POLITECNICO DI TORINO
Repository ISTITUZIONALE

Energy-GNoME: A living database of selected materials for energy applications

Original
Energy-GNoME: A living database of selected materials for energy applications / De Angelis, P., Barletta, G., Trezza, G., Asinari, P., Chiavazzo, E.. - In: ENERGY AND AI. - ISSN 2666-5468. - (2025). [10.1016/j.egyai.2025.100605]

Availability:
This version is available at: 11583/3003161 since: 2025-09-19T03:29:12Z

Publisher:
Elsevier

Published
DOI:10.1016/j.egyai.2025.100605

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

25 June 2026

Energy and AI 22 (2025) 100605

Contents lists available at ScienceDirect

Energy and AI

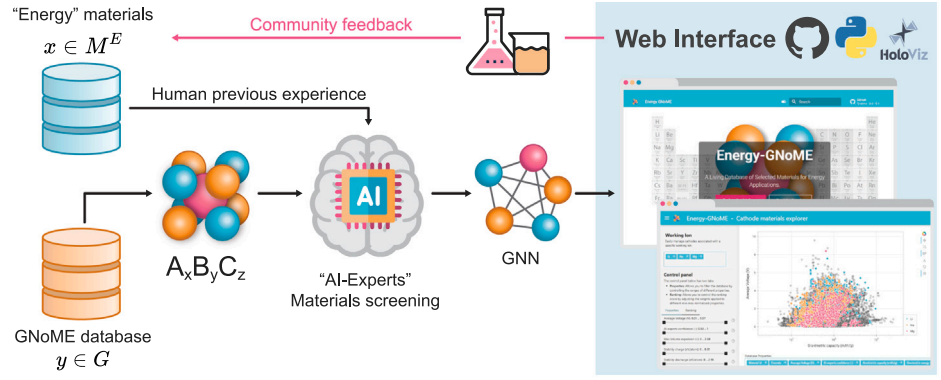
journal homepage: www.elsevier.com/locate/egyai

Energy-GNoME: A living database of selected materials for energy applications☆

Paolo De Angelis [a], Giulio Barletta [a], Giovanni Trezza [a], Pietro Asinari [a,b], Eliodoro Chiavazzo [a,*]

[a] Department of Energy, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy
[b] INRIM, Istituto Nazionale di Ricerca Metrologica, Strada delle Cacce 91, Torino, 10135, Italy

GRAPHICAL ABSTRACT



HIGHLIGHTS

- Fully AI-driven protocol to screen relevant materials for energy applications.
- 38,500 new candidate materials for thermoelectric, PV and batteries.
- User friendly Web App ensures easy access to the novel materials.
- Energy-GNoME database is refined iteratively through research community feedback.

ARTICLE INFO

Dataset link: https://github.com/paolodeangelis/Energy-GNoME/, https://doi.org/10.5281/zenodo.14338533, https://paolodeangelis.github.io/Energy-GNoME/

Keywords:
Energy materials
Artificial Intelligence

ABSTRACT

Artificial Intelligence (AI) in materials science is driving significant advancements in the discovery of advanced materials for energy applications. The recent GNoME protocol identifies over 380,000 novel stable crystals. From this, we identify over 38,500 materials with potential as energy materials forming the core of the Energy-GNoME database. Our unique combination of Machine Learning (ML) and Deep Learning (DL) tools mitigates cross-domain data bias using feature spaces, thus identifying potential candidates for thermoelectric materials, novel battery cathodes, and novel perovskites. First, classifiers with both structural and compositional features detect domains of applicability, where we expect enhanced reliability of regressors. Here, regressors are trained

☆ This article is part of a Special issue entitled: 'AI for Energy Materials' published in Energy and AI.

* Corresponding author.
E-mail address: eliodoro.chiavazzo@polito.it (E. Chiavazzo).

https://doi.org/10.1016/j.egyai.2025.100605
Received 27 February 2025; Received in revised form 5 August 2025; Accepted 1 September 2025
Available online 16 September 2025
2666-5468/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Machine Learning
Deep Learning
Thermoelectric
Battery
Perovskite

to predict key materials properties, like thermoelectric figure of merit (zT), band gap (E_g), and cathode voltage (ΔV_c). This method significantly narrows the pool of potential candidates, serving as an efficient guide for experimental and computational chemistry investigations and accelerating the discovery of materials suited for electricity generation, energy storage and conversion.

1. Introduction

The growing commitment to environmental sustainability and preservation has catalyzed a shift towards a green economy, emphasizing usage of Renewable Energy Sources (RES), decarbonization strategies, and sustainable resource management to ensure long-term ecological balance and economic resilience [1]. In this context, energy-related materials play a central role in driving the transition to a new, eco-friendly industrial paradigm. Materials for renewable energy conversion — such as perovskites for photovoltaic (PV) solar cells [2,3], materials for efficient energy usage — such as thermoelectric [4,5] materials, along with materials for energy storage devices — like cathode materials for batteries [6–8] — are the key to attenuate the intermittent nature of RES, unlocking the full potential of clean energy and achieving the overarching goal of minimizing our environmental footprint while shifting to a sustainable green economy [9–11].

Advancements in these fields are strictly correlated to the discovery of novel materials with enhanced properties. Most of these physical and chemical properties can be accurately determined by first principles methods based on Density Functional Theory (DFT) [12,13]. However, an exhaustive screening of such innovative materials with traditional strategies is to date impractical due to the high-dimensional composition spaces and the often unaffordable computational cost of DFT simulations. Furthermore, efforts in investigating hypothetical materials typically rely heavily on the intuition of the researcher for identifying promising candidates, as well as on heuristics with limited extrapolation capacities on unseen samples [14–16].

Despite these difficulties, the efforts required to address the task of materials discovery have been greatly reduced over the last years by the development of high-throughput platforms and of data-driven Machine Learning (ML) techniques. Indeed, in recent years, ML techniques have greatly impacted the way research [17,18] and industry [19–21] approach to several applications, including those in the energy field [22–24]. Coupled with the creation of extensive materials databases such as Materials Project (MP) [25], the Inorganic Crystal Structure Database (ICSD) [26], the Open Quantum Materials Database (OQMD) [27], NOMAD [28], and AFLOWLIB [29], these advanced tools have matured to unlock new potential in the materials discovery process [30,31].

The combination of high-throughput computational methods and ML approaches has been successfully applied in recent research to predict novel materials and determine key properties, driving innovation in energy storage, generation, and conversion. Fanourgakis et al. [32] applied ML methodologies to screen a wide virtual space of hypothetical Metal-Organic Frameworks (MOFs), introducing a universal strategy employing the "atom types" as the only descriptors to predict the MOFs' adsorption capacities. A similar work was carried out by Trezza et al. [33]: the authors exploited ML regressors trained to predict MOFs' adsorption capacities to establish a minimal set of important crystallographic features, and investigated the role of such "genetic code" when using Sequential Learning (SL) algorithms. Nandy et al. [14] exploited Natural Language Processing (NLP) procedures to leverage the available MOF literature, obtaining stability measures and thermal decomposition temperatures for structurally characterized MOFs. Furthermore, the authors trained Artificial Neural Network (ANN) models to predict solvent removal stability and thermal stability. Cerqueira et al. [34] created a computational dataset for conventional, i.e. Bardeen–Cooper–Schrieffer, superconductors containing *ab initio* electron–phonon calculations, which was also used for training a ML model to identify superconducting compounds with a critical temperature T_c greater than 5 K, taking as input features the compositional, structural, and ground-state properties. Moses et al. [35] leveraged data retrieved from the MP database to train deep Neural Network (DNN) models able to predict the change in volume and the average voltage of battery electrode materials during the charging and discharging processes. The authors also investigated the screening capabilities outside of the training dataset. Rutt et al. [36] applied a computational screening approach to identify promising multivalent cathodes, demonstrating the importance of evaluating both relative stability and ion mobility in materials not initially containing the working ion of interest. Here a high-throughput material exploration strategy was applied, in which the MP crystal candidates were iteratively defined in 4 steps, starting by screening the materials with relative stability above $0.2\,\mathrm{eV\,atom^{-1}}$, and selecting the crystals showing reducible potential with respect to Mg ions. The authors then proceeded with insertion site identification, and finally measured the migration path using approximate Nudged Elastic Band (ApproxNEB) algorithms [37]. Wang et al. [38] developed a computational band gap database of single and double perovskites based on highly accurate DFT calculations, and used it to identify an accurate expression to predict the band gap. This model was then employed to screen Pb-free perovskites in the MP database, finding 14 unreported crystals potentially suitable for PV applications. Kim et al. [39] trained a Random Forest (RF) based ML model on the whole OQMD to identify novel quaternary Heusler compounds. The model was employed to screen 3.2 million possible structures, predicting their stability and identifying 303 promising compositions, of which 55 were confirmed to yield stable compounds through DFT calculations. Kang et al. [40] used ML to characterize the heat of explosion of potential candidates of energetic materials, based on the constituent elements-averaged cohesive energy and on the oxygen balance. The authors applied the model to perform a high-level screening of over 140 million molecules in the PubChem database [41], followed by a theoretical fine-level screening which eventually identified 262 molecular candidates with the required properties. Rao et al. [42] investigated the compositional design of high-entropy alloys, proposing an active learning framework which combines a generative model, regression ensemble, physics-driven learning, and experiments. The authors demonstrated the framework's capabilities in the design of high-entropy Invar alloys with low thermal expansion coefficient.

In addition to these previous screening material successes, the application of ML in materials science has unlocked a vast and largely untapped resource: the Graph Networks for Materials Exploration (GNoME) database [43]. GNoME is an Artificial Intelligence (AI) driven platform designed to explore the vast chemical space through an iterative pipeline that combines active learning algorithms and Graph Neural Networks (GNNs). This process generates and filters numerous candidate solid state materials using a GNN trained and validated with DFT to predict formation energies. This active learning framework enables GNoME to continuously refine its predictions, culminating in the discovery of over 2.2 million stable materials. Remarkably, it has identified over 380,000 novel stable crystals, which reside on the updated convex hull of formation energies. To the best of our knowledge, this vast database of materials has not yet been screened to identify potential materials for disparate energy applications.

Therefore, in this study, we aim to perform a preliminary screening of the GNoME database to identify potential materials for further numerical or experimental investigation within three relevant domains in the energy field: thermoelectric materials, perovskites, and batteries. As

detailed below, the adopted protocol is general and additional domains of application can be investigated in the near future.

Specifically, we adopt specialized datasets available in the literature for training, validating and testing proper ML regressors towards the prediction of relevant properties of interest across the above three domains. A straightforward practice would consist in using those models to directly predict such properties of interest over the GNoME materials. However, the specialized datasets utilized for training represent only a localized subset of the entire materials space. As a consequence, the trained ML models — being not extrapolative [16] — are able to reliably forecast the corresponding property of interest only for those GNoME samples that fall within the same localized subset as the training materials. To take into account this biased nature of specialized datasets, we adopt the protocol recently proposed and validated by some of the authors of this work [44]. Specifically, it consists of a set of binary classifier-based filters, trained over samples from the specialized dataset (class 1) and random subsets from a *less biased* general-purpose database like MP (class 0). By applying these classifiers to the GNoME materials, we can effectively rule out samples for which the regression models are likely to provide unreliable predictions, thereby we expect an enhanced reliability and accuracy of our screening process.

The protocol identified 13,057 thermoelectric candidate materials, 4259 perovskite candidates for PV applications, and 21,243 cathode material candidates for lithium and eight *post-lithium* kind batteries.

The primary objective of this work is to introduce and validate the proposed protocol. As such, the list of screened materials is also expected to be refined and updated in the future when further experimental evidence and additional knowledge on the specialized database become available. While predictive performance varies across different cases, this study aims to demonstrate the methodology through practical examples and assess its potential for future discoveries. In this spirit, we refer to the Energy-GNoME as a *living* database.

This article is structured as follows: Section 2 includes a brief overview of the proposed protocol and workflow and presents the findings of our AI-driven screening process for various energy-related materials, including candidates for thermoelectric, perovskites, and cathodes applications. Section 3 discusses the significance and limitations of these results, their potential applications in the energy sector, and the further development of these methods. Section 4 details the computational approaches, data handling, and ML protocols used to predict material properties and identify promising candidates within the GNoME database.

2. Results

The proposed method — introduced in the next Section 2.1 and detailed in Section 4 — requires three key components: a specialized energy materials database containing experimental measurements or numerical predictions of desired properties, a significantly less biased general-purpose materials database, and a set of unexplored materials. The selection of the last two components is straightforward. We utilize the Materials Project (MP) database, which, from its deployment in 2018 to today, reached a cardinality of over 150,000 materials (of which approximately 34,000 stable and 23,000 experimentally observed) characterized using DFT. For the unexplored materials set, we use the recent GNoME database, which includes materials that have yet to be experimentally synthesized or numerically simulated for specific energy applications of interest in this work. Thus, the primary challenge is to obtain reliable and comprehensive energy material data for various applications. In the following Subsections, we introduce the adopted protocol and present the results of the ML-based screening, along with the identified potential candidates, for three case-studies corresponding to three classes of materials with significant energy relevance: thermoelectric materials, perovskites, and cathode materials. We also report the ML models predictions for the related properties: figure of merit for thermoelectric materials, band gap for perovskites, and reduction potential for cathode materials.

2.1. Protocol overview

We propose that there exists a region E within the high-dimensional feature space of all materials, containing all materials suited for a given energy application. The intersection between E and a general-purpose database of known materials (e.g., MP) contains known materials for energy applications. This "energy-specific" subset $M^E = M \cap E$ and its complement $M \setminus E$ are leveraged by the AI-experts to demarcate the boundary of E. This approach allows for the identification of the intersection between the dataset of unexplored materials G (namely the GNoME database) and E, $G^E = G \cap E$, i.e., the subset of crystals in G that share properties with $M \cap E$. Here, regression yields more reliable results than when applied to the entire set G.

For a detailed explanation of the protocol and methods, please refer to Section 4. Additionally, Fig. 7 visually represents the relationships among these different sets within the high-dimensional material feature space aforementioned.

The discovery of energy materials within the GNoME database by identifying the G^E subset is achieved through the designed protocol illustrated in Fig. 1. The process involves two distinct workflows: the training workflow and the prediction workflow, indicated in Fig. 1 with a gray dashed line and a black solid line, respectively.

