitioning the dataset into distinct training and validation subsets, facilitating iterative model evaluation across diverse data permutations.

Addressing duplicate or conflicting data

In instances where disparate performance data were reported for ostensibly identical materials within the literature, the primary principle guiding our approach was to meticulously investigate whether these discrepancies originated from variations in experimental conditions (e.g., distinct testing temperatures, current densities, electrolyte compositions, or specific processing protocols). Following careful review, if data for the same material under purportedly identical experimental conditions nevertheless exhibited variations, these discrepancies were generally indicative of potential measurement errors, batch-to-batch variations in sample preparation, or reporting inconsistencies within the literature. In such scenarios, the subsequent hierarchical approaches were employed:

Prioritization: Prioritization was accorded to data sourced from higher-impact, authoritative journals, or datasets demonstrably widely cited by subsequent research, reflecting their greater reliability and scientific consensus.

Statistical averaging/median: Where multiple reliable but marginally divergent data points existed, their average or median value was considered for inclusion. The median, in particular, was favored for its enhanced robustness against potential outliers.

Conservative estimation: In specific scenarios, to bolster model robustness and prevent overly optimistic predictions, a more conservative estimate was adopted. For instance, if the optimization objective was specific energy, a lower reported value might be selected.

Outlier analysis: Data points exhibiting significant deviations were subjected to rigorous outlier analysis. This process, critically informed by domain-specific expertise, guided decisions regarding their correction, removal, or further meticulous scrutiny.

Addressing data heterogeneity challenges

The collection of training data from heterogeneous literature sources inherently poses challenges to data uniformity, due to potential variations in experimental details across research groups and the limited ability of selected features to fully capture such variance. To meticulously address these issues and ensure robust model training, a multi-layered strategy was employed.

Initially, data acquisition prioritized high-impact and authoritative journals, or those extensively cited in subsequent literature. This established a foundation of reliability and helped mitigate inconsistencies arising from less rigorously reported studies. Furthermore, once it was confirmed that observed data discrepancies were not attributable to identifiable variations in experimental conditions (e.g., measurement error, batch-to-batch variability, or reporting inconsistencies), a nuanced approach to data treatment was adopted.

In instances of slight numerical discrepancies among multiple reliable data points, the median was preferred over the mean owing to its greater robustness against potential outliers, which might stem from subtle, unaccounted-for experimental differences. For energy density, conservative estimates were additionally employed to mitigate the risk of overly optimistic model predictions in the presence of inherent data scatter.

Concurrently, a rigorous outlier analysis was conducted on data points exhibiting significant deviations. Guided by domain expertise, decisions were made regarding their correction, removal, or further investigation, thereby directly addressing potential anomalies stemming from diverse experimental protocols. Ultimately, to further optimize the dataset, minimize intrinsic noise, and enhance the model's generalizability to core material systems, a select number of complex multi-element alloy samples—often associated with higher inherent variability and more challenging experimental standardization—were systematically excluded.

Detailed data cleaning process

The data cleaning process comprised the following sequential steps:

Missing value handling: Upon initial compilation, a systematic assessment for missing values was performed. A total of 103 entries were identified as lacking target values. To uphold data quality and facilitate effective model training, these entries were systematically excluded from the dataset.

Exclusion of complex alloy samples: To streamline the dataset representation and enable the ML model to more accurately capture the performance patterns of fundamental alloy systems, highly complex five-element and six-element alloy samples were systematically excluded. This accounted for 92 multi-element entries. Their removal was justified by their inherent compositional complexity, which typically leads to performance characteristics significantly divergent from more prevalent binary, ternary, or quaternary alloys. Furthermore, as they constituted a minor fraction of the overall dataset, their exclusion aimed to optimize the model's generalization capability and enhance interpretability.

Data Standardization: Throughout the entire data cleaning process, meticulous attention was paid to ensuring the consistency of all numerical data units and standardizing data formats, thereby optimizing the dataset for subsequent ML model processing.

Exclusion of higher-order Mg alloys

This work intentionally excluded higher-order Mg alloys (quinary and above) from the initial dataset primarily due to insufficient statistical representativeness. The scarcity of data points for these systems would introduce noise, risk overfitting, and degrade the robustness and generalization ability of ML models. Moreover, higher-order alloys often exhibit distinct strengthening mechanisms and microstructural evolutions (e.g., high-entropy effects) compared to lower-order systems. Mixing fundamentally different metallurgical systems without adequate data could lead to misleading correlations and hinder accurate model training.

This study focuses on establishing a robust, interpretable ML methodology for screening new Mg anode materials using simpler alloy systems (binary, ternary, quaternary). This strategic exclusion ensures foundational reliability, allowing to rigorously validate the approach and lay a solid basis for future methodological extensions to more complex compositional spaces.

While higher-order Mg alloys represent a promising frontier, their data scarcity is a current challenge. We propose an iterative methodology: combining high-throughput theoretical screening (e.g., DFT, CALPHAD) to generate virtual datasets, followed by targeted experimental verification. This experimental feedback will continuously calibrate and optimize the model, forming a closed-loop of prediction and refinement. Future work will integrate deep learning and multi-scale data fusion to accelerate the discovery and application of these advanced alloys.

Feature preparation

The alloy composition and experimental testing conditions are recognized as critical parameters influencing the specific energy of Mg-air batteries. Features that describe the electronic and physical properties of Mg alloys are utilized, enabling a more comprehensive understanding of their characteristics. Importantly, these features can be quantified without the need for intricate and time-consuming quantum chemical calculations, thereby ensuring more accessible and practical. The types of features employed in this study are detailed in **Tables S1 and S2**. As shown in **Table S1**, the electronic and physical properties of these alloys are categorized into 36 distinct classes, with each class yielding 3 distinct features derived from Formulas (1)-(3), where a_i represents the molar ratios of the *i*th constituent elements of Mg alloys, x_i represents the property of the *i*th element, $\bar{x}$ denotes the mean value of property, δ_x signifies the difference of property between the specific element in alloys and mean value of alloys, x_{Mg} represents the property of pure Mg, and $\delta_{x.Mg}$ indicates the difference of property between the specific element in alloys and Mg element.

