nature portfolio

Peer Review File

Predicting Thermodynamic Stability of Inorganic Compounds Using Ensemble Machine Learning Based on Electron Configuration

Corresponding Author: Professor Jianxin Wang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In materials science, predicting the thermodynamic stability of inorganic compounds is crucial. Traditionally, this stability has been assessed through experimental methods or computational approaches like density functional theory (DFT), which are often time-consuming and resource-intensive. Recent advancements in machine learning (ML) present promising solutions to these challenges, as discussed in previous works cited by the authors. This study is motivated by the need for more efficient and accurate methods to predict the thermodynamic stability of inorganic compounds. However, based on the key findings and model performances, reviewer did not find the results to be particularly ground-breaking or exciting. Therefore, I do not suggest this paper for publication in Nature Communications.
The primary objective of this research is to develop a machine learning framework that accurately predicts the thermodynamic stability of inorganic compounds using electron configuration information. However, the authors need to provide more detailed explanations of the Electron Configuration Convolutional Neural Network (ECCNN), including diagrams to aid understanding, and enhance the clarity of figures illustrating the model architectures. While the authors conducted several case studies using this model, they do not effectively address the research challenges in these materials, and the reviewer cannot discern the necessity of this model for solving the stated problems. Additionally, upon checking the code, the reviewer found that it still requires debugging to run the demo. Therefore, the authors need to address these issues.

(Remarks on code availability)

The code can be run after debugging. However, the results of the paper cannot be replicated as the database was not provided in the GitHub repository.

Reviewer #2

(Remarks to the Author)
Overall, the discussion is well-structured and clear. The authors have done a good job explaining the motivation behind their work, the challenges in the field, and how their proposed method addresses these challenges. However, it could benefit from more technical details and context to support the claims made. Moreover, my feeling is that a graphical representation of the used metrics would help the general audience to better understand their relevance.

Minor editorial suggestions
Fig 1.: Nodes should be in the layer above the connections.
Fig 7: the 2 and 3 of Al2O3 collide with the surrounding box.
Consider refraining from using the term “impressive” in the assessment of the models all over the manuscript.

Scientific technical questions and remarks:
The authors mainly use AUC to assess the models` performance, however, there is no addressing of how they avoided the known problem of poor classification performance, where AUC incorporates irrelevant areas. Furthermore, relying solely on

the outcome of the AUC would result in a lack of precision and negative predictive value information due to the fact that AUC focuses on sensitivity and specificity but does not provide information regarding precision or negative predictive value. For the ablation study section it would be more representative to present the outcome of the other tested metrics mentioned within the manuscript, similar to the presentation of performance in the prediction in unknown space section. Could you please elaborate on that in the manuscript?

Related to the point that ECSG outperforms other models in terms of classification metrics and sample efficiency. However, it would be beneficial to include more specific details about these base-level models for comparison. For example, further details about how the nature and hypothesis used in construction of the tested models is influencing the classification metrics and sample efficiency. Could you please comment on this?

The authors explain that their method expands the parameter space and reduces the error between predictions and ground truth. However, the explanation could benefit from more technical details. How exactly does the method combine multiple models? How will a different method of combining the different models affect the propagation of errors? Moreover, how does it ensure the complementarity of different base-level models? Could you please add a paragraph to make this clearer?

Starting from line 640 “Our method also offers advantages in integrating heterogeneous data in materials science and engineering. Data in these fields often exhibit heterogeneity, encompassing numerical table data, spectra data, and image data.” This is a significant contribution. However, there is no description or details of how the proposed method integrated the heterogeneous data within any of the tested cases or models. How does the model handle different types of data? How does it preserve data integrity?

For future work, the authors mention plans to apply ECSG to other material properties. It would be an interesting element in the final paragraph of the manuscript to address the following question: What specific properties will you investigate? What challenges do you anticipate?