The *training workflow* goes through all the steps to instruct all the ML models, starting from the specialized "energy" materials database M^E. The training set data is used to qualify the AI-expert algorithms to classify and, therefore, hypothetically identify the boundary, ∂E, of the energy material region in the n-dimensional space and to train regression ML algorithms to predict the specific property of interest for each class of energy materials. In the proposed case study, we have chosen the figure of merit (zT) for thermoelectric materials, the band gap (E_g) for perovskites, and the reduction potential (ΔV) for cathode materials. This workflow ends with assessing the performance of all the models. The available data for materials in energy applications do not always include structural information, such as crystallographic files (e.g., CIF or XYZ files) or unique string identifiers like the International Chemical Identifier (InChI). Consequently, for the regression, a conditional OR switch splits the workflow into two pipelines: one optimized for materials with complete structural information, and another designed to handle cases with only compositional data. If structural data is available, the workflow follows the "structure pipeline" (top blue box in Fig. 1) that uses the graph representation of the material to train a committee of four E(3)NNs to predict the material property of interest. If only composition information is available, the data flow goes through the "composition pipeline" (center blue box in Fig. 1), which uses the descriptors array to train four Gradient Boosted Decision Trees (GBDT) models instead. On the other hand, the AI-experts — implemented as a committee of ten binary GBDT classifiers — are always trained on composition-based descriptors derived from the chemical formula; whenever a crystal structure is available, we enrich this input with additional structure-based features, enabling the models to handle *polymorphism* and achieve higher predictive accuracy while exposing biases in the specialized energy-materials database (M^E).

Once all the models are trained and validated, the data flow transitions to the *prediction workflow*, which begins from the GNoME database G. All the data points $y \in G$ are processed through the "screening pipeline" (bottom orange box in Fig. 1) where the AI-experts are consulted to compute the probability that the crystal shares the same biases as those in the specialized training set, i.e. $P(y \in M^E)$. According to our hypothesis, this probability also coincides with the likelihood of falling inside the energy material region E ($P(y \in M^E) = P(y \in E)$). All the crystals that have passed the screening process — i.e., those with an average probability from the AI-experts higher than 50% — continue to the "regressors pipeline". Here, depending on the specialized databases used for training, the materials are featurized either by converting them into graphs if a committee of E(3)NNs was trained or into descriptor arrays if a committee of GBDTs was used. The trained regressors then

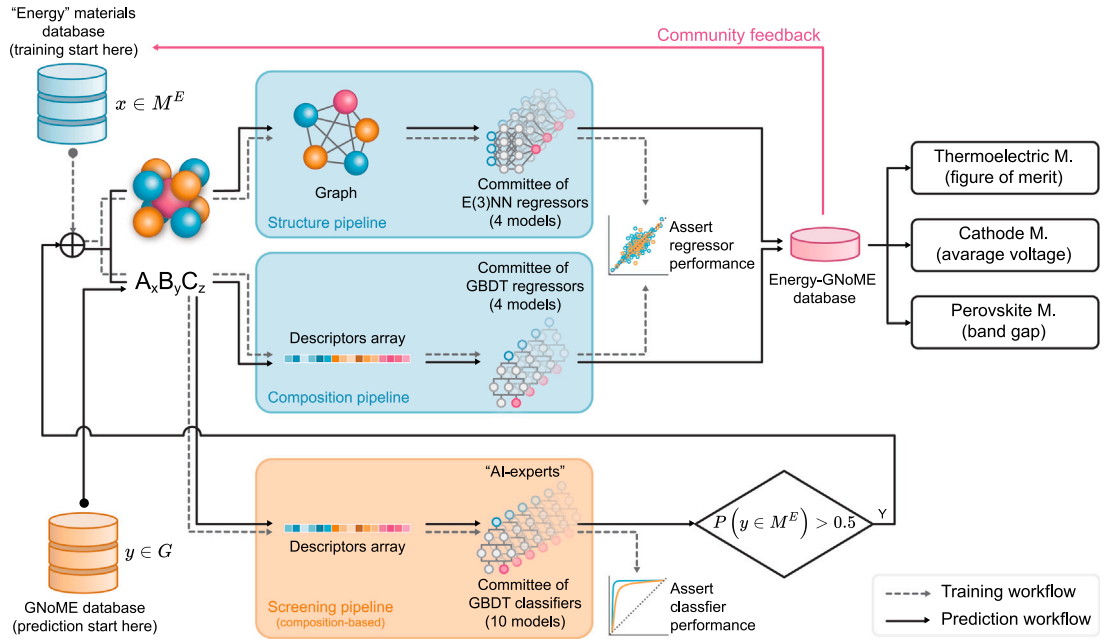


Fig. 1. The schematic shows the protocol for creating the Energy-GNoME database, illustrating training (gray dashed line) and predictive (black solid line) phases. Training begins with the cyan database and ends with ML model evaluations (e.g., parity plots, ROC curves). Feature extraction depends on material storage in the "Energy" database M^E and may use composition- or structure-based pipelines, as indicated by the OR switch symbol $\oplus$. The *structure pipeline* applies a graph representation, while the *composition pipeline* uses chemical descriptors, each feeding a committee of E(3)NN or GBDT. Concurrently, the *screening pipeline* (orange box) trains GBDT classifiers — "AI-experts" — to identify M^E-like materials. In prediction mode, screened GNoME materials (y) with over 50% likelihood ($P(y \in M) > 0.5$) of matching M^E materials' (x) biases enter the regressor pipeline to predict properties. These candidates, with predicted properties, are added to the Energy-GNoME database, initiating a continuous active learning cycle (see magenta arrow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

predict the property related to the specialized energy material under investigation only for the screened materials. In this way, we have the computational benefit of working with a significantly smaller dataset than the entire GNoME database, reducing computational costs and resources.

Finally, the candidates with the predicted properties are stored in the Energy-GNoME database. The resulting database can then be evaluated, refined, and validated by both the computational and experimental community, expanding the initial specialized energy material database used to train the workflow. Consequently, the entire workflow can be rerun, thereby improving both the screening and prediction accuracy, initiating an iterative process that makes the Energy-GNoME a *living database*.

2.2. Protocol verification

Before applying the protocol to the GNoME database, we first verify its effectiveness by assessing its performance on completely unseen materials. Three databases are used: the Experimentally Synthesized Thermoelectric Materials (ESTM) database [45], M^E; an unspecialized database built by randomly sampling from the MP database and removing any overlaps with ESTM, $MP \setminus M^E$; and the verification database, Y, built from the other two by removing 34 materials from M^E and 196 from $MP \setminus M^E$. Including all their temperature-dependent variations, the 34 unique ESTM materials result in 196 entries, while the 196 unique MP materials (each repeated at six different temperatures) result in a total of 1176 entries. Importantly, all these materials are entirely excluded from the datasets used to train, validate, and test the models, ensuring they are never exposed to the classifiers or regressors during any stage of model development.

The verification workflow is illustrated in Fig. 2a: after training, the AI-experts classify the materials in the Y database as thermoelectric candidates or not. Remarkably, all 196 entries from the ESTM are correctly classified (Blind True Positive, BTP), with an average classification probability $P > 0.95$ across the 10 classifiers. For these materials, the screening is followed by the prediction of the thermoelectric figure of merit zT. As shown in Fig. 2b, the regressors achieve strong predictive accuracy ($R^2 = 0.84$).

Notably, 5 MP unique materials (namely Cu_2GeHgS_4, Bi_3Pd_8, $AsCu_3S_3$, $NbNiTe_5$, and $NdNiSb_3$) escape the screening and are classified as thermoelectric candidates (Blind False Positive, BFP, $P > 0.5$). However, due to the generic nature of MP, some of these may be thermoelectric materials that have not yet been experimentally investigated. This result is further interpreted using the Euler–Venn diagram in Fig. 2c.

Here, the specialized dataset M^E represents experimentally known thermoelectrics, while the unspecialized dataset $MP \setminus M^E$ consists of crystals that are mostly non-thermoelectric although we cannot exclude *a priori* that a few thermoelectric materials could also be present. The set T represents the broader space of all thermoelectric materials, both known and unknown, such that $M^E \subset T$. The intersection of the validation database Y with T identifies the screening outcomes: BTP, BFP, and BTN (Blind True Negative). Indeed, Cu_2GeHgS_4 (whose crystal structure is depicted in Fig. 2c) has already been recognized for its thermoelectric properties, as reported by Gorai et al. [46].

Interestingly, this suggests that the protocol is able to capture materials with latent thermoelectric potential that are not explicitly labeled in ESTM.

These results serve as a critical sanity check, confirming that the protocol effectively identifies thermoelectric materials it has never encountered before while maintaining a low false-positive rate. This reinforces the robustness of the screening process and its applicability to discovering new thermoelectric candidates.

2.3. Thermoelectrics (figure of merit zT)

When an electrical current is supplied, thermoelectric materials are able to generate a temperature gradient while also releasing Joule heat;

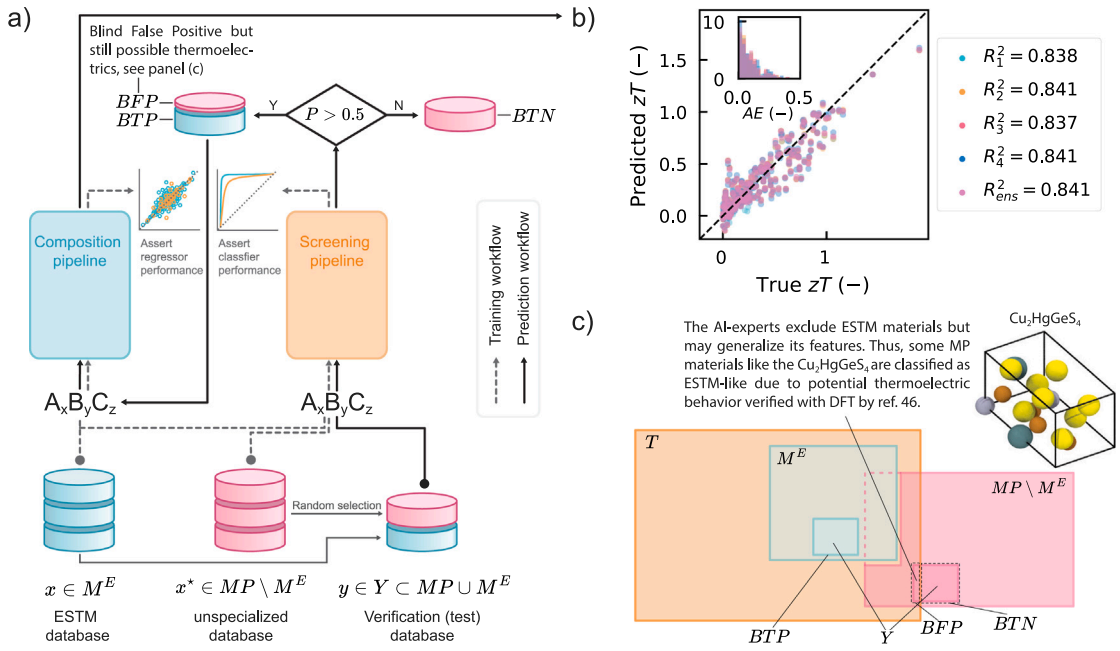


Fig. 2. The schematics in (a) illustrate the training (gray dashed line) and predictive (black solid line) phases during the verification of the hypothesis. Three databases were used: the ESTM database, M^E (in cyan); an unspecialized database, $MP \setminus M^E$ (in magenta); and the verification database, Y. The screening process successfully recognizes all the ESTM materials in Y (BTP), and the performance of the regressors (4 models + ensemble) on these 34 materials (196 entries) is reported in the parity plot in (b), along with the absolute error AE distribution density in the top left inset. The Euler–Venn diagram in (c) illustrates classification outcomes explication. The specialized database M^E and the unspecialized database $MP \setminus M^E$ are shown alongside the set T (orange) representing all thermoelectric materials. The small sets in M^E and $MP \setminus M^E$ correspond to the materials in the validation database Y, and their intersections with T identify the screening outcomes BTP, BFP, and BTN. The crystal in (c), which belongs to the BFP set, is a thermoelectric according to Gorai et al. [46]. Atom colors follow the extended CPK [47] scheme by Jmol [48]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

vice versa, a temperature gradient can generate an open-circuit voltage, allowing these materials to play the role of electric generators [49]. As a result, thermoelectric-based devices may utilize a range of heat sources, such as solar radiation and industrial waste heat, making them an interesting asset for the advancement of sustainable and energy-efficient technologies. The effectiveness of a material in thermoelectric systems is determined by the dimensionless thermoelectric figure of merit $zT = (S^2\sigma/\kappa)T$, with S denoting the Seebeck coefficient, σ the electrical conductivity, κ the thermal conductivity, all varying with the temperature T.

Here, we aim at finding new potential thermoelectric materials within the GNoME database following the same aforementioned protocol. As data source, we use the ESTM database [45], which contains experimental zT values of 869 unique materials measured at different temperatures T. In this case only the composition is available and, as such, we extract a set of 145 composition-based features for those 869 brute formulae by means of Matminer [50]. Specifically, as detailed by Ward et al. [51], these descriptors include stoichiometric features, statistics on elemental properties, characteristics related to electronic structure, and specific attributes for ionic compounds. Furthermore, since on average each material in ESTM comes with 6 distinct values of the temperature T — in general different across the database — we make T act as the 146th feature. However, to prevent potential unfairness that may arise from having the same material — only with a different T — over various random splits across training/testing sets, we ensure that all instances of the same material (coming with various T values) are included in the same split. In order to take into account the 146th feature for materials not coming from ESTM (i.e., MP), for each of them we create 6 replicas, each with one of 6 evenly spaced temperatures T, namely 300 K, 430 K, 560 K, 690 K, 820 K, 950 K.

To further assess the reliability of our workflow, 34 materials from ESTM are excluded from training and testing and instead reserved for an independent verification step of the entire protocol. As such, those materials are not seen by the models at any stage before the final predictions, ensuring a rigorous evaluation of the protocol's ability to generalize beyond its training data. On average, the 4 regression models show reasonably high performance in test ($R^2 \approx 0.80$, see Table 1) and the 10 classifiers turn out to be highly skilled as well (AUC $\lesssim 1$). The metrics of the individual regression and classification models are reported in Table S1 and Table S2 respectively. For further details about pre-processing and featurization, ML models training, and the whole protocol, refer to Section 4.