For category *i*, Formula (1) computes the average electronic or physical properties ($\bar{x}$) by summing values of *i* property (x_i) according to the molar ratios (a_i) of the constituent elements in Mg alloys, resulting in feature "$i.m$". For instance, *Eea* represents electron affinity; thus, the average electron affinity of the alloy is calculated using Formula (1) and is designated as feature "*Eea.m*". Similarly, Formula (2) is employed to evaluate the variations in electronic and physical properties among the elements in magnesium alloys, referred to as feature "i". Furthermore, Formula (3) assesses the performance quality differences between the doped elements and magnesium within the alloys, designated as feature "$i.Mg$".

Feature $i.m$: $$\overline{\mathrm{x}} = \sum a_i x_i \quad (1)$$

Feature i: $$\delta_x = \sqrt{\sum a_i \left(1 - \frac{x_i}{\overline{x}}\right)^2} \quad (2)$$

Feature $i.Mg$: $$\delta_{x.Mg} = \sqrt{\sum a_i \left(1 - \frac{x_i}{x_{Mg}}\right)^2} \quad (3)$$

Note that *Eea* is calculated exclusively using Formula 1 and Formula 2, whereas *Smis*, mixing entropy, is calculated solely with Formula 1. As a result, 105 features are generated to describe the electronic and physical properties of alloys. The mole fraction of 27 elements (listed in **Table S2**) is taken into account as features. Furthermore, features representing six different thermal treatment processes were incorporated, which significantly influences the microstructure and macroscopic properties of the alloys. The type of cathode catalyst is also included, given its pivotal role in the oxygen reduction reaction within the battery. For these parameters, one-hot encoding was applied to properly represent their distinct nominal categories. The current density and discharge time, critical experimental parameters, are included as well, as they are closely related to the power output and energy density of the battery. In summary, we have assembled a comprehensive set of 144 features. These features provide ML model with a holistic perspective to understand and predict the performance of different alloys when they are used as anodes in Mg-air batteries.

Table S1 The list of physical features for predicting specific energy.

Physical properties[a)]			
Abb.	**Description**	**Abb.**	**Description**
XP	Pauling electronegativities	D	Averaged density
I1	First ionization energies	Dl	Averaged density in liquid
I2	Second ionization energies	Tm	Averaged melting temperature calculated by the rule of mixtures
I3	Third ionization energies	K	Averaged thermal conductivity
AW	Atomic weight	R	Averaged resistivity
VEC	Valence electrons	C	Averaged electrical conductivity
A1	Resulting first lattice constant calculated by the rule of mixtures	Hf	Heat of fusion
A2	Resulting second lattice constant calculated by the rule of mixtures	Cs	Specific heat capacity
A3	Resulting third lattice constant calculated by the rule of mixtures	Cm	Molar heat capacity
Rm	Atomic radii	Smis	Mixing entropy
Rc	Covalent radii	H1	Mohs hardness
Eea	Electron affinity	H2	Brinell hardness
MV	Molar Volume	M1	Bulk modulus
AN	Atomic number	M2	Shear modulus
SEN	The number of electrons in the second outermost shell	M3	Young modulus
PR	Poisson ratio	MS3	Volume magnetic susceptibility
MS1	Mass magnetic susceptibility	S	Standard entropy
MS2	Molar magnetic susceptibility	H	Standard enthalpy

[a)] The element parameters come from http://www.periodictable.com.

Table S2 The list of experimental features for predicting specific energy.

	Experimental parameters		**Experimental parameters**
Abb.	**Description**	**Abb.**	**Description**
Mg/%	The molar fraction of Mg	Ag/%	The molar fraction of Ag
Al/%	The molar fraction of Al	Er/%	The molar fraction of Er
Zn/%	The molar fraction of Zn	Ge/%	The molar fraction of Ge
Y/%	The molar fraction of Y	Cd/%	The molar fraction of Cd
Zr/%	The molar fraction of Zr	Cu/%	The molar fraction of Cu
Sm/%	The molar fraction of Sm	Pr/%	The molar fraction of Pr
Mn/%	The molar fraction of Mn	Th/%	The molar fraction of Th
Gd/%	The molar fraction of Gd	P1	Homogenization
Ca/%	The molar fraction of Ca	P2	Hot extrusion

In/%	The molar fraction of In	P3	Hot rolling
Bi/%	The molar fraction of Bi	P4	Quenching
Li/%	The molar fraction of Li	P5	As-cast
Sn/%	The molar fraction of Sn	P6	Solution
Ce/%	The molar fraction of Ce	MA	Current density
Ga/%	The molar fraction of Ga	Num	The number of elements in alloys
Si/%	The molar fraction of Si	CT	Discharge time
Ba/%	The molar fraction of Ba	Cat.Pt	Pt electrode
Nd/%	The molar fraction of Nd	Cat.Mn	MnO_2
Pb/%	The molar fraction of Pb	Cat.Ag	Ag electrode
La/%	The molar fraction of La		

SHAP analysis

Shapley Additive Explanations (SHAP) is a model interpretation method based on Shapley values from game theory. A key feature of the SHAP model is its additivity, which means that the model output can be expressed as the sum of each feature contributions. Formally, given an input x, SHAP model can be represented as follows: $f(x) = \Phi_0 + \sum_{i=1}^{M} \Phi_i$ (4), where $f(x)$ is the model output, Φ_0 denotes the baseline value (e.g., the average prediction of the model) and Φ_i represents the SHAP value for the *i*th feature.

Table S3 The prediction accuracy of XGB models after screening features by RFE-Lasso.

Features	**R^2**			**MAE**		
	Train	**Val**	**Test**	**Train**	**Val**	**Test**
144 features	0.96	0.81	0.51	48	141	146
100 features	0.97	0.71	0.56	41	155	144
90 features	0.97	0.77	0.76	36	149	114
80 features	0.97	0.79	0.42	39	146	183
70 features	0.97	0.78	0.61	39	146	142
60 features	0.98	0.79	0.81	35	137	93
50 features	**0.98**	**0.79**	**0.82**	**20**	**134**	**106**
40 features	0.97	0.75	0.79	43	157	120
39 features	0.97	0.78	0.68	36	120	146
38 features	0.97	0.81	0.63	27	130	131
37 features	**0.97**	**0.81**	**0.78**	**22**	**120**	**119**
36 features	0.97	0.75	0.76	21	154	117

35 features	0.97	0.74	0.72	26	158	131

Table S4 The parameter settings of models after screening features by RFE-Lasso.

Feature number	**Data partitioning**	**RFE-Lasso**	**XGB**
144 features	test_size=0.15, random_state =3		random_state=0, n_estimators=31, max_depth =7
100 features	test_size=0.15, random_state=1	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=100, step=1)	random_state=0, n_estimators=31, max_depth =8
90 features	test_size=0.15, random_state=5	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=90, step=1)	random_state=0, n_estimators=31, max_depth =8
80 features	test_size=0.15, random_state=5	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=80, step=1)	random_state=0, n_estimators=31, max_depth =8
70 features	test_size=0.15, random_state=5	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=70, step=1)	random_state=0, n_estimators=31, max_depth =8
60 features	test_size=0.15, random_state=5	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=60, step=1)	random_state=0, n_estimators=31, max_depth =8
50 features	test_size=0.15, random_state=5	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=50, step=1)	random_state=0, n_estimators=71, max_depth =8
40 features	test_size=0.15, random_state=5	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=40, step=1)	random_state=0, n_estimators=31, max_depth =8
39 features	test_size=0.15, random_state=0	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=39, step=1)	random_state=0, n_estimators=31, max_depth =8
38 features	test_size=0.15, random_state=2	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=38, step=1)	random_state=0, n_estimators=61, max_depth =8
37 features	test_size=0.15, random_state=0	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=37, step=1)	random_state=0, n_estimators=61, max_depth =9
36 features	test_size=0.15, random_state=5	Lasso(alpha=1), RFE(estimator=lasso, n_features_to_select=36, step=1)	random_state=0, n_estimators=91, max_depth =8
35 features	test_size=0.15, random_state=5	Lasso(alpha=1), RFE(estimator=lasso,	random_state=0, n_estimators=61,

n_features_to_select=35, step=1)	max_depth =8

Table S5 The detailed meanings of the 37 features selected by RFE-Lasso.

Order	Feature name	Definition	Feature importance
1	Al/%	The molar fraction of Al	13.3
2	Zn/%	The molar fraction of Zn	7.0
3	Y/%	The molar fraction of Y	0.8
4	Zr/%	The molar fraction of Zr	0.1
5	Sm/%	The molar fraction of Sm	1.2
6	Mn/%	The molar fraction of Mn	2.6
7	Gd/%	The molar fraction of Gd	0.3
8	Ca/%	The molar fraction of Ca	7.4
9	In/%	The molar fraction of In	1.0
10	Ce/%	The molar fraction of Ce	1.2
11	Ba/%	The molar fraction of Ba	0.1
12	Pb/%	The molar fraction of Pb	0.0
13	Er/%	The molar fraction of Er	0.4
14	Cd/%	The molar fraction of Cd	0.2
15	VEC	Valence electron deviation	21.3
16	D	Averaged density deviation	30.0
17	A2	Second lattice constant deviation	7.4
18	Cm	Molar heat capacity deviation	11.0
19	Eea	Electron affinity deviation	76.4
20	H1.m	Averaged mohs hardness	20.3
21	H2	Brinell hardness deviation	28.4
22	M2.Mg	Difference in shear modulus of alloy with pure Mg	25.7
23	PR	Poisson ratio deviation	10.8
24	MS1	Mass magnetic susceptibility deviation	9.8
25	MS2	Molar magnetic susceptibility deviation	13.1
26	SEN.m	The averaged number of electrons in the second outermost shell	16.0
27	S.Mg	Standard entropy deviation	8.6
28	P1	Homogenization	4.1
29	P2	Hot extrusion	1.6
30	P3	Hot rolling	3.0
31	P4	Quenching	5.9
32	P5	As-cast	38.3
33	MA	Current density	116.3
34	Cat.Pt	Pt electrode	5.1
35	Cat.Mn	MnO_2	8.1
36	Cat.Ag	Ag electrode	0.0