By incorporating T as an additional feature, we predict the average classification probabilities and the average zT values for each of the energy material candidates within the GNoME database, across 6 replicas at the same 6 evenly spaced temperatures mentioned above. As a result, we identify 13,069 unique GNoME compositions, corresponding to 50,779 T-based replica samples, showing an average probability $P > 0.5$ to fall within the materials space of the ESTM database. Among those, at $T = 950$ K, $SrBi_3Te_{10}$ is predicted to exhibit an average $zT = 1.34$ (over the 4 regressors) with an average classification probability (over the 10 classifiers) $P = 1.00$; $BaBi_3Te_{10}$ is predicted to exhibit an average $zT = 1.35$ (over the 4 regressors) with an average classification probability (over the 10 classifiers) $P = 1.00$; Ag_8Se_3S is predicted to exhibit an average $zT = 1.37$ (over the 4 regressors) with an average classification probability (over the 10 classifiers) $P = 0.99$ (see Fig. 3).

2.4. Perovskites (band gap E_g)

Perovskite solar cells have gained extensive popularity due to their high absorption coefficient, high charge carrier mobility, controllable band gap, and ease and low cost of fabrication [52,53]. In fact, the suitability of a perovskite as photovoltaic material is most importantly determined by its band gap E_g. It is worth noting that the possibility of engineering synthetic perovskites gives rise to a vast compositional

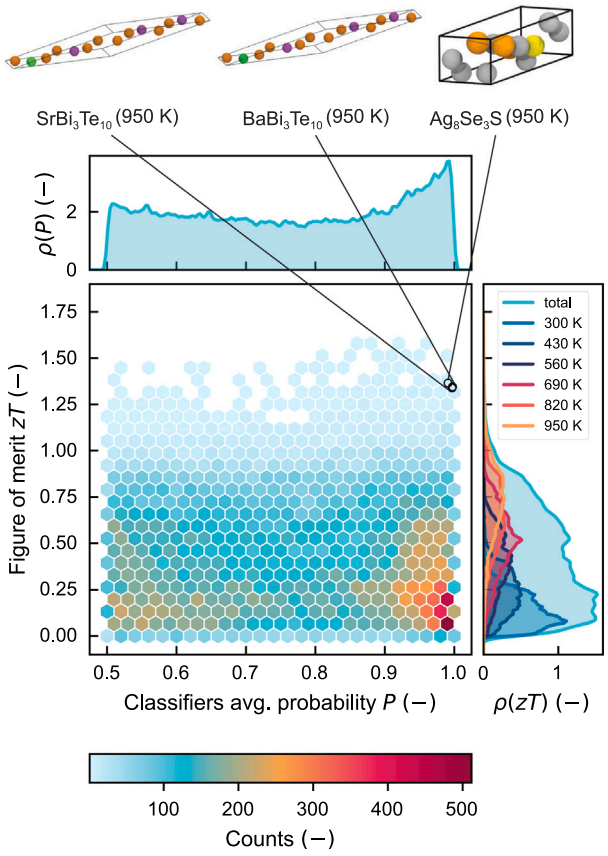


Fig. 3. Hexagonal plot of thermoelectric candidate materials in the Energy-GNoME database. The hexagon colors represent material counts per region as indicated by the color bar. Density distributions, ρ, are shown on the plot's top and right, calculated using Gaussian KDE for the average AI-expert probability, P, and predicted figure of merit, zT. The thermoelectric performance was assessed across six temperatures, with combined zT values displayed in a color-coded distribution on the right. The crystal structures above show three notable candidates among the top-ranked screened thermoelectric materials as determined by $R^T(y)$ (see Section 4.6.1). Atom colors follow the extended CPK [47] scheme by Jmol [48]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
The table reports the total number of thermoelectrics ($|M^E|$, with $|\cdot|$ denoting the cardinality) and a summary of regressor models testing performance (coefficient of determination R^2 and root mean square error RMSE) for the committee of 4 GBDT (ensemble). The models predict the thermoelectric figure of merit zT.

$\lvert M^E\rvert$ (−)	R^2 (−)	RMSE (−)
835 (4865 replicas)	0.800	0.150

materials space available for exploration, allowing researchers to fine-tune their properties for specific applications [54]. Hence, a reliable and cheap methodology to determine potential interesting structures would be a strong tool to investigate such space.

In this case study, we apply our protocol to identify new potential perovskite materials suitable for PV applications within the GNoME database. As a primary data source, we leverage the Materials Project (MP) database [25], which provides the structure along with the computed properties of more than 150,000 solid-state materials, of which ~23,000 where experimentally synthesized. From the MP database, we extract 641 unique perovskites with structures exhibiting a band gap

Table 2
The table reports the total number of specialized materials ($|M^E|$) and a summary of regressor models testing performance (coefficient of determination R^2 and root mean square error RMSE) for the committee of 4 E(3)NNs (ensemble). The models predict one key perovskite property for PV applications, namely the band gap E_g.

	$\lvert M^E\rvert$ (−)	R^2 (−)	RMSE (eV)
Pure	641	0.65	0.422
Mixed	81,895	0.90	0.496
Mixed[a]	81,895	0.78	0.312

[a] For the mixed models, metrics are provided for the entire testing set as well as for the domain of interest (non-magnetic perovskites with $1 \times 10^{-3}\,\text{eV} < E_g \leq 2.5\,\text{eV}$).

E_g within the range $1 \times 10^{-3}\,\text{eV} < E_g \leq 2.5\,\text{eV}$ and classified as non-magnetic, making them promising candidates for PV applications [55–57]. This choice is not intended to imply that materials with extremely narrow band gaps (e.g., $< 0.5\,\text{eV}$) are optimal for photovoltaics, but rather to conservatively avoid the exclusion of potentially promising candidates due to possible DFT-related underestimations of band gaps in large databases such as the Materials Project. The lower bound of $1{\times}10^{-3}$ eV serves as a soft threshold to filter out metallic materials while retaining semiconductors that may be inaccurately characterized. Subsequent regression and ranking steps in our protocol further prioritize candidates with more suitable band gap values (e.g., near the optimal photovoltaic range of 1.1–1.6 eV). Using this dataset, we train and test four regression models (hereafter referred to as the "pure models") to predict the band gap E_g. Taking advantage of the MP database's detailed crystal structure information, we utilize an Euclidean Equivariant Neural Network (E(3)NN) architecture, following the structure-based prediction pipeline (Fig. 1) described in the Methods (see Section 4).

As the E_g is a property shared by all materials, to improve generalization, we expand the training dataset by incorporating the entire set of non-metallic materials from MP, without restricting band gap or magnetic properties. This broader dataset, which includes both perovskites and non-perovskites, is used to train an additional set of four E(3)NN regression models (hereafter referred to as the "mixed models"), enabling the prediction of E_g across a wider chemical and structural space. Additionally, we identify a subset of the 641 perovskites containing 576 samples with unique chemical formula, and select 620 random non-perovskite materials, from which we extract a set of 694 composition-based and structure-based features by means of Matminer [50]. For further details, refer to Section 4.2. These features are used to train and test 10 classifiers (AI-experts) able to discriminate perovskites from other materials in the MP database.

On average, the pure and mixed regressors exhibit different levels of predictive performance in testing, with the pure models showing reasonable accuracy (coefficient of determination $R^2_{\text{pure}} \approx 0.65$ and the mixed models achieving strong predictive capability $R^2_{\text{mixed}} \approx 0.78$, see Table 2). This improvement comes at a substantial computational cost. Training the mixed models is considerably more resource-intensive, requiring approximately 10 min per epoch, compared to just 5 s per epoch for the pure models. This trade-off highlights the balance between accuracy and efficiency when scaling the training dataset. The metrics of the individual pure and mixed regressors are reported in Table S3 and Table S4, respectively. The AI-experts are highly skilled (AUC ≈ 0.98). The metrics of the individual classification models are reported in Table S5. For further details about pre-processing and featurization, ML models training, and the whole protocol, refer to Section 4.

As a result, we identify 4259 GNoME materials showing an average probability $P > 0.5$ to fall within the materials space of the perovskites included in the MP database. Fig. 4 shows the potential perovskite candidates in the Energy-GNoME database, along with the classifier committee's average probability and the average predictions of the individual E_g values, obtained through either the pure or the mixed regressor models.

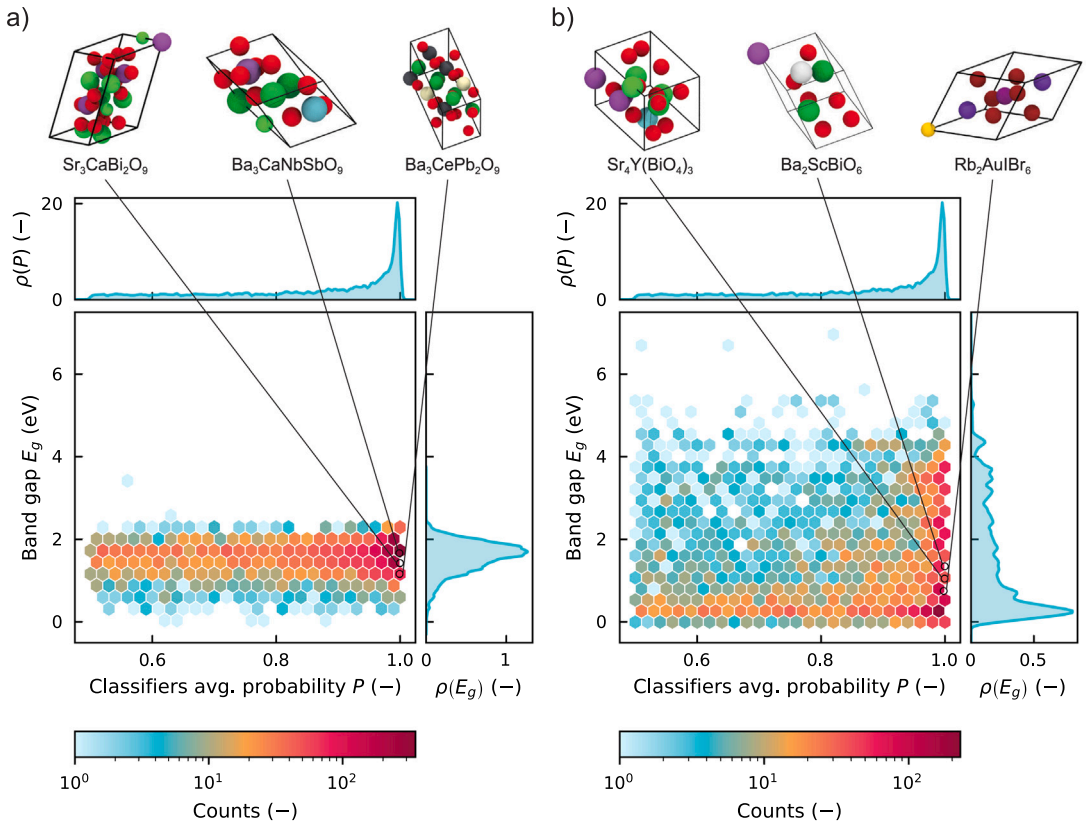


Fig. 4. Hexagonal plot of perovskite candidates materials in the Energy-GNoME database. The hexagon colors represent material counts per region on a logarithmic scale, as indicated by the color bar. Density distributions, ρ, are shown on the plot's top and right, calculated using Gaussian KDE for the average AI-expert probability, P, and the predicted band gap, E_g. For E_g, results are displayed from regressors trained on (a) perovskite data alone and (b) an augmented dataset. The crystal structures above show three notable candidates among the top-ranked perovskite materials, as determined by $R^P(y)$ (see Section 4.6.2). Atom colors follow the extended CPK [47] scheme by Jmol [48]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Among the most promising candidates according to the "pure models" predictions and the ranking function Eq. (5) (described in Section 4.6.2), interesting materials are $Sr_3CaBi_2O_9$, $Ba_3CaNbSbO_9$, and $Ba_3CePb_2O_9$. In particular, the regressor committee predicts $Sr_3CaBi_2O_9$ to exhibit an average $E_g = 0.57$ eV, $Ba_3CaNbSbO_9$ to exhibit an average $E_g = 2.49$ eV, and $Ba_3CePb_2O_9$ to exhibit an average $E_g = 0.65$ eV. All three materials have an average classification probability (over the 10 AI-experts) $P = 1.00$.

On the other hand, among the most promising candidates according to the "mixed models" predictions and the ranking function Eq. (5), interesting materials are $Sr_4Y(BiO_4)_3$, Ba_2ScBiO_6, and Rb_2AuIBr_6. The regressor committee predicts $Sr_4Y(BiO_4)_3$ to exhibit an average $E_g = 1.05$ eV, Ba_2ScBiO_6 to exhibit an average $E_g = 1.35$ eV, and Rb_2AuIBr_6 to exhibit an average $E_g = 0.75$ eV. All three materials have an average classification probability (over the 10 AI-experts) $P = 1.00$.

It is worth noticing that a proper value of the band gap only represents a necessary condition for a material to be a potential candidate in PV applications. Additional screening based on other relevant properties (e.g. optical absorption), aiming at further narrowing the set of potential new materials, is beyond the scope of this work and can be pursued in the near future.

2.5. Cathodes (average voltage ΔV_c)

The search for new cathode materials is a critical focus in the electrochemical community as the demand for next-generation batteries continues to rise. Indeed, the widespread use of batteries in electronic devices, electric vehicles, and for energy storage during RES's surplus production is driving the demand for new battery technologies that are safer, more reliable, cost-effective, and sustainable, even moving beyond the conventional Lithium-ion Battery (LIB) technology [58–60].

In this Subsection, we present the results of our ML-based screening protocol applied to the GNoME database using the "Battery Explore" specialized database from Materials Project [25]. This specialized database — at the time of this study — consisted of 3985 batteries, each comprising a pair of charge and discharge cathode materials and intermediate stable phases used to compute the average voltage potential. The database includes intercalation-type cathode materials designed for nine different monovalent and multivalent working ions: Li, Na, Mg, K, Ca, Cs, Al, Rb, and Y.