37	Num	The number of elements in alloys	6.0

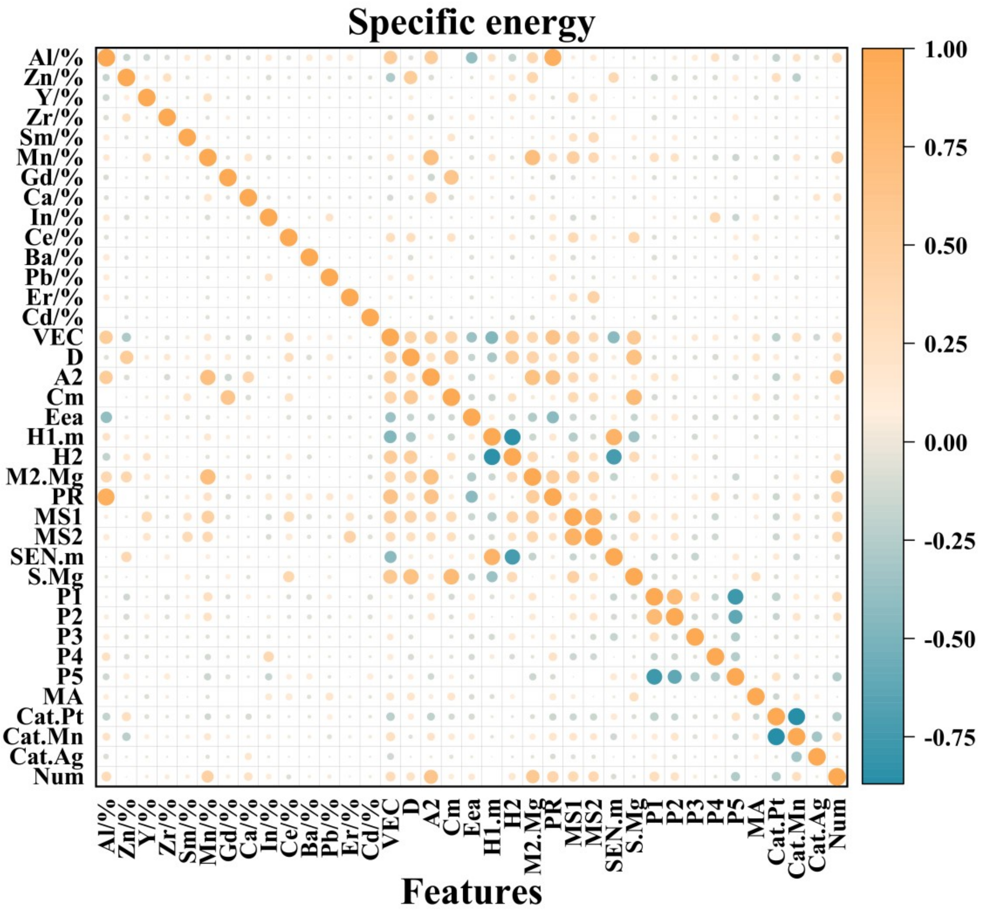

Fig. S1. The heatmap of correlation coefficients (*Ci,j*) between every two features calculated via the Pearson method.

Performance comparison of models

To rigorously validate the effectiveness of the RFE-Lasso method in feature selection, its performance was comparatively assessed against RFE-XGB and Lasso, with the number of retained features consistently fixed at 37. As depicted in **Fig. S2a**, RFE-Lasso exhibited markedly superior feature selection capabilities, achieving a **Test** dataset R^2 of 0.78. This performance significantly surpassed RFE-XGB ($R^2 = 0.51$) and Lasso ($R^2 = 0.63$), primarily due to its synergistic integration of the advantages inherent to both RFE and Lasso. Notably, this method proved particularly efficacious for predicting Mg alloys containing elements not present in the original training dataset.

Subsequently, a comparative analysis was undertaken to evaluate the applicability and predictive performance of five prevalent ML algorithms—XGBoost (XGB), Random Forest (RF), Gradient Boosting Regression (GBR), Support Vector Machine (SVM), and Multilayer Perceptron (MLP)—on the Mg-air battery dataset. The parameter settings of these models are presented in **Tables S6**. As illustrated in **Fig. S2b**, XGBoost consistently emerged as the highest-performing algorithm, yielding the most robust R^2 values on the **Val** set (0.81), the **Test** set (0.78), and across 10-fold cross-validation (average $R^2 = 0.76$). In stark contrast, while RF performed reasonably well on the **Val** set and during cross-validation, its **Test** dataset R^2 declined significantly to 0.58, indicating a suboptimal capacity for generalization beyond the training data. Furthermore, GBR, SVM, and MLP consistently demonstrated lower predictive performance across both the initial and **Test** datasets. These findings unequivocally establish XGB as the most reliable and effective model for predicting battery properties within this study, offering superior robustness and generalization capabilities compared to the other evaluated algorithms.

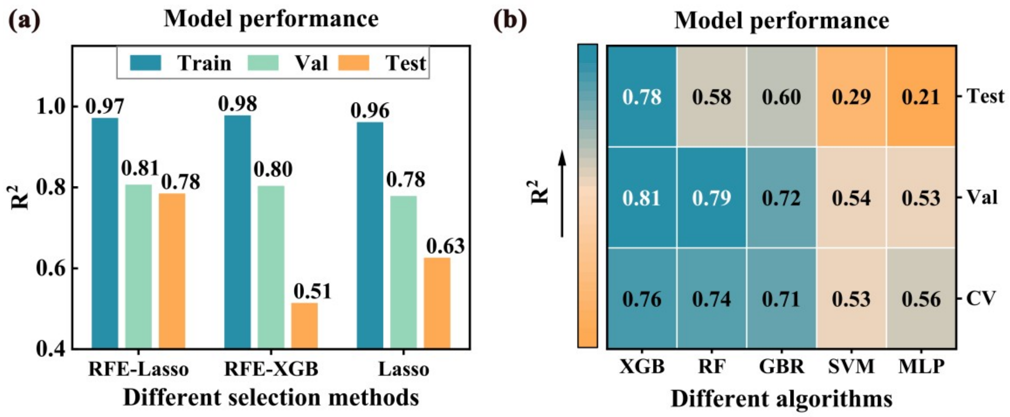

Fig. S2 Performance comparison of models: (a) models employing three distinct feature selection methods; (b) models utilizing different algorithms.

Table S6 The parameter settings for different algorithms.

Algorithms	**Parameters**
XGB	random_state = 0, n_estimators = 61, max_depth = 9
RF	n_estimators = 31, max_depth = 19, bootstrap = True, random_state = 47
GBR	learning_rate = 0.28, random_state = 47, n_estimators = 60
SVM	Kernel = 'rbf', C = 1300.0, epsilon = 1e-05, gamma = 0.02
MLP	max_iter=80, hidden_layer_sizes = [64, 64], alpha = 0.58, activation = 'relu', solver = 'lbfgs', random_state = 30

Detailed calculation of "main SHAP values"

To precisely quantify the individual contribution of each feature to the model's prediction, "Main SHAP values" were derived using the SHAP interaction module. This approach is efficient for distinguishing the independent impact of a feature from its synergistic or antagonistic interactions with other features, thereby enhancing the interpretability of ML model.

The SHAP interaction module specifically computes a square interaction matrix for each datapoint, where off-diagonal elements ($SHAP_{i,j}$) represent the interaction effect between feature *i* and feature *j*. A positive $SHAP_{i,j}$ indicates a synergistic effect, implying that the joint presence of features *i* and *j* enhances the target prediction. Conversely, a negative value signifies an antagonistic effect, where their combined influence diminishes the target prediction.

The "Main SHAP values" are specifically defined by the diagonal elements of this matrix, $SHAP_{i,i}$. These self-interaction terms quantify the isolated, individual contribution of feature *i* to the model's output. By design, $SHAP_{i,i}$ reflects the unique marginal impact of feature *i*, effectively filtering out the influence of other variables. Consequently, it represents the pure, independent contribution of that single feature.

Therefore, for each sample in the dataset and for all 37 features, the SHAP interaction module was utilized to compute these interaction matrices. The "Main SHAP value" for any given feature was then directly extracted from the corresponding diagonal element ($SHAP_{i,i}$) of the corresponding matrix. This methodology provides a direct and explicit measure of each feature's independent predictive power across all samples, which is crucial for robust scientific interpretation.

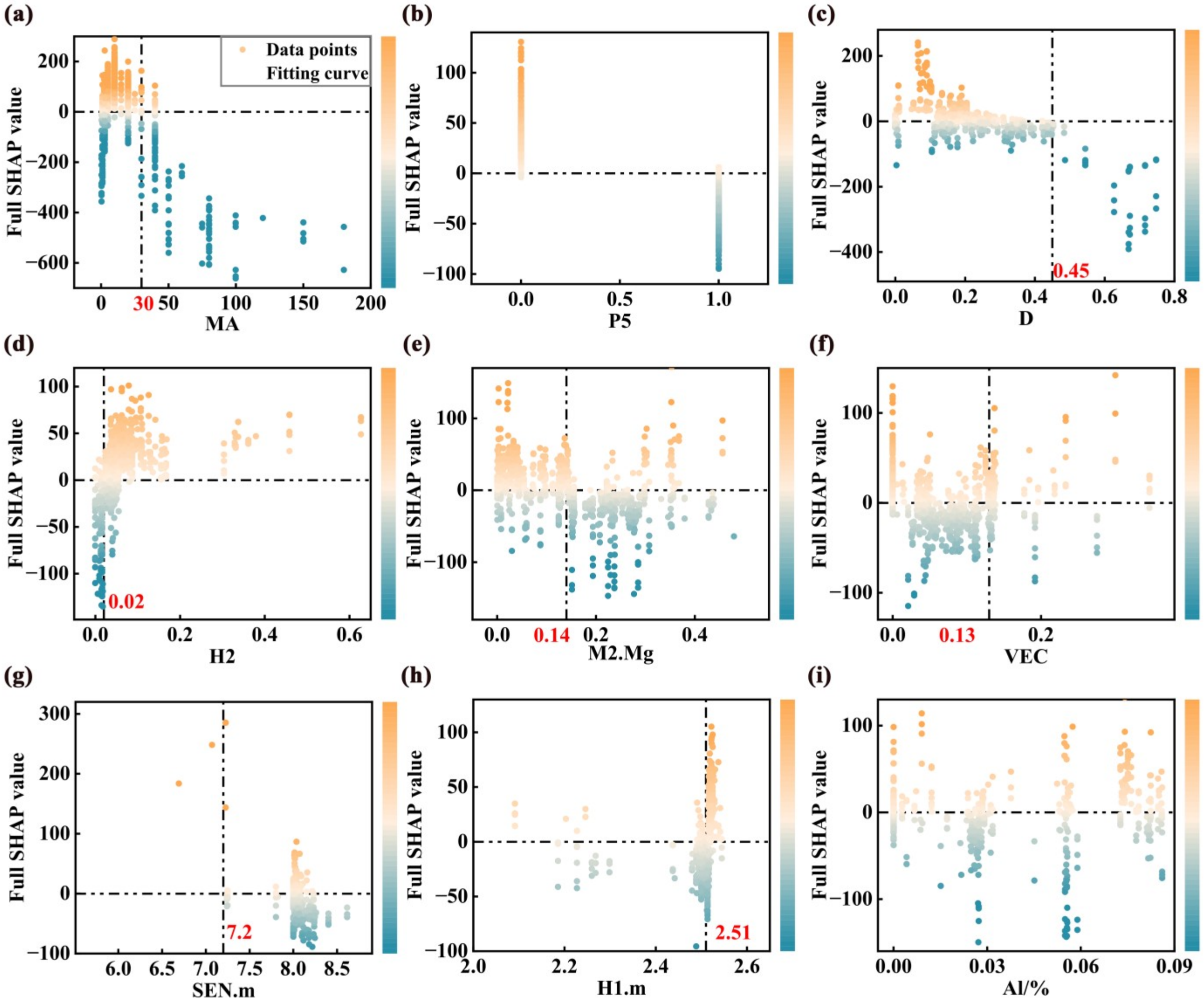

Fig. S3. The scatter plot of full SHAP values before filtering out the interaction effects: (a) MA; (b) P5; (c) D; (d) H2; (e) M2.Mg; (f) VEC; (g) SEN.m; (h) H1.m; (i) Al/%.