This case study aims to identify possible new cathode material candidates for lithium and *post-lithium* batteries within the GNoME database. Critical parameters present in the MP database for batteries include the average voltage (ΔV_c), maximum relative volume difference ($\max(\Delta\text{Vol})$), stability of the charged state (ΔE_{charge}), and stability of the discharged state ($\Delta E_{\text{discharge}}$). Following the protocol we are presenting, we trained four regressors. Similar to the perovskite case, we can also access the crystal structure of the materials by querying the MP database, therefore also here we use E(3)NN, following the structure-based pipeline of the dataflow (Fig. 1). However, instead of focusing on a single property, we repeated the training for all four target cathode properties, resulting in a total of 16 models. Furthermore, we tailored these 16 E(3)NN models for each of the nine working ions under consideration, culminating in a total of 144 trained models. For the AI-experts, we trained 10 classifiers for each working ion (totaling 90 classifiers). We used MP materials containing the working ion element,

Table 3
The table reports the total number of materials per working ion ($|M^E|$) and a summary of regressor models testing performance (coefficient of determination R^2 and root mean square error RMSE) for the committee of 4 E(3)NNs (ensemble). The models predict four key cathode properties: average cathode potential (ΔV_c), maximum relative volume difference ($\max(\Delta\text{Vol})$), stability of the charged state (ΔE_{charge}), and stability of the discharged state ($\Delta E_{\text{discharge}}$).

	$\lvert M^E\rvert$	ΔV_c		$\max(\Delta\text{Vol})$		ΔE_{charge}		$\Delta E_{\text{discharge}}$	
		R^2 (−)	RMSE (V)	R^2 (−)	RMSE (m^3/m^3)	R^2 (−)	RMSE (eV/atom)	R^2 (−)	RMSE (eV/atom)
Li	2440	0.733	0.563	0.442	0.030	0.315	0.041	0.263	0.034
Na	309	0.606	0.744	0.205	0.055	0.432	0.049	−0.112	0.034
Mg	423	0.619	0.876	0.335	0.040	0.491	0.069	0.669	0.091
K	107	−0.173	1.023	0.440	0.109	0.545	0.028	−0.115	0.092
Ca	435	0.695	0.654	0.550	0.043	0.477	0.076	0.276	0.072
Cs	33	0.594	0.865	−0.246	0.094	0.082	0.017	−0.199	0.033
Al	95	0.783	0.571	−0.186	0.059	0.335	0.158	0.596	0.076
Rb	50	−0.944	2.196	0.040	0.192	−0.079	0.065	−0.335	0.102
Y	93	0.365	0.757	0.488	0.096	0.493	0.122	0.600	0.092

Table 4
The table reports the total number of candidate cathodes identified by AI-experts within GNoME per working ion ($|G^E|$) and a summary of classification testing performance (AUC of Receiver Operating Characteristic (ROC) and Precision of the classifiers) for the committee of 10 GBDT AI-experts (ensemble).

	AUC (−)	Precision (−)	$\lvert G^E\rvert$
Li	0.897	0.764	413
Na	0.890	0.811	779
Mg	0.946	0.864	888
K	0.844	0.706	1327
Ca	0.917	0.791	1128
Cs	0.458	0.400	4634
Al	0.887	0.857	6280
Rb	0.886	0.857	3559
Y	0.833	0.733	2235

but not overlapping with the training database, as a *less biased* dataset. The two datasets, representing the two classes for the classification are then translated into the materials space using descriptors made of 694 composition-based and structure-based features by means of Matminer. A detailed explanation of pre-processing and featurization of all ML models, training and testing settings, and the whole protocol layout is provided in Section 4.

Table 3 reports the total amount of data available in MP for training and the relative ensemble of E(3)NNs models performance for each cathode class evaluated in the testing set (20% of the specialized database M^E). The results for each one of the 16 models are reported in the supplementary Tables S6-S9. The performance of each regressor varies across working ions due to differences in dataset size and years of investigation for each cathode type. Li cathodes are the most prevalent in the MP database, with 2,440 crystal materials. For these, the committee of models shows that the average voltage is the property best predicted by the E(3)NN model, with $R^2 = 0.733$. In contrast, the stability energies ΔE_{charge} and $\Delta E_{\text{discharge}}$ show lower performance, with $R^2 = 0.315$ and $R^2 = 0.263$, respectively. These results highlight that, despite E(3)NN's proven ability to predict energy, forces [61], and phonon density of states [62], a large energy material database is crucial. Therefore, the predictions presented here should be interpreted with caution, as further expansion of the cathode material database is necessary to enhance model performance. The worst cases, which seem extremely hard to interpolate, are K, Cs, Al, Rb, and Y cathodes due to their small training set sizes (<150).

On the other hand, despite the small sample size, the classifiers demonstrate a higher skill in identifying materials belonging to the energy materials space than the more general MP database. Table 4 reports the AUC and precision metrics on the testing set for each working ion specialized AI-expert ensemble. Most AI-expert ensembles achieve an AUC > 0.85, a clear sign of their robust discriminative capability. However, the cathodes for K, Cs, and Y ion batteries present the most challenging sets of materials for the AI-experts to classify. Detailed metric scores for all 90 classifiers on the test and training sets are reported in Table S10 in the supplementary information.

As a result of the screening process, we identify 21,243 GNoME materials with an average probability $P > 0.5$ belonging to the cathode materials space. Table 4 shows the number of materials, $|G^E|$, for each specific monovalent and multivalent ion battery. An interesting result is the inverse proportionality between the number of screened materials and the training set size. This may indicate that when the protocol is applied to deeply investigate cathode materials, the AI-experts are more skilled in finding biases in the database, narrowing the boundaries of the cathode materials space. Consequently, the false positive rate when working with small training sets is likely higher compared to cathodes for Li, Na, and Mg ion batteries. Another possible interpretation is that, in relatively large datasets indicating deep historical investigation, the remaining unexplored materials are fewer than in the case of new *post-lithium* cathodes, which explains the numerous candidates found for these classes.

Regarding the protocol's performance for this case study, we must consider the empirical evidence showing that the accuracy of ML models increases with the size and variety of the training set, independent of the model architecture's complexity [63]. Therefore, the materials science community can numerically or experimentally validate the possible candidates identified by the AI-experts and enrich the starting specialized materials database. As depicted in the protocol Fig. 1, these initial results represent just the starting point. The next round of "active learning" will likely enhance the accuracy and precision of both the regressor and classifier models.

In Fig. 5, we report all the new potential cathodes for the nine working ions. In addition to the four properties predicted by the E(3)NN committee, we compute additional properties for the pure cathode material that can aid in decision-making for further investigation. Namely, we compute the gravimetric and volumetric capacities and energies based on the E(3)NN predictions, the primitive unit cell, and the number of active elements within the cell. It is important to note that here, the terms gravimetric and volumetric capacities refer to the mass and volume of the cathode material alone, so as not to be confused with the volumetric and gravimetric properties of the entire battery system (which includes the cathode, anode, electrolyte, etc.). We compute the gravimetric capacity using the following equation:

$$q_g = \frac{F}{3.6} \cdot \frac{\sum_{i=X} q_i}{\sum_i m_i}, \tag{1}$$

where q_g is the gravimetric capacity in mA h g^{-1}, F is the Faraday constant in C mol^{-1}, m_i is the molar mass of each atom inside the material unit cell in g mol^{-1}, and q_i is the charge carried by the working ion element inside the crystal unit, estimated using the most probable oxidation state of the atom using the bond valence sum method as

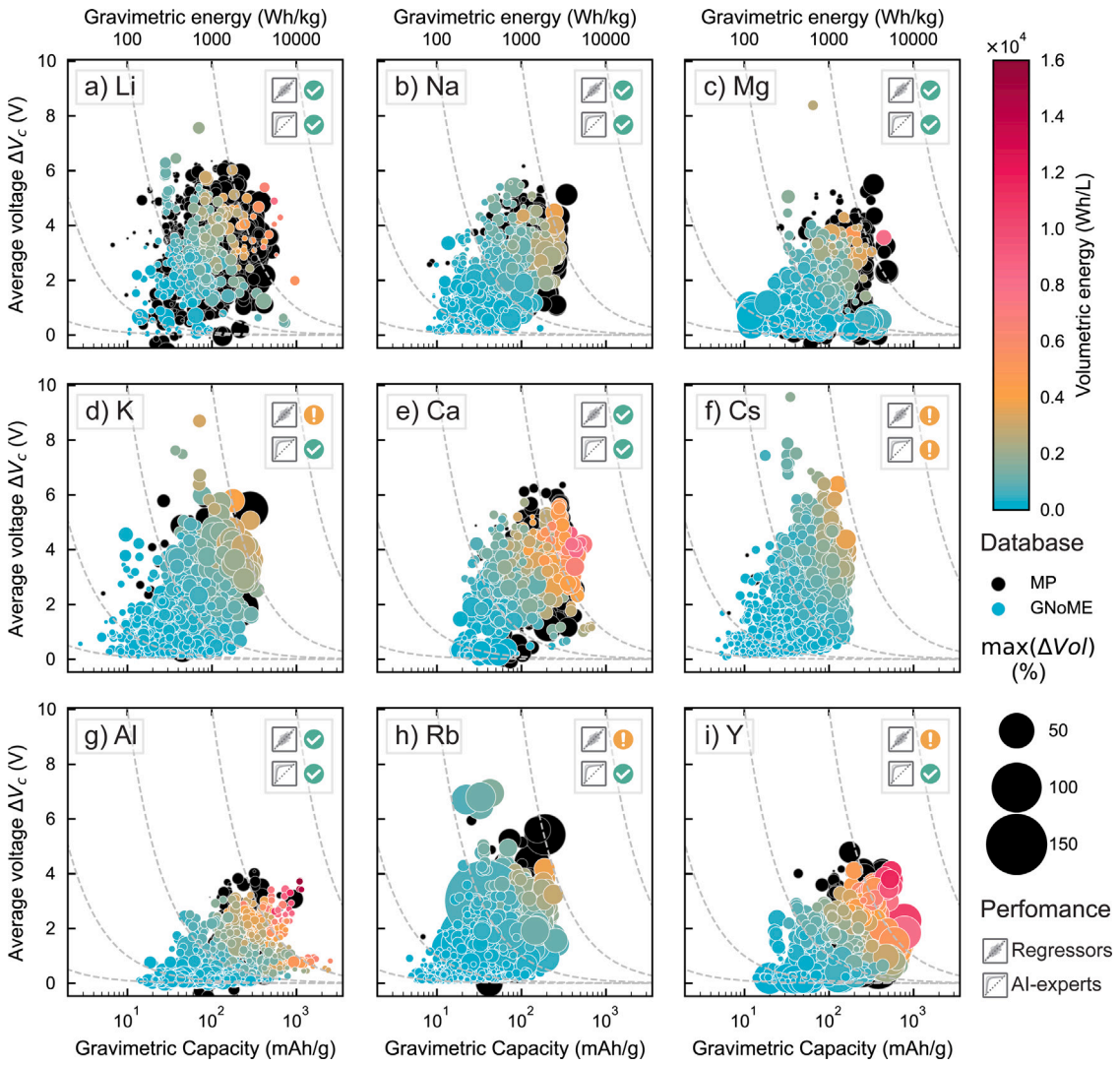


Fig. 5. Candidates for battery cathode materials with various charge carriers. Li (a), Na (b), Mg (c), K (d), Ca (e), Cs (f), Al (g), Rb (h), and Y (i). Each candidate is represented as a point in the scatter plot, showing theoretical gravimetric capacity ($mAh\,g^{-1}$, logarithmic scale) versus predicted average voltage difference (V) relative to pure element oxidation potential (X/X^{n+} with X being the working ion). Gray dashed hyperbolas indicate the predicted gravimetric energy ($Wh\,kg^{-1}$), noted on the upper axis. Dot size represents predicted maximum volume expansion, and the dots are color-coded to represent the predicted volumetric energy ($Wh\,L^{-1}$). Due to dataset limitations, model performance varies (see Section 2.5). The top right corner legends indicate prediction reliability: a green checkmark (✓) for models with R^2 and AUC above 0.5, showing higher accuracy, and a yellow warning (!) for models below this threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

detailed in Methods Section 4.6.3. The factor 3.6 in Eq. (1) serves as a unit conversion constant, transforming $A\,s\,g^{-1}$ to $mA\,h\,g^{-1}$. Similarly, we compute the volumetric capacity as follows:

$$q_v = \frac{F}{3.6\times10^{-3}} \cdot \frac{\sum_{i=X} q_i}{|\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})|}, \tag{2}$$

where q_v is the volumetric capacity in $mA\,h\,L^{-1}$, F is the Faraday constant in $C\,mol^{-1}$, **a**, **b**, and **c** are the crystallographic axis vectors with norms measured in m, and q_i is the charge carried by the working ion element inside the crystal unit as in Eq. (1). The factor 3.6×10^{-3} in Eq. (2) serves as a unit conversion constant, converting As/m^3 to $mA\,h\,L^{-1}$. The relative gravimetric and volumetric energy is then computed using the predicted average voltage ΔV_c:

$$E_g = \Delta V_c \cdot q_g, \quad E_v = \frac{\Delta V_c \cdot q_v}{1000}, \tag{3}$$

where E_g is the gravimetric energy in $Wh\,kg^{-1}$, q_g is the gravimetric capacity computed as in Eq. (1), E_v is the volumetric energy in $Wh\,L^{-1}$, and q_v is the volumetric capacity computed as in Eq. (2). The division E_v by 1000 converts the units from $mWh\,L^{-1}$ to $Wh\,L^{-1}$ since the volumetric capacity q_v is measured in $mA\,h\,L^{-1}$.