Table S7 The threshhold value of screening conditions.

Abb.	**Descriptor**	**Threshold value**
A	MA	MA ≤ 30
B	Eea	Eea ≥ 13.5
C	P5	P5 = 0
D	D	D ≤ 0.45
E	H2	H2 ≥ 0.02
F	M2.Mg	M2.Mg ≤ 0.14 or M2.Mg ≥ 0.30
G	VEC	VEC ≤ 0.05 or VEC ≥ 0.13
H	H1.m	H1.m ≥ 2.51
B + D	Eea, D	Eea ≥ 13.5, D ≤ 0.45
B + F	Eea, M2.Mg	Eea ≥ 13.5, M2.Mg ≤ 0.14 or M2.Mg ≥ 0.30
B + G	Eea, VEC	Eea ≥ 13.5, VEC ≤ 0.05 or VEC ≥ 0.13
A + B + F	MA, Eea, M2.Mg	MA ≤ 30, Eea ≥ 13.5, M2.Mg ≤ 0.14 or M2.Mg ≥ 0.30
A + B + G	MA, Eea, VEC	MA ≤ 30, Eea ≥ 13.5, VEC ≤ 0.05 or VEC ≥ 0.13
B + F + G	Eea, M2.Mg, VEC	Eea ≥ 13.5, M2.Mg ≤ 0.14 or M2.Mg ≥ 0.30, VEC ≤ 0.05 or VEC ≥ 0.13
A + B + F + G	MA, Eea, M2.Mg, VEC	MA ≤ 30, Eea ≥ 13.5, M2.Mg ≤ 0.14 or M2.Mg ≥ 0.30, VEC ≤ 0.05 or VEC ≥ 0.13
A + B + C + F	MA, Eea, P5, M2.Mg	MA ≤ 30, Eea ≥ 13.5, P5 = 0, M2.Mg ≤ 0.14 or M2.Mg ≥ 0.30

Table S8 The screening conditions validated using the initial dataset comprising 1,024 samples for establishing the ML model.

Abb.	**Descriptor**	**True**			**False**		
		Data points	Good	Bad	Data points	Good	Bad
A	MA	878	15.1%	84.9%	146	2.7%	97.3%
B	Eea	283	24.7%	75.3%	741	9.0%	91.0%
C	P5	480	17.1%	82.9%	544	10.1%	89.9%
D	D	981	13.6%	86.4%	43	9.3%	90.7%
E	H2	819	14.3%	85.7%	205	9.8%	90.2%
F	M2.Mg	634	18.5%	81.5%	390	5.1%	94.9%

G	VEC	589	17.7%	82.3%	435	7.6%	92.4%
H	H1.m	367	9.3%	90.7%	657	15.7%	84.3%
B + D	Eea, D	276	25.4%	74.6%	748	9.0%	91.0%
B + F	Eea, M2.Mg	**223**	**30.9%**	**69.1%**	**801**	**8.5%**	**91.5%**
B + G	Eea, VEC	230	29.1%	70.9%	794	8.8%	91.2%
A + B + F	MA, Eea, M2.Mg	**196**	**35.2%**	**64.8%**	**828**	**8.2%**	**91.8%**
A + B + G	MA, Eea, VEC	199	33.7%	66.3%	825	8.5%	91.5%
B + F + G	Eea, M2.Mg, VEC	196	33.7%	66.3%	828	8.6%	91.4%
A + B + F + G	MA, Eea, M2.Mg, VEC	**174**	**37.9%**	**62.1%**	**850**	**8.4%**	**91.6%**
A + B + C + F	MA, Eea, P5, M2.Mg	123	33.3%	66.7%	901	10.7%	89.3%

Features B (Eea) and F (M2.Mg) are top 2 effective screening conditions with larger "Good" yields in "Ture" samples (**Table S8**). Feature A (MA) emerges as the most effective to evaluate the "Bad" yield in "False" sample. Although these 8 screening conditions all exhibit high "Bad" yields in "False" sample, their "Good" yields in "Ture" samples are insufficient to reliably determine the potential anodes. Therefore, it is not reliable and inadequate to identify the high-quality anodes on the threshold value achieved from a single descriptor.

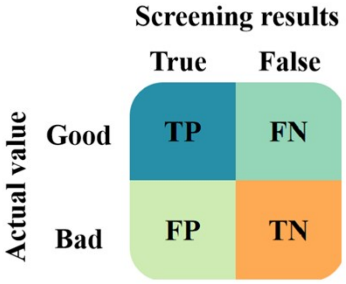

Fig. S4. The confusion matrix of screening results.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2TP}{2TP + FN + FP} \tag{8}$$

Table S9 The accuracy, presion, recall and F1-score of screening results in original dataset.