Moving our analysis to Li and Na ion batteries, we report in Fig. 6 the distribution of all the Li and Na potential cathodes with respect to the classifier committee's average probability P and the average predicted voltage ΔV_c. Among the candidates found using the ranking function Eq. (6), we identify $Li_{12}ZnCoO_8$, $LiFePO_4$ and Li_5CoF_8. $Li_{12}ZnCoO_8$ has a predicted voltage of 1.99 V and shows an extremely high gravimetric capacity of $958\,mA\,h\,g^{-1}$, which exceeds the highest capacity commercially available cathode $LiMnO_2$ ($285\,mA\,h\,g^{-1}$) and is below the under-investigation Li_2S cathode [38] with the theoretical gravimetric capacity of $1675\,mA\,h\,g^{-1}$. It shows a layered structure common to other cobalt oxide cathodes like $LiCoO_2$. The model also identifies Li_5CoF_8 as a possible LIB cathode. According to the E(3)NN models, the average voltage is 4.89 V, resulting in a theoretical specific energy extremely high at $2670\,Wh\,kg^{-1}$. However, from the crystal structure, the positions of the lithium ions do not show a fully layered structure, and the high voltage makes it impractical since it exceeds the electrochemical window of common electrolyte compositions (~4.5 V

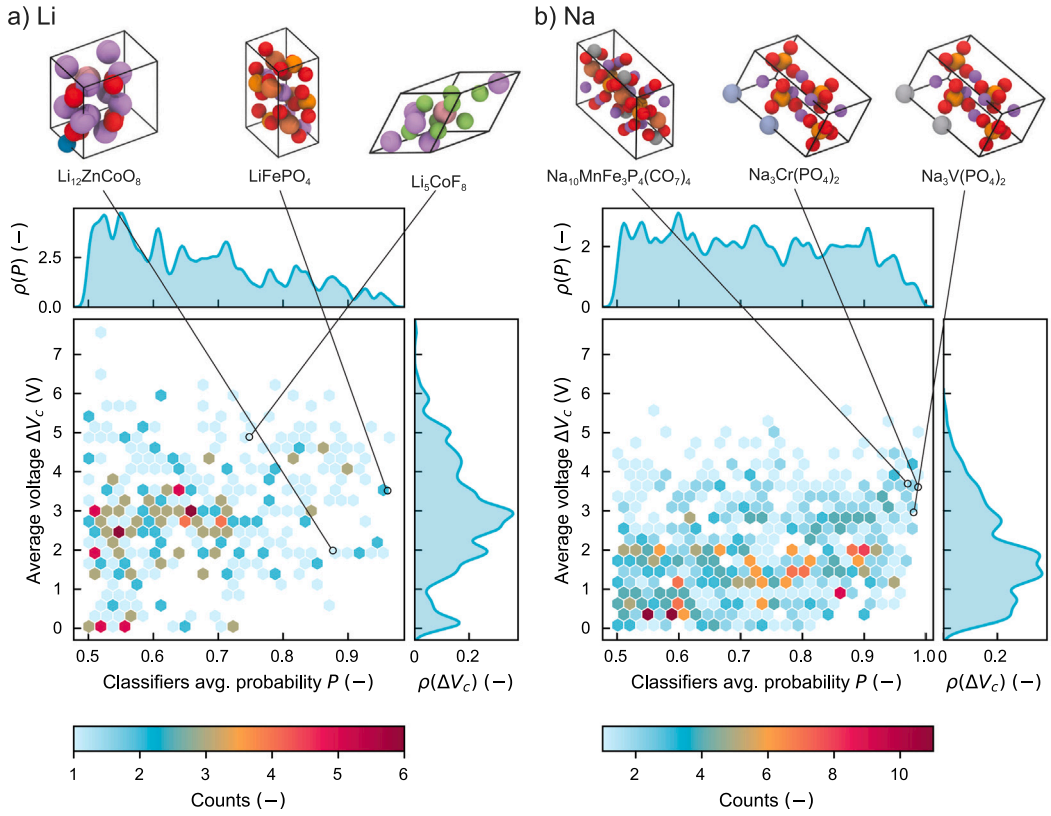


Fig. 6. Hexagonal plot of cathode candidate materials in the Energy-GNoME database for (a) Li-ion and (b) Na-ion batteries. The hexagon colors represent material counts per region, as indicated by the color bar. Density distributions, ρ, are shown on the plot's top and right, calculated using Gaussian KDE for the average AI-expert probability, P, and the average reduction potential, ΔV. Three high-ranking screened cathode materials are shown as primitive crystal units above, identified using $R^B(y)$ (see Section 4.6.3). Atom colors follow the extended CPK [47] scheme by Jmol [48]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

upper limit for Fluoroethylene-carbonate (FEC) and $LiPF_6$ electrolyte for high-voltage cathodes [64]). This incomplete compliance with the requirements for a good LIB cathode is also predicted by the AI-experts' average probability $P = 0.75$ (i.e., there is a 25% chance that this is not a suitable cathode). A peculiar result is the identification and presence of $LiFePO_4$ in GNoME, which is already present in the MP database and has been experimentally investigated, explaining the high AI-experts' average probability of $P = 0.96$. Upon examining the data in the crystal files of GNoME and the MP database, the only difference between the two is a Euclidean rotation, which explains the good E(3)NN prediction of the expected voltage 3.52 V (3.79 V for the cathode in the MP database).

For the Na cathode candidates, using the same ranking function, we identify $Na_{10}MnFe_3P_4(CO_7)_4$, $Na_3Cr(PO_4)_2$, and $Na_3V(PO_4)_2$. $Na_{10}MnFe_3P_4(CO_7)_4$ has a predicted voltage of 3.70 V and shows good gravimetric capacity of 250 mA h g^{-1} and energy density of 924 W h kg^{-1}. It shows a typical olivine structure with Na aligned in channels. Many Na-ion battery cathode materials show similar transition metal and polyanionic frameworks, such as manganese and iron combined with phosphate, like the Na super ionic conductor (NASICON) type $Na_4Fe_3(PO_4)2(P_2O_7)$ cathode [65]. This explains the high probability that the AI-experts associate with it ($P = 0.97$). However, the presence of the carbon-oxygen groups CO_7 is quite exotic, and to our knowledge, has never been observed in Na cathodes. The other two top-ranked materials, $Na_3V(PO_4)_2$ and $Na_3Cr(PO_4)_2$, are extremely similar, where the V in the first is substituted with Cr, which is next in the periodic table. Indeed, they show very close gravimetric capacities of 259.5 mA h g^{-1} and 258.6 mA h g^{-1}, respectively. The E(3)NN committee predicted different average voltages: 2.96 V for $Na_3V(PO_4)_2$ and 3.61 V for $Na_3Cr(PO_4)_2$. The $Na_3V(PO_4)_2$ is also similar to an already under-investigation cathode, the NASICON-type polyanion sodium vanadium phosphate (NVP) $Na_3V_2(PO_4)_3$ [66], which explains the very high probability over the 10 classifiers ($P = 0.98$).

3. Discussion

The AI protocol presented in this work demonstrates a computationally highly efficient approach for screening vast unexplored material databases such as GNoME to identify promising candidates for energy applications.

As it has been recently demonstrated in Ref. [44], specialized datasets can be affected by detrimental biases due to human investigation history: here we apply this idea to known energy-related materials, such as thermoelectric materials, perovskites, and electrochemical battery cathodes. In fact, it is reasonable to assume that the materials published in the literature and later included in specialized databases were not randomly selected and tested (either numerically or experimentally) over time, but rather carefully chosen based on prior knowledge so as to increase the likelihood that such materials would exhibit high performance. As a consequence, this accurate selection of materials leads to an uneven and non-uniform representation of the materials space within a given database, ultimately resulting in an anthropogenic bias [67]. These biases are invisible in a low-dimensional space like the measurable property space of materials, but we show that they can be identified and exploited by training skilled AI-expert classifiers. Such classifiers operate in high-dimensional feature spaces for screening. Indeed, by leveraging complex feature interactions rather than simple acceptance criteria applied to individual features, our method provides a more comprehensive screening process. This is

particularly evident when compared to previous approaches, such as the one used by Cerqueira et al. [34] for superconductors.

In this work, we trained a committee of 10 classifiers, named AI-experts, that learn the biases of the training set and replace human experience in selecting possible energy material candidates. Additionally, we trained a committee of 4 decision trees or equivariant neural networks, depending on the database quality, to predict relevant properties for energy materials. These properties include figure of merit for thermoelectric materials, band gap for perovskites, and average voltage for cathode materials in batteries.

The first key advantage is that, given the poor extrapolation capability of ML tools, the property predictions are expected to be more robust and reliable within the feature space shared with the training data.

Another important advantage of this method is its efficiency in narrowing down the candidate pool size. This targeted approach allows for a more efficient exploration of the energy materials space, which can potentially save a significant amount of time and resources during experimental and numerical validation.

However, a limitation remains in the inability to explicitly measure the false-positive rate for the identified materials, as no direct method for *a priori* quantification has been integrated into the current protocol. While this limitation does not diminish the overall effectiveness of the protocol, it highlights an area for potential improvement. For instance, one way to tackle this issue could be having an additional screening step with a specialized AI-expert (classifier) employed to screen over other essential properties of materials like "synthesizability" and "(eco)toxicity" of novel materials. At present, these properties are only implicitly embedded in the AI-expert score rather than being explicitly quantified.

We see this work as the first step of a continuous community effort. Further improvements and developments are expected in the future with the very likely discovery of new stable materials and knowledge advancement on the properties of known materials. Along this direction, a natural next step for improvement lies in expanding the training dataset. While in the current work the accuracy of each trained model is not the main focus, we do expect that, as more materials become experimentally validated and incorporated into the training set, the accuracy and precision of both the classifiers and regressors will improve. This planned refinement will enhance the feature space, leading to more accurate and narrowly focused predictions. Thus, this protocol calls for cooperative "active learning" by the community, which will accelerate the discovery of new and high-performing energy materials.

Finally, it is worth noting that — for the sake of simplicity and without a loss of generality of the described methods — in the current work we have predominantly focused on physical and chemical figures of merits. In future developments, we aim to expand the screening process to incorporate additional factors, such as the expected (eco)toxicity and sustainability. The latter could potentially be achieved through an additional classification step leveraging existing databases [68].

4. Methods

This Section details our heuristics-driven protocol for evaluating potential new materials for energy applications.

Our approach is motivated by recent advancements in materials discovery research while addressing critical limitations in applying ML in property prediction for materials.

The first advancement is increasing reliance on ML techniques to accelerate materials discovery and optimization processes, which has led to the publication of several specialized databases for various classes of materials for energy applications [18,25–29,69–71], such as the three used in this study. The second advancement stems from the publication of an active learning method by Merchant et al. [43] that facilitated the discovery of over 380,000 new stable crystal structures, culminating in the GNoME database.

Despite these advancements offering opportunities to identify new energy materials within the GNoME database, we face two significant challenges related to the limitations of current ML models in materials property prediction.

The first limitation involves the reliability of ML models when predicting properties in high-dimensional feature spaces. Many models encode the material's properties in these high-dimensional spaces, relying on regressors with thousands or even millions of degrees of freedom. Consequently, these models often act as a "black box", creating interpretability challenges and significant uncertainty when applied beyond their training domains. This issue, where ML models struggle to extrapolate, has been documented by Shimakawa et al. [72]. They empirically showed that ML tools perform well when interpolating within the boundaries of the training set but perform poorly when extrapolating beyond them. The second limitation is the inherent bias present in these databases used for training [44]. The source of these biases may be internal due to the material classes' nature or external due to human-driven selection processes. Indeed, most materials in these databases were discovered through selection processes guided by human knowledge and intuition, introducing biases into the datasets [44,73]. To address and mitigate these issues, under our assumptions, we propose applying regressors only to subsets of data points that share similar biases with the training set so that the model works inside the interpolated region [44].

To illustrate this hypothesis, we present the conceptual framework in Fig. 7, where we assume that the GNoME dataset (G), the Materials Project database (M), and a specific subset for energy applications (M^E) are being used. Typically, these datasets differ by orders of magnitude in size, with GNoME containing approximately 380,000 materials, MP containing around 34,000 stable materials, and energy-specific subsets usually comprising thousands of data points. The energy-specific subset $M^E = M \cap E$ represents the intersection of known materials from the MP dataset and a hypothetical set of all possible materials suited for a given energy application, expressed as E. This hypothetical set E is defined not only by the application-specific requirements but also by historical biases inherent in the materials discovery process, including factors such as synthesis feasibility, experimental limitations, and researchers' prior knowledge.

Our primary hypothesis is that the energy-relevant region in the materials space overlaps with the GNoME dataset, indicated as $G^E = G \cap E$. If we can accurately estimate the boundary of this subset (∂E), we can effectively screen materials in GNoME that have high potential for use in energy applications. This boundary estimation is crucial because it will make the regressor predictions more reliable and allow us to focus our computational resources on the most promising candidates within the vast GNoME dataset. Here, to implement this approach, we propose training a committee of classifiers — "AI-experts" — assigned with the task of distinguishing materials in the energy-specific subset (M^E) from the remaining part of the broader MP dataset ($M \setminus E = M \setminus M^E$). These classifiers aim to simulate the decision-making processes of human experts in the field, incorporating the knowledge-derived biases we hypothesize exist in the energy-related subset. Crucially, the distinction between these three main sets (G, M, and E) is challenging in low-dimensional spaces that can only be represented by few theoretical or experimentally measured properties of the crystal, as illustrated in Fig. 7a. Our second hypothesis is that the separation between these sets becomes more evident in high-dimensional feature spaces, where materials are represented by a more comprehensive set of descriptors, as shown in the ideal case displayed in Fig. 7b. These higher-dimensional spaces are more suitable for ML tasks, as they can potentially reveal patterns and relationships not evident in lower-dimensional representations [51,74].

Building on the overview of the entire protocol provided in Section 2.1, the following Sections will detail our methodology, including the technical explanation of the pre-processing and dataset featurization steps, the screening process, and the architecture and training of the regressor.

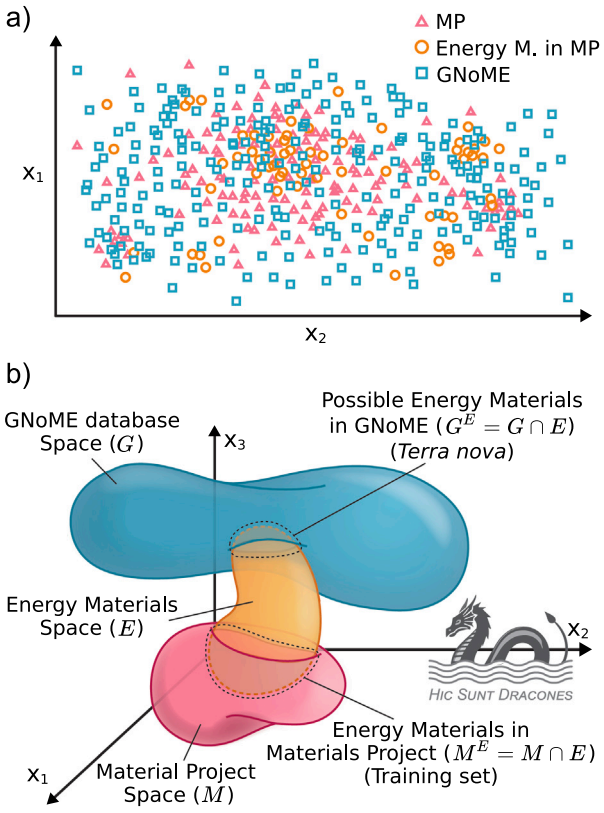


Fig. 7. Conceptual illustration of the method. In low-dimensional space (a), materials are plotted with experimental or chemical properties, creating scattered, often unclassifiable points. Here, we illustrate the hypothetical case with data from the MP (red triangles), energy-related MP data for battery or perovskite studies (orange circles), and GNoME data (green squares). In high-dimensional feature space (b), generated from chemical and structural descriptors, distinct n-dimensional regions emerge for MP (M) and GNoME (G) data. We hypothesize the existence of an orange region E, where all energy-related materials (e.g., cathodes, perovskites) reside. The AI-experts use the intersection $M \cap E$ and the remaining MP data, $M \setminus E$, to define the boundary of E. This enables the identification of $G \cap E$, the crystals in GNoME with similar properties to $M \cap E$, where regression is more reliable than when applied to the whole G set.

4.1. Data pre-processing

The data pre-processing employed in this work consists of building and cleaning the specialized energy material database.

Various sources were used for the three test cases: Materials Project and published literature databases, depending on the specific energy material class under investigation. All materials available in the MP database were fetched using its API (application programming interface) [75] through the python library `mp-api` to efficiently retrieve relevant information. In other cases, we obtain the database from published literature additional material or associated repositories.

All the specialized "energy" material databases were then cleaned to enhance data quality. This process encompassed handling missing values, standardizing units, and unifying file formats. Due to the diverse nature of the training databases used in this study, data cleaning procedures were specifically tailored to each dataset. Detailed descriptions of these case-specific procedures can be found in Section 4.6.

4.2. Materials featurization

Depending on the model used for property prediction, we have two possible ways to translate materials into machine-readable data. When the data flows through the *structure pipeline*, for regression tasks we perform a Structural Encoding (SE) process proposed by Chen et al. [62], which involves creating a periodic graph representation of the atomic structure of the crystal, with chemical information embedded in the graph using one-hot encoding. The graph representation consists of N_n nodes A_i, each representing an atom in the crystal unit, and N_e edges e_{ij} storing the interatomic distances r_{ij} between neighboring atoms within a cutoff radius r_{cut} set to 5 Å. Periodic Boundary Conditions (PBC) are applied to identify neighboring atoms, resulting in a periodic graph. Chemical information is then encoded at each node using one-hot representation of the atomic element and mass. This results in a node feature vector $\mathbf{x}_i = \{x_j\} \in \mathbb{R}^{118}$, with the jth element defined as $x_j = m_i \cdot \delta_{ij=Z_i}$, where δ_{ij} is the Kronecker delta, m_i is the atomic mass of atom A_i, and Z_i is the atomic number of A_i. For example, if the node A_i represents a Li atom, the corresponding vector would be $\mathbf{x}_i = \{0, 0, 6.94, 0, \ldots, 0\}^\top$. For classification tasks, we featurize each material with a descriptor vector $\mathbf{x}_i \in \mathbb{R}^{694}$, encompassing both composition- and structure-based features. Such composition-based descriptors include stoichiometric distribution moments, fractional presence of each element within the compound, average number of electrons in each orbital, features related to possible oxidation states, and elemental properties obtained from the Materials-Agnostic Platform for Informatics and Exploration (Magpie) database [51]. Conversely, the structure-based ones come from the Jarvis-ML descriptors [76], considering cell and chemical composition.

The database fed to the *composition pipeline* is featurized based on the chemical composition of the materials using the `matminer` Python library [50]. In this case, we rely for both regression and classification tasks only on 145 (146 considering the temperature) chemical composition-based features, including stoichiometric features, elemental property statistics, and electronic structure characteristics. Further details on both featurization processes can be found in Section 4.6.1

4.3. Regressors training

From the periodic graph representation of the crystal, we utilize a tailored version of the E(3)NN model originally proposed by Chen et al. [62] for predicting the phonon density of states (pDOS) across various crystal structures. This model initially applies a linear transformation to reduce the 118 node features to 64 embedded chemical features. The data then passes through two layers of graph "convolutions and gating" equivariant operations. The convolution kernel used is a product of learnable radial functions and spherical harmonics of the form:

$$K_m^{(l)}(\mathbf{r}_{ij}) = R(r_{ij}) Y_m^{(l)}(\hat{\mathbf{r}}_{ij})$$

where $\mathbf{r}_{ij}$ is the distance vector between ith and jth atoms, r_{ij} and $\hat{\mathbf{r}}_{ij} = \mathbf{r}_{ij}/r_{ij}$ its associated norm and direction vector. Therefore, a crucial set of hyperparameters includes the maximum order of the spherical harmonics, which was set to $l_{\max} = 3$, and the radial function defined as a fully connected neural network (FNN):

$$R(r_{ij}) = \sum_h W_{kh} \sigma \left(\sum_q W_{hq} B_q \left(r_{ij} \right) \right),$$

where W is the weight matrix of the input and hidden layer, B_q are the radial basis functions, and σ is the activation function. We use 10 equally distanced (from 0 Å to $r_{\text{cut}} = 5$ Å) Gaussian radial basis functions, 100 neurons for the hidden layer, and the Sigmoid Linear Unit (SiLU) as the activation function ($\sigma(x) = x(1 + e^{-x})^{-1}$). The data then passes through a final single graph convolution layer before summing all the resulting embedded features tensors from all atoms into a single one (sum-pooling). Then, the output passes through a Rectified Linear Unit (ReLU) activation layer, and the average of the ReLU outputs (mean-pooling) is the final scalar quantity. For additional details on the mathematics of the GNN, we invite the reader to refer to the original work by Chen et al. [62], the `e3nn` Python library articles and documentation [77,78], and related publications [61,79–81]. The

model is then trained using the AdamW optimizer [82], with an exponentially decaying learning rate and L_1-norm (Mean Absolute Error (MAE)) as loss function. The training behavior primarily depends on the initial learning rate η_0, the weight decay λ for the optimizer, and the exponential base of the learning rate decay β ($\eta_i = \eta_0 \beta^i$ for ith epoch). Specifically, for the cathode materials, we found optimal $\eta_0 = 1 \times 10^{-3}$, $\lambda = 0.1$, and $\beta = 0.99$; for the perovskites, we used instead $\eta_0 = 0.005$, $\lambda = 0.05$, and $\beta = 0.99$. It has been observed that, during training, the number of epochs required to reach the minimum of the loss function varies depending on the material class. For cathode materials, 100 epochs are sufficient for convergence. However, for both pure and mixed perovskite models, 500 epochs are necessary. This reflects the increased complexity of learning band gap predictions, whether specifically for perovskites or across a broader range of materials in the mixed models. We then select the model with the minimum loss on the test set.

For the composition-based regression models, we adopt the GBDT method [83]. Specifically, we set up a pipeline by means of the `GradientBoostingRegressor` object based on the python ML-library `scikit-learn` [84]. In particular, before the actual training, we perform reduction of data dimensionality by adopting Recursive Feature Elimination (RFE) [85]. The key hyperparameters of the resulting ML models are the number of boosting stages to perform (N_b) and the number of selected features to use after the RFE step (N_f). For each AI-expert training, we perform a k-fold (with $k = 4$ number of split datasets) cross-validated grid search on the hyperparameter space:

$$(N_b, N_f) \in \{50, 100, 250, 500\} \times \left\{ \frac{|M^E|}{40}, \frac{|M^E|}{20}, \frac{|M^E|}{10} \right\}.$$

Regardless of the material representation, we train a committee of 4 regressors. This approach provides robust predictions of various material properties and offers the additional benefit of measuring the average deviation among the 4 models' predictions. This average deviation is not a direct measure of the uncertainty: in the absence of other *a priori* knowledge about the properties, it serves as an indicator of the reliability of the predictions.

4.4. AI-experts training

We hypothesize that the specialized "energy" material set M^E is affected by biases due to the specific criteria used to select these materials, which in turn stem from the experts' knowledge. By leveraging these biases, we can train ML classifiers — referred to as AI-experts — to operate in the high-dimensional space of composition-based descriptors, effectively distinguishing materials belonging to M^E from a randomly constructed set $M^{NE} \subset M \setminus M^E$. To achieve this, we define two classes for our binary classification task:

- *Class 1*: Materials within the specialized set M^E.
- *Class 0*: A randomly selected set of materials M^{NE} from the Materials Project database that do not overlap with M^E (i.e., $M^{NE} \subset M \setminus M^E$), removing possible *polymorph* crystals, and ensuring that the two classes have the same cardinality $|M^E| = |M^{NE}|$ to maintain balance between the classes.

Our AI-experts consist of a committee of 10 binary classifiers. For this work, we found GBDTs [83] efficient and sufficiently accurate. Using the python ML library `scikit-learn` [84] pipeline construction, we pre-processed the data before feeding it to the GBDT (`GradientBoostingClassifier` object in `scikit-learn`). Indeed, before feeding the data into the GBDT classifiers, we standardized the input features to ensure zero average and unit variance for all features. Then, the data dimension is reduced by performing RFE [85], to constrain that the dimension of the composition-based descriptor is an order of magnitude lower than the cardinality of the specialized material set M^E. The key hyperparameters of the resulting ML models are the number of boosting stages to perform (N_b) and the number of selected features to use after the RFE step (N_f). For each AI-expert training, we performed a k-fold (with $k = 4$ number of split datasets) cross-validated grid search on the hyperparameter space:

$$(N_b, N_f) \in \{50, 100, 250, 500\} \times \left\{ \frac{|M^E|}{40}, \frac{|M^E|}{20}, \frac{|M^E|}{10} \right\}.$$

From the resulting skilled AI-experts, for each material we obtain a value between 0 and 1, which can be interpreted as a probability P_i that the input material x falls inside the region of the investigated energy material set M^E (class 1). Under our hypothesis, this also represents the probability of the input material x being inside the energy material region E. This can also be interpreted as the AI-experts approximating the boundary ∂E with a smooth transition, represented by the classifier's output.

4.5. Screening

The specialized AI-experts can then be used to screen the GNoME materials $y \in G$. After the composition featurization described earlier and classification processing, we can impose the acceptance criteria:

$$P(y \in M^E) = \frac{1}{N} \sum_{i=1}^{N} P_i(y \in M^E) \approx P(y \in E) > 0.5,$$

where $N = 10$ is the number of AI-experts, and P_i is the prediction of the single AI-expert interpreted as the probability of the GNoME material y belonging to "class 1", which by construction is equivalent to the probability $P_i(y \in M^E) \approx P_i(y \in E)$. This ensemble approach allows us to effectively utilize the biases present in M^E to identify potential candidate materials within G that share similar characteristics, which should attenuate the extrapolation problem of the regression models and the benefits of reducing the pool size of materials to investigate as potential new materials.

4.6. Case-specific pre- and post-processing

4.6.1. Thermoelectric materials

For the thermoelectric materials investigation, within the ESTM database, after normalizing the stoichiometry of the formulae, some of the materials appeared multiple times with the same T value, along with different measured zT values. In such cases we retained the average zT value only for those instances showing a RSD over such zT values less than 20%. This brought the number of materials from the original 870 to 869 ($Ag_{100}Bi_{63}Nb_7Sb_{30}Se_{200}$ is the only material reported with two highly different zT values for each of the temperatures it is listed with, namely 323 K, 423 K, 523 K, 623 K, 723 K, 823 K), along with a reduction of the T-based replicas from 5101 to 5061.

Also, the materials for "class 0" were randomly selected within MP, with the only constraint that these compositions were not already present in the ESTM database.

As already mentioned above, such compounds are featurized with 145 composition-based features encompassing stoichiometric attributes (based on the ratios of elements), elemental property statistics (including the mean, absolute deviation, minimum, and maximum of 22 atomic properties such as atomic number and atomic radii), electronic structure attributes (which represent the average fraction of electrons in the s, p, d, and f valence shells for all elements in the compound), and ionic compound attributes (indicating the possibility of forming an ionic compound, assuming all elements exist in a single oxidation state), adding the temperature T which plays as the 146th feature.

After applying the AI-based expert screening and using the regression models for predicting the figure of merit zT, we identified

a subset of materials (approximately 4.9% of the screened database) for which the regressor predicted negative zT values at lower temperatures, which is unphysical. We therefore excluded these materials from further analysis. This behavior occurs because our training dataset contains measurements only at temperatures at which the ESTM materials exhibit thermoelectric properties, while explicitly excluding data points for the same materials at temperatures where $zT = 0$, i.e., when they lack thermoelectric properties. Consequently, the regressor models were never exposed to such scenarios during training, leading, in some cases, to purely extrapolative predictions and unphysical (negative) zT values.

For plotting and ranking, we used the following function:

$$\begin{aligned} R^T(x) = {} & w_1 \cdot n\left(P(x)\right) \\ & + w_2 \cdot n\left(zT(x)\right) \\ & + w_3 \cdot n\left(\sigma\left(zT(x)\right)\right), \end{aligned} \tag{4}$$

where:

$n(x)$ is the min–max normalization function $n(x) = \frac{x - \min(x)}{\max(x) - \min(x)}$;
$P(x)$ is the average probability from the AI-experts;
$zT(x)$ is the "figure of merit" predicted by the regressors;
$\sigma\left(zT(x)\right)$ is the standard deviation of the predicted "figure of merit";
$\left[w_1, w_2, w_3\right]$ are arbitrarily chosen weights, respectively equal to [2, 3, −1].

Note that using negative weights, w_i, penalizes that specific candidate property — in this case study, the deviation of the regressor committee prediction, $\sigma\left(zT(x)\right)$.