Abb.	Description	Accuracy	Precision	Recall	F1-score
A	MA ≤ 30	26.9%	15.1%	97.1%	26.2%
B	Eea ≥ 13.5	72.7%	24.7%	51.1%	33.3%
C	P5 = 0	55.8%	17.1%	59.9%	26.6%
D	D ≤ 0.45	16.8%	13.6%	97.1%	23.8%
E	H2 ≥ 0.02	29.5%	14.3%	85.4%	24.5%
F	M2.Mg ≤ 0.14 or M2.Mg ≥ 0.30	47.6%	18.5%	85.4%	30.4%
G	VEC ≤ 0.05 or VEC ≥ 0.13	49.4%	17.7%	75.9%	28.7%
H	H1.m ≥ 2.51	57.4%	9.3%	24.8%	13.5%
B+D		73.3%	25.4%	51.1%	33.9%
B+F		**78.3%**	**30.9%**	**50.4%**	**38.3%**
B+G		77.2%	29.1%	48.9%	36.5%
A+B+F		**81.0%**	**35.2%**	**50.4%**	**41.4%**
A+B+G		80.3%	33.7%	48.9%	39.9%
B+F+G		80.4%	33.7%	48.2%	39.6%
A+B+F+G		**82.5%**	**37.9%**	**48.2%**	**42.4%**
A+B+C+F		82.6%	33.3%	29.9%	31.5%

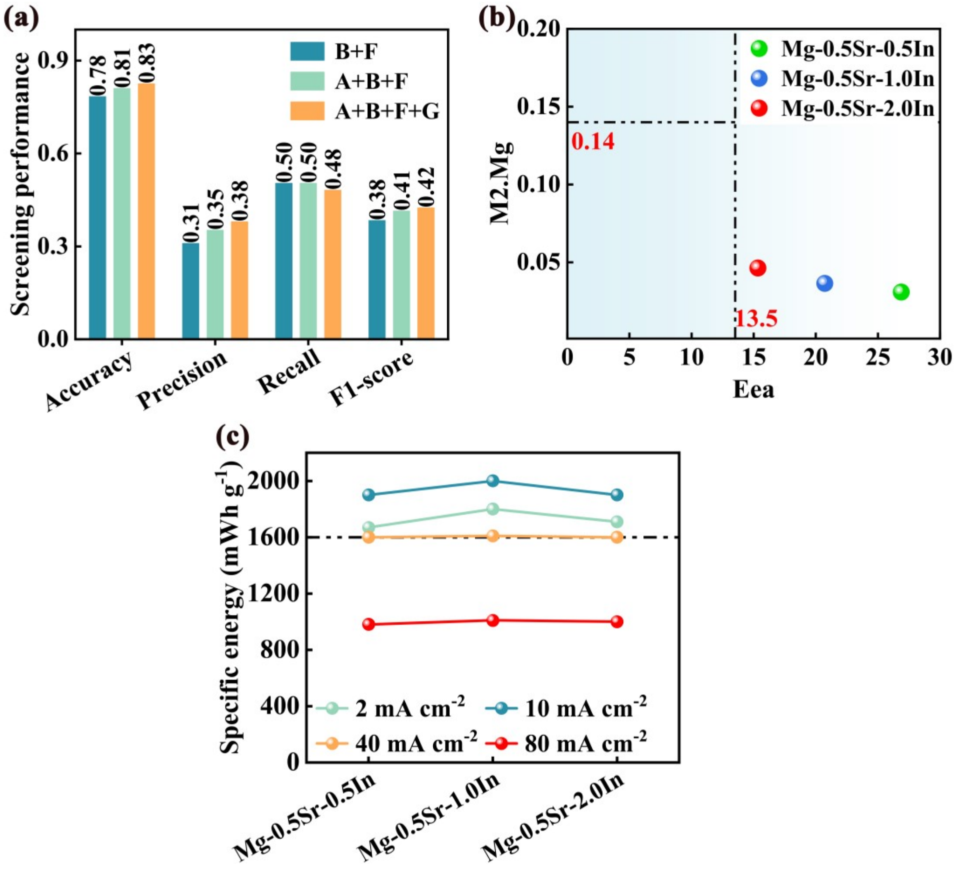

Fig. S5. Screening results and model predictions: (a) the accuracy, presion, recall and F1-score of screening results in original dataset; (b) Eea and M2.Mg value of 12 data points derived from literature data; (c) actual specific energy of 12 samples derived from literature data.

Screening performance on ternary Mg-Sr-In alloys

Additionally, the effectiveness of these screening criteria is systematically evaluated by four key metrics in machine learning classification: accuracy, precision, recall, and the F1-score. Detailed methodologies for calculating accuracy, precision, recall, and the F1-score, are provided in **Fig. S4**, Equations **(5)** - (**8**), and detailed values are given in **Table S9**. The definitions of the four metrics are as follows: TP (True Positive) denotes the number of positive samples correctly predicted as positive; TN (True Negative) denotes the number of negative samples correctly predicted as negative; FP (False Positive) denotes the number of negative samples incorrectly predicted as positive; FN (False Negative) denotes the number of positive samples incorrectly predicted as negative. Accuracy represents the proportion of samples that are correctly predicted out of the total sample set. Precision indicates the ratio of actual positives in the predicted positives. Recall denotes the proportion of actual positive samples that are correctly identified as positive. The F1-score is the harmonic mean of precision and recall, balancing both metrics.

As illustrated in **Fig. S5a**, the accuracy for each combination is approximately 0.80, demonstrating that precise screening can be accomplished with straightforward combinations of several descriptors. Considering both the screening performance and concise, the combination of B + F is selected as the screening condition. Furthermore, it has been verified whether the B+F screening condition can be extended to Mg-Sr-In ternary alloys through literature data. The Mg-0.5Sr-0.5In, Mg-0.5Sr-1.0In, and Mg-0.5Sr-2.0In alloys derived from the literature meet the threshold values of the B+F condition (**Fig. S5b**), and thus they should exhibit excellent battery performance. Indeed, as reported in the literature, the aforementioned three alloys can achieve a specific energy of at least 1600 mWh g^{-1} at current densities not exceeding 40 mA cm^{-2} (**Fig. S5c**).

Table S10 ICP results for the chemical composition of the investigated alloys (wt.%).