4.6.2. Perovskite

For the perovskite materials investigation, the specialized materials in MP were selected with the constraint of being suitable for PV applications. First, the material must not be a metal, as the valence and conduction bands must not overlap. Additionally, the material must not possess magnetic properties: these would enhance the probability of self-trapping of charge carriers, thus resulting in reduced carrier mobility and increased recombination rates [86,87]. Finally, the material must display a band gap in the range $1 \times 10^{-3}\,\text{eV} < E_g \leq 2.5\,\text{eV}$ [55–57]. For the mixed models in regression, the non-perovskite materials were selected without applying a constraint on the magnetic properties or on the band gap range, except for ensuring that they were non-metallic. For the training of the AI-experts, the materials for "class 0" were randomly selected without any constraints. The adoption of the same constraints would have had a negative impact on the classification capabilities of the models, as having only non-metallic and non-magnetic materials would have resulted in a partial loss of the knowledge leveraged by the AI-experts. At the other end of the spectrum, including only metallic and magnetic materials in "class 0" would have introduced an undesired bias that materials not displaying these properties have higher probabilities of being considered possible candidates.

For plotting and ranking, we used the following function:

$$\begin{aligned} R^P(x) = {} & w_1 \cdot n\left(P(x)\right) \\ & + w_2 \cdot n\left(\sigma\left(P(x)\right)\right) \\ & + w_3 \cdot n\left(\sigma\left(E_g(x)\right)\right) \\ & + w_4 \cdot n\left(\left|E_g(x) - 1.34\right|\right), \end{aligned} \tag{5}$$

where:

$n(x)$ is the min–max normalization function $n(x) = \frac{x - \min(x)}{\max(x) - \min(x)}$;
$P(x)$ is the average probability from the AI-experts;
$\sigma\left(P(x)\right)$ is the standard deviation of the AI-experts prediction;
$\sigma\left(E_g(x)\right)$ is the standard deviation of the predicted band gap;
$\left|E_g(x) - 1.34\right|$ is the distance of the predicted band gap from the ideal 1.34 eV value [88,89];
$\left[w_1, w_2, w_3, w_4\right]$ are arbitrarily chosen weights, respectively equal to [4, −2, −2, −1].

Note that using negative weights, w_i, penalizes that specific candidate property — in this case study, the deviation of the regressor committee prediction, $\sigma\left(zT(x)\right)$, and AI-experts $\sigma\left(P(x)\right)$ and the distance for ideal 1.34 eV band gap.

4.6.3. Cathodes

For the cathode materials database, a crucial preprocessing step was outlier detection to eliminate possible data anomalies. Given the small training sets available for some working ions, noise produced by irregularities could strongly degrade ML model performance. We employed the Interquartile Range (IQR) method [90]. For each property of interest for our material class, we calculated the first quartile (Q_1) and third quartile (Q_3), and determined the IQR as $\text{IQR} = Q_3 - Q_1$. Thus the data points that fell outside the range $\left[Q_1 - k \cdot \text{IQR}, Q_3 + k \cdot \text{IQR}\right]$ were removed since they were considered a possible outlier. The scale value k for the IQR is usually set at 1.5. Since an actual high-performance material can be mistaken as an outlier, we set k conservatively to 3.0 to retain a broader range of data.

For the cathode materials investigation, the materials for "class 0" were randomly selected with the constraint that the working-ion (Li, Na, Mg, K, Ca, Cs, Al, Rb, Y) is present in the composition. This constraint was implemented to prevent the classifier from developing a bias that materials containing the specific working-ion are automatically considered as possible candidates, which could potentially overestimate their number. This condition was also applied after the AI-experts' classification process.

To determine the oxidation state for each working ion within the unit crystal cell of the candidate cathodes, we employ the Python library `pymatgen` [91], specifically its `oxi_state_guesses` function. The algorithm first computes the bond valence sum V_i of all symmetrically distinct ith sites using the method and parameters defined by O'Keefe and Brese [92] with the formula:

$$V_i = \sum_{j}^{j \in N} e^{\frac{R_{ij} - d_{ij}}{b}},$$

where b is a "universal" constant equal to 0.37 Å, d_{ij} is the bond length between the ith atom and the jth atom in the neighborhood set N, and R_{ij} is the bond valence parameter defined as:

$$R_{ij} = \left(r_i + r_j - \frac{r_i r_j \left(\sqrt{c_i} - \sqrt{c_j}\right)^2}{c_i r_i + c_j r_j} \right) \left(1 - \delta_{ij}\right).$$

Here, δ_{ij} is the Kronecker delta, r_i is the "size" parameter, and c_i is a second parameter related to electronegativity, both tabulated for each element. The posterior probabilities of all oxidation states X^{n+} of the atom element is computed using Bayesian inference:

$$P\left(X^{n+} | V_i\right) = P\left(V_i | X^{n+}\right) \cdot P\left(X^{n+}\right),$$

where the likelihood probability $P\left(V_i | X^{n+}\right)$ is modeled as a Gaussian distribution with mean and standard deviation determined from an analysis of the Inorganic Crystal Structure Database (ICSD), and the prior probability $P\left(X^{n+},\right)$ is tabulated and derived from a frequency analysis of the ICSD. Finally, the algorithm computes and selects the most probable oxidation state combination that results in a charge-neutral cell.

For plotting and ranking, we used the following function:

$$\begin{aligned} R^B(x) = & w_1 \cdot n\left(P(x)\right) \\ & + w_2 \cdot n\left(\sigma\left(P(x)\right)\right) \\ & + w_3 \cdot n\left(\sigma\left(\Delta V_c(x)\right)\right) \\ & + w_4 \cdot n\left(q(x)\right) \\ & + w_5 \cdot n\left(e(x)\right) \\ & + w_6 \cdot n\left(\max\left(\Delta \mathrm{Vol}(x)\right)\right) \\ & + w_7 \cdot n\left(\Delta E_{\mathrm{charge}}(x)\right) \\ & + w_8 \cdot n\left(\Delta E_{\mathrm{discharge}}(x)\right). \end{aligned} \tag{6}$$

where:

$n(x)$ is the min–max normalization function $n(x) = \frac{x - \min(x)}{\max(x) - \min(x)}$;
$P(x)$ is the average probability from the AI-experts;
$\sigma\left(P(x)\right)$ is the standard deviation of the AI-experts prediction;
$\sigma\left(\Delta V_c(x)\right)$ is the standard deviation of the "average voltage" of the regressor prediction;
$q(x)$ is the predicted "gravimetric capacity";
$e(x)$ is the predicted "gravimetric energy";
$\max\left(\Delta \mathrm{Vol}(x)\right)$ is the predicted maximum voltage change during the phase transition;
$\Delta E_{\mathrm{charge}}(x)$ is the predicted "energy above the hull" of the charged state (i.e. the stability of charge);
$\Delta E_{\mathrm{discharge}}(x)$ is the predicted "energy above the hull" of the discharged state (i.e. the stability of discharge);
$\left[w_1, \ldots, w_8\right]$ are arbitrarily chosen weights, respectively equal to [4, −2, −2, 2, 2, −0.5, −0.5, −0.5].

Note that using negative weights, w_i, penalizes that specific candidate property — in this case study, the deviation of the regressor committee prediction for the average voltage, $\sigma\left(V_c(x)\right)$, and AI-experts $\sigma\left(P(x)\right)$, the predicted expansion, $\max\left(\Delta \mathrm{Vol}(x)\right)$, and the instability in charged, $\Delta E_{\mathrm{charge}}(x)$, and discharged states, $\Delta E_{\mathrm{discharge}}(x)$.

CRediT authorship contribution statement

Paolo De Angelis: Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Giulio Barletta:** Writing – original draft, Software, Investigation, Formal analysis, Data curation. **Giovanni Trezza:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pietro Asinari:** Writing – review & editing, Supervision, Funding acquisition. **Eliodoro Chiavazzo:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Code availability

The entire workflow is managed through Jupyter notebooks and Python libraries with additional installation requirements and routines provided and maintained in the GitHub repository (https://github.com/paolodeangelis/Energy-GNoME/) and Zenodo repository (https://doi.org/10.5281/zenodo.14338533).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Eliodoro Chiavazzo reports financial support was provided by Ministero dell'Università e della Ricerca (MUR). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

G.T. and E.C. acknowledge funding under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3-Call for tender No. 1561 of 11.10.2022 of Ministero dell'Università e della Ricerca (MUR), funded by the European Union NextGenerationEU. We acknowledge ISCRA (IsB29_NEXT-LIB) for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.egyai.2025.100605.

Data availability

All GNoME database screening results (Energy-GNoME), including metadata and crystal structure files, are available in our GitHub (https://github.com/paolodeangelis/Energy-GNoME/) and Zenodo (https://doi.org/10.5281/zenodo.14338533) repositories. Furthermore, the Energy-GNoME database can be interactively explored using our web application at https://paolodeangelis.github.io/Energy-GNoME/.

References

[1] Sun H, Edziah BK, Sun C, Kporsu AK. Institutional quality, green innovation and energy efficiency. Energy Policy 2019;135:111002. http://dx.doi.org/10.1016/j.enpol.2019.111002.
[2] Kojima A, Teshima K, Shirai Y, Miyasaka T. Organometal Halide Perovskites as visible-light sensitizers for photovoltaic cells. J Am Chem Soc 2009;131:6050–1. http://dx.doi.org/10.1021/ja809598r.
[3] Green MA, Ho-Baillie A, Snaith HJ. The emergence of perovskite solar cells. Nat Photonics 2014;8:506–14. http://dx.doi.org/10.1038/nphoton.2014.134.
[4] Snyder GJ, Toberer ES. Complex thermoelectric materials. Nat Mater 2008;7:105–14. http://dx.doi.org/10.1038/nmat2090.
[5] Bell LE. Cooling, heating, generating power, and recovering waste heat with thermoelectric systems. Sci 2008;321:1457–61. http://dx.doi.org/10.1126/science.1158899.
[6] Whittingham MS. Electrical energy storage and intercalation chemistry. Sci 1976;192:1126–7. http://dx.doi.org/10.1126/science.192.4244.1126.
[7] Mizushima K, Jones PC, Wiseman PJ, Goodenough JB. LixCoO2 (0<x<-1): A new cathode material for batteries of high energy density. Mater Res Bull 1980;15:783–9. http://dx.doi.org/10.1016/0025-5408(80)90012-4.
[8] Lazzari M, Scrosati B. A cyclable lithium organic electrolyte cell based on two intercalation electrodes. J Electrochem Soc 1980;127:773–4. http://dx.doi.org/10.1149/1.2129753.
[9] Kittner N, Lill F, Kammen DM. Energy storage deployment and innovation for the clean energy transition. Nat Energy 2017;2:1–6. http://dx.doi.org/10.1038/nenergy.2017.125.
[10] Halkos GE, Gkampoura E-C. Reviewing usage, potentials, and limitations of renewable energy sources. Energies 2020;13:2906. http://dx.doi.org/10.3390/en13112906.
[11] Nadeem F, Hussain SMS, Tiwari PK, Goswami AK, Ustun TS. Comparative review of energy storage systems, their roles, and impacts on future power systems. IEEE Access 2019;7:4555–85. http://dx.doi.org/10.1109/ACCESS.2018.2888497.
[12] Hohenberg P, Kohn W. Inhomogeneous electron gas. Phys Rev 1964;136:B864–71. http://dx.doi.org/10.1103/PhysRev.136.B864.
[13] Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. Phys Rev 1965;140:A1133–8. http://dx.doi.org/10.1103/PhysRev.140.A1133.
[14] Nandy A, Terrones G, Arunachalam N, Duan C, Kastner DW, Kulik HJ. MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. Sci Data 2022;9:74. http://dx.doi.org/10.1038/s41597-022-01181-0.
[15] Back S, Aspuru-Guzik A, Ceriotti M, Gryn'ova G, Grzybowski B, Gu G Ho, et al. Accelerated chemical science with AI. Digit Discov 2024;3:23–33. http://dx.doi.org/10.1039/D3DD00213F.
[16] Schrier J, Norquist AJ, Buonassisi T, Brgoch J. In pursuit of the exceptional: Research directions for machine learning in chemical and materials science. J Am Chem Soc 2023;145:21699–716. http://dx.doi.org/10.1021/jacs.3c04783.
[17] Zdeborová L. New tool in the box. Nat Phys 2017;13:420–1. http://dx.doi.org/10.1038/nphys4053.

[18] Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: Status, challenges, and perspectives. Adv Sci 2019;6:1900808. http://dx.doi.org/10.1002/advs.201900808.

[19] Lee J, Bagheri B, Kao H-A. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. Manuf Lett 2015;3:18–23. http://dx.doi.org/10.1016/j.mfglet.2014.12.001.

[20] Casini M, De Angelis P, Porrati M, Vigo P, Fasano M, Chiavazzo E, et al. Machine Learning and image analysis towards improved energy management in Industry 4.0: a practical case study on quality control. Energy Effic 2024;17:48. http://dx.doi.org/10.1007/s12053-024-10228-7.

[21] Li J, Herdem MS, Nathwani J, Wen JZ. Methods and applications for Artificial Intelligence, big data, internet of things, and blockchain in smart energy management. Energy AI 2023;11:100208. http://dx.doi.org/10.1016/j.egyai.2022.100208.

[22] Casini M, De Angelis P, Chiavazzo E, Bergamasco L. Current trends on the use of deep learning methods for image analysis in energy applications. Energy AI 2024;15:100330. http://dx.doi.org/10.1016/j.egyai.2023.100330.

[23] Koroteev D, Tekic Z. Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. Energy AI 2021;3:100041. http://dx.doi.org/10.1016/j.egyai.2020.100041.

[24] Liu Y, Esan OC, Pan Z, An L. Machine learning for advanced energy materials. Energy AI 2021;3:100049. http://dx.doi.org/10.1016/j.egyai.2021.100049.

[25] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Mater 2013;1:011002. http://dx.doi.org/10.1063/1.4812323.

[26] Hellenbrandt M. The inorganic crystal structure database (ICSD)—Present and future. Crystallogr Rev 2004;10:17–22. http://dx.doi.org/10.1080/08893110410001664882.

[27] Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). JOM 2013;65:1501–9. http://dx.doi.org/10.1007/s11837-013-0755-4.

[28] Draxl C, Scheffler M. The NOMAD laboratory: from data sharing to artificial intelligence. J Phys Mater 2019;2:036001. http://dx.doi.org/10.1088/2515-7639/ab13bb.

[29] Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor RH, et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. Comput Mater Sci 2012;58:227–35. http://dx.doi.org/10.1016/j.commatsci.2012.02.002.

[30] Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nat 2018;559:547–55. http://dx.doi.org/10.1038/s41586-018-0337-2.

[31] Tabor DP, Roch LM, Saikin SK, Kreisbeck C, Sheberla D, Montoya JH, et al. Accelerating the discovery of materials for clean energy in the era of smart automation. Nat Rev Mater 2018;3:5–20. http://dx.doi.org/10.1038/s41578-018-0005-z.

[32] Fanourgakis GS, Gkagkas K, Tylianakis E, Froudakis GE. A universal machine learning algorithm for large-scale screening of materials. J Am Chem Soc 2020;142:3814–22. http://dx.doi.org/10.1021/jacs.9b11084.

[33] Trezza G, Bergamasco L, Fasano M, Chiavazzo E. Minimal crystallographic descriptors of sorption properties in hypothetical MOFs and role in sequential learning optimization. Npj Comput Mater 2022;8:1–14. http://dx.doi.org/10.1038/s41524-022-00806-7.

[34] Cerqueira TFT, Sanna A, Marques MAL. Sampling the materials space for conventional superconducting compounds. Adv Mater 2024;36:2307085. http://dx.doi.org/10.1002/adma.202307085.

[35] Moses IA, Joshi RP, Ozdemir B, Kumar N, Eickholt J, Barone V. Machine learning screening of metal-ion battery electrode materials. ACS Appl Mater Interfaces 2021;13:53355–62. http://dx.doi.org/10.1021/acsami.1c04627.

[36] Rutt A, Shen J-X, Horton M, Kim J, Lin J, Persson KA. Expanding the material search space for multivalent cathodes. ACS Appl Mater Interfaces 2022;14:44367–76. http://dx.doi.org/10.1021/acsami.2c11733.

[37] Rong Z, Kitchaev D, Canepa P, Huang W, Ceder G. An efficient algorithm for finding the minimum energy path for cation migration in ionic materials. J Chem Phys 2016;145:074112. http://dx.doi.org/10.1063/1.4960790.

[38] Wang P, Kateris N, Li B, Zhang Y, Luo J, Wang C, et al. High-performance lithium–sulfur batteries via molecular complexation. J Am Chem Soc 2023;145:18865–76. http://dx.doi.org/10.1021/jacs.3c05209.

[39] Kim K, Ward L, He J, Krishna A, Agrawal A, Wolverton C. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds. Phys Rev Mater 2018;2:123801. http://dx.doi.org/10.1103/PhysRevMaterials.2.123801.

[40] Kang P, Liu Z, Abou-Rachid H, Guo H. Machine-learning assisted screening of energetic materials. J Phys Chem 2020;124:5341–51. http://dx.doi.org/10.1021/acs.jpca.0c02647.

[41] Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 2009;37:W623–33. http://dx.doi.org/10.1093/nar/gkp456.

[42] Rao Z, Tung P-Y, Xie R, Wei Y, Zhang H, Ferrari A, et al. Machine learning–enabled high-entropy alloy discovery. Sci 2022;378:78–85. http://dx.doi.org/10.1126/science.abo4940.

[43] Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nat 2023;624:80–5. http://dx.doi.org/10.1038/s41586-023-06735-9.

[44] Trezza G, Chiavazzo E. Classification-based detection and quantification of cross-domain data bias in materials discovery. J Chem Inf Model 2025;65:1747–61. http://dx.doi.org/10.1021/acs.jcim.4c01766.

[45] Na GS, Chang H. A public database of thermoelectric materials and system-identified material representation for data-driven discovery. Npj Comput Mater 2022;8:1–11. http://dx.doi.org/10.1038/s41524-022-00897-2.

[46] Gorai P, Toberer ES, Stevanović V. Thermoelectricity in transition metal compounds: the role of spin disorder. Phys Chem Chem Phys 2016;18:31777–86. http://dx.doi.org/10.1039/C6CP06943F.

[47] Corey RB, Pauling L. Molecular models of amino acids, peptides, and proteins. Rev Sci Instrum 1953;24:621–7. http://dx.doi.org/10.1063/1.1770803.

[48] Jmol: an open-source Java viewer for chemical structures in 3D. 2024, URL https://jmol.sourceforge.net/, (Accessed 09 August 2024).

[49] Sootsman J, Chung D, Kanatzidis M. New and old concepts in thermoelectric materials. Angew Chem Int Ed 2009;48:8616–39. http://dx.doi.org/10.1002/anie.200900598.

[50] Ward L, Dunn A, Faghaninia A, Zimmermann NER, Bajaj S, Wang Q, et al. Matminer: An open source toolkit for materials data mining. Comput Mater Sci 2018;152:60–9. http://dx.doi.org/10.1016/j.commatsci.2018.05.018.

[51] Ward L, Agrawal A, Choudhary A, Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials. Npj Comput Mater 2016;2:1–7. http://dx.doi.org/10.1038/npjcompumats.2016.28.

[52] Roy P, Kumar Sinha N, Tiwari S, Khare A. A review on perovskite solar cells: Evolution of architecture, fabrication techniques, commercialization issues and status. Sol Energy 2020;198:665–88. http://dx.doi.org/10.1016/j.solener.2020.01.080.

[53] Nair S, Patel SB, Gohel JV. Recent trends in efficiency-stability improvement in perovskite solar cells. Mater Today Energy 2020;17:100449. http://dx.doi.org/10.1016/j.mtener.2020.100449.

[54] Osterrieder T, Schmitt F, Lüer L, Wagner J, Heumüller T, Hauch J, et al. Autonomous optimization of an organic solar cell in a 4-dimensional parameter space. Energy Env Sci 2023;16:3984–93. http://dx.doi.org/10.1039/D3EE02027D.

[55] Hu Z, Lin Z, Su J, Zhang J, Chang J, Hao Y. A review on energy band-gap engineering for perovskite photovoltaics. Sol RRL 2019;3:1900304. http://dx.doi.org/10.1002/solr.201900304.

[56] Hörantner MT, Leijtens T, Ziffer ME, Eperon GE, Christoforo MG, McGehee MD, et al. The potential of multijunction perovskite solar cells. ACS Energy Lett 2017;2:2506–13. http://dx.doi.org/10.1021/acsenergylett.7b00647.

[57] Liu X, Wu Z, Fu X, Tang L, Li J, Gong J, et al. Highly efficient wide-band-gap perovskite solar cells fabricated by sequential deposition method. Nano Energy 2021;86:106114. http://dx.doi.org/10.1016/j.nanoen.2021.106114.

[58] Larcher D, Tarascon J-M. Towards greener and more sustainable batteries for electrical energy storage. Nat Chem 2015;7:19–29. http://dx.doi.org/10.1038/nchem.2085.

[59] Canepa P, Sai Gautam G, Hannah DC, Malik R, Liu M, Gallagher KG, et al. Odyssey of multivalent cathode materials: Open questions and future challenges. Chem Rev 2017;117:4287–341. http://dx.doi.org/10.1021/acs.chemrev.6b00614.

[60] Manthiram A. A reflection on lithium-ion battery cathode chemistry. Nat Commun 2020;11:1550. http://dx.doi.org/10.1038/s41467-020-15355-0.

[61] Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nat Commun 2022;13:2453. http://dx.doi.org/10.1038/s41467-022-29939-5.

[62] Chen Z, Andrejevic N, Smidt T, Ding Z, Xu Q, Chi Y, et al. Direct prediction of phonon density of states with euclidean neural networks. Adv Sci 2021;8:2004214. http://dx.doi.org/10.1002/advs.202004214.

[63] Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation. 2001, http://dx.doi.org/10.3115/1073012.1073017.

[64] Teufl T, Pritzl DJ, Hartmann L, Solchenbach S, Mendez M, Gasteiger H. Implications of the thermal stability of fec-based electrolytes for Li-Ion batteries. J Electrochem Soc 2023. http://dx.doi.org/10.1149/1945-7111/acbc52.

[65] Subaşı Y, Altenschmidt L, Lindgren F, Ericsson T, Häggström L, Tai C-W, et al. Synthesis and characterization of a crystalline Na4Fe3(PO4)2(P2O7) cathode material for sodium-ion batteries. J Mater Chem 2024;12:23506–17. http://dx.doi.org/10.1039/D4TA03554B.

[66] He F, Kang J, Liu T, Deng H, Zhong B, Sun Y, et al. Research progress on electrochemical properties of $Na_3V_2(PO_4)_3$ as cathode material for Sodium-Ion batteries. Ind Eng Chem Res 2023;62:3444–64. http://dx.doi.org/10.1021/acs.iecr.2c04054.

[67] Jia X, Lynch A, Huang Y, Danielson M, Lang'at I, Milder A, et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. Nat 2019;573:251–5. http://dx.doi.org/10.1038/s41586-019-1540-5.

[68] Mancardi G, Mikolajczyk A, Annapoorani VK, Bahl A, Blekos K, Burk J, et al. A computational view on nanomaterial intrinsic and extrinsic features for nanosafety and sustainability. Mater Today 2023;67:344–70. http://dx.doi.org/10.1016/j.mattod.2023.05.029.

[69] Talirz L, Kumbhar S, Passaro E, Yakutovich AV, Granata V, Gargiulo F, et al. Materials cloud, a platform for open computational science. Sci Data 2020;7:299. http://dx.doi.org/10.1038/s41597-020-00637-5.

[70] Winther KT, Hoffmann MJ, Boes JR, Mamun O, Bajdich M, Bligaard T. Catalysis-Hub.org, an open electronic structure database for surface reactions. Sci Data 2019;6:75. http://dx.doi.org/10.1038/s41597-019-0081-y.

[71] Zhou C, Wu S. Medium- and high-temperature latent heat thermal energy storage: Material database, system review, and corrosivity assessment. Int J Energy Res 2019;43:621–61. http://dx.doi.org/10.1002/er.4216.

[72] Shimakawa H, Kumada A, Sato M. Extrapolative prediction of small-data molecular property using quantum mechanics-assisted machine learning. Npj Comput Mater 2024;10:1–14. http://dx.doi.org/10.1038/s41524-023-01194-2.

[73] Jain A, Hautier G, Ong SP, Persson K. New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships. J Mater Res 2016;31:977–94. http://dx.doi.org/10.1557/jmr.2016.80.

[74] Manzhos S, Tsuda S, Ihara M. Machine learning in computational chemistry: interplay between (non)linearity, basis sets, and dimensionality. Phys Chem Chem Phys 2023;25:1546–55. http://dx.doi.org/10.1039/D2CP04155C.

[75] Ong SP, Cholia S, Jain A, Brafman M, Gunter D, Ceder G, et al. The materials application programming interface (API): A simple, flexible and efficient API for materials data based on REpresentational state transfer (REST) principles. Comput Mater Sci 2015;97:209–15. http://dx.doi.org/10.1016/j.commatsci.2014.10.037.

[76] Choudhary K, DeCost B, Tavazza F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. Phys Rev Mater 2018;2:083801. http://dx.doi.org/10.1103/PhysRevMaterials.2.083801.

[77] Geiger M, Smidt T. E3nn: euclidean neural networks,. 2022, http://dx.doi.org/10.48550/arXiv.2207.09453, arXiv:2207.09453.

[78] Geiger M, Smidt T, M A, Miller BK, Boomsma W, Dice B, et al. E3nn. Zenodo data. 2022, http://dx.doi.org/10.5281/zenodo.6459381, URL https://zenodo.org/records/6459381.

[79] Thomas N, Smidt T, Kearnes S, Yang L, Li L, Kohlhoff K, et al. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. 2018, http://dx.doi.org/10.48550/arXiv.1802.08219, arXiv:1802.08219.

[80] Weiler M, Geiger M, Welling M, Boomsma W, Cohen T. 3D steerable CNNs: Learning rotationally equivariant features in volumetric data. 2018, http://dx.doi.org/10.48550/arXiv.1807.02547, arXiv:1807.02547.

[81] Kondor R, Lin Z, Trivedi S. Clebsch-Gordan nets: a fully Fourier space spherical convolutional neural network. 2018, http://dx.doi.org/10.48550/arXiv.1806.09231, arXiv:1806.09231.

[82] Loshchilov I, Hutter F. Decoupled weight decay regularization. 2019, http://dx.doi.org/10.48550/arXiv.1711.05101, arXiv:1711.05101.

[83] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001;29. http://dx.doi.org/10.1214/aos/1013203451.

[84] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[85] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422. http://dx.doi.org/10.1023/A:1012487302797.

[86] Haas C. Magnetic semiconductors. CRC Crit Rev Solid State Sci 1970;1:47–98. http://dx.doi.org/10.1080/10408437008243418.

[87] Ren L, Wang Y, Wang M, Wang S, Zhao Y, Cazorla C, et al. Tuning magnetism and photocurrent in Mn-doped organic–inorganic Perovskites. J Phys Chem Lett 2020;11:2577–84. http://dx.doi.org/10.1021/acs.jpclett.0c00034.

[88] Konstantakou M, Stergiopoulos T. A critical review on tin halide perovskite solar cells. J Mater Chem 2017;5:11518–49. http://dx.doi.org/10.1039/C7TA00929A.

[89] Zhou X, Zhang L, Wang X, Liu C, Chen S, Zhang M, et al. Highly efficient and stable GABr-modified ideal-bandgap (1.35 eV) Sn/Pb Perovskite solar cells achieve 20.63% efficiency with a record small V_{oc} deficit of 0.33 V. Adv Mater 2020;32:1908107. http://dx.doi.org/10.1002/adma.201908107.

[90] Upton GJG, Cook I. Understanding statistics. 1. publ ed.. Oxford: Oxford University Press; 1996.

[91] Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, et al. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. Comput Mater Sci 2013;68:314–9. http://dx.doi.org/10.1016/j.commatsci.2012.10.028.

[92] O'Keefe M, Brese NE. Atom sizes and bond lengths in molecules and crystals. J Am Chem Soc 1991;113:3226–9. http://dx.doi.org/10.1021/ja00009a002.