Alloys	Mg	Sr
Mg-0.1Sr	Bal.	0.09
Mg-0.5Sr	Bal.	0.42
Mg-1Sr	Bal.	1.01

Discharge performance

After discharge, specific energy of the Mg-air battery were calculated using the following equations:

$$Specific\ \ energy = \frac{\int_0^t I\times U\times \Delta t}{W_{total}} \tag{9}$$

$$W_{theo} = \frac{I\times t}{F\times \sum(\frac{x_i\times n_i}{m_i})} \tag{10}$$

where W_{theo} (g) means the theoretical weight loss of anode, W_{total} (g) means the actual weight loss of anode before and after discharge; I (A) is applied current, t (h) is time of discharge; U (V) represents the voltage of discharge; F is Faraday constant (26.8 A h mol^{-1}), x_i, n_i, *and* m_i represent the mass fraction, number of exchanged electrons, and atomic weight, respectively, of each element in alloys.

Full cell discharge test

The discharge performance of Mg alloys in 3.5 wt.% NaCl solution was measured in a laboratory-assembled full cell. The exposed surface area of anode was approximately 1 × 1 cm^2. The cathode was the commercial air cathode with MnO_2/C as catalyst (2 × 2 cm^2). The cell voltage of Mg-air battery was tested at applied current densities of 1, 2.5, 5, and 10 mA cm^{-2}. During the discharge process, the electrolyte was stirred with the same rate. After the full cell discharge test, the surface of Mg anode was cleaned by 200 g L^{-1} CrO_3 + 10 g L^{-1} $AgNO_3$ solution to remove the discharge products and account for the weight loss.

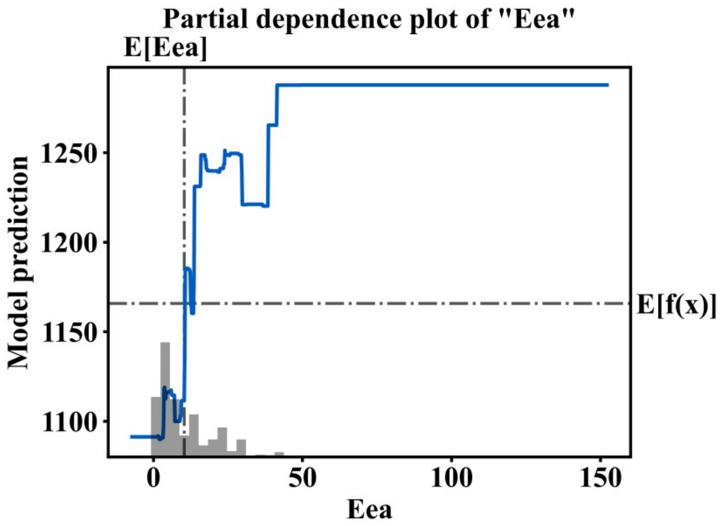

Fig. S6. Partial dependency plot of Eea feature.

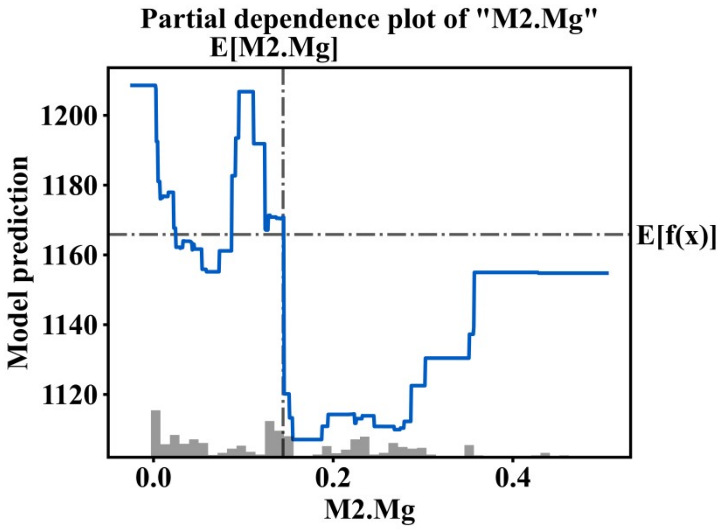

Fig. S7. Partial dependency plot of M2.Mg feature.

The availability for high-throughput screening

The superior screening performance of the combined Eea and M2.Mg descriptors (referred to as B + F in Table 1) is evident from their ability to correctly identify 30.9% of high-performing samples (true positives) and accurately reject 91.5% of low-performing samples (true negatives). This integrated descriptor approach thus serves as an efficient screening condition for high-throughput materials discovery.

Crucially, Eea and M2.Mg are readily accessible for high-throughput screening and the prediction of novel materials. Unlike approaches relying on computationally expensive theoretical calculations (e.g., density functional theory, DFT), these descriptors are derived from direct formulaic calculations. Such calculations only require fundamental physical parameters of constituent elements, such as atomic radii and shear moduli, resulting in a minimal computational cost.

This inherent low cost makes them highly suitable for the high-throughput screening of numerous virtual materials. Specifically, our study demonstrates that when an alloy fails to satisfy the proposed criteria (i.e., Eea ≥ 13.5 and (M2.Mg ≤ 0.14 or M2.Mg ≥ 0.3)), there is a

91.5% probability that a magnesium-air battery utilizing it as an anode will exhibit a specific energy below 1600 mWh g^{-1}. This high negative predictive power establishes these descriptors as potent tools for efficiently eliminating a significant portion of unqualified candidate materials.

While alloys satisfying these conditions exhibit a 30.9% probability of exceeding 1600 mWh g^{-1}—a moderate success rate for proactively identifying high-energy density materials—their exceptionally low computational burden nonetheless renders them ideal for preliminary screening. They effectively function as a rapid first-pass filter, substantially reducing the pool of materials that would otherwise necessitate more computationally intensive and precise theoretical calculations, thereby significantly accelerating the overall material discovery process.