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
The authors present a new machine learning framework, called ECSG (Electron Configuration models with Stacked Generalization), for the prediction of materials stability, in which three composition-based sub-models Magpie, Roost, and a new architecture ECCNN (Electron Configuration Convolutional Neural Network) are combined by stacked generalization into a super-learner. The idea behind the ECSG framework is to limit the bias inherent in sub-models’ constructions of the composition-property relationship by unifying the models into a single framework. The authors demonstrate the ability of ECSG to outperform each of the three sub-models as well as several other composition-based machine learning models in predicting the stability of inorganic compounds from three large DFT databases (MP, OQMD, and JARVIS). Additionally, they demonstrate the ability of ECSG to predict new stable compounds in the largely unexplored family of double perovskite oxides. The ECSG framework is clearly impressive and useful, although it appears to be only marginally better than some models, and is missing consideration of crystal structure. The manuscript is well written, and I believe it is suitable for publication. I have some questions and comments about the work, which are below.

1) The authors use only composition-based models rather than structure-based ones. Their justification is that composition-based models are advantageous over structure-based models in that it is easier to acquire compositional information compared to detailed structure data. However, DFT databases (MP, OQMD, and JARVIS) contain crystal structures for every entry. Why not include this structure information into the ECSG model? This would enable us to compare ECSG with existing structure-based models such as CGCNN (Xie and Grossman, Phys. Rev. Lett. 120, 145301, 2018). Also, if the model is composition-based, would ECSG give us the same stability for two different crystal structures having the same chemical composition? If so, how would we know that Na2WNiO6, one of the stable perovskite oxides recommended by ECSG, does not have a polymorph that is lower in energy than the perovskite structure?

2) The “TRUE” perovskite oxides in Supplementary Table 6 are predicted to lie on the convex hull of the Materials Project database, correct? I wonder how many of these compounds would also lie on the convex hull of the OQMD and JARVIS databases, as these databases have different compounds in them, although many compounds are identical between the databases.

3) Personally, I would find it helpful if I had some idea of how the AUC and other scores translate to the actual numbers of compounds predicted to be stable or not. Without this I do not have a sense of which scores are good or bad. For example, in Table 1, the AUC score of ECSG is 0.887 whereas for RF it is 0.862, so it appears to me that ECSG is only marginally better than RF.

4) This may be nitpicky, but how exactly is ΔHd defined? If it is the "energy above the hull", then it can never be negative, but the "TRUE" DFT values in Supplementary Table 6 are negative. I think in this case ΔHd is the energy above the convex hull of compounds in the Materials Project at that moment in time.

5) Does the "stability probability" outputted by ECSG translate to the real probability that the compound is stable? For example, should I expect 20% of compounds with probability of 0.2 to be stable? If not, then I do not see the value of the distribution in Figure 5a, except that the compounds with "probability" greater than 0.9 represent a tiny fraction of all candidates. It is impressive that 25 of the 35 predicted-stable perovskite oxides were confirmed to be stable by DFT, but I have to wonder whether other composition-based models would perform nearly as well as ECSG in this case.

(Remarks on code availability)
I had to install a different set of Torch libraries that were compatible with my GPU machine, but otherwise, I found the instructions in the README file to be straightforward. I did not try to reproduce the numbers in the manuscript, but the code appears to produce stability predictions for hypothetical compositions.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have made great efforts in thoroughly addressing the comments and concerns raised by the reviewers. It is evident that they have thoughtfully incorporated the feedback, enhancing both the clarity and depth of the manuscript. I am pleased to note that the majority of the questions have been comprehensively answered, significantly improving the quality of the work.

Overall, the paper now demonstrates a high standard of scholarship , making it well-suited for publication. I look forward to seeing it contribute to the field.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have responsed comprehensively to my feedback and questions. Thanks!

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
The authors have responded comprehensively to our feedback and questions. Thanks!

Reviewer #4

(Remarks to the Author)
I have reviewed the authors' responses to reviewers and changes to the manuscript, and I find them to all be satisfactory. I appreciate the detailed answers to questions and feedback, as well as the additional clarifications and analyses that have been added to the manuscript. For these reasons, I strongly recommended publication of this work to Nature Communications.

Thank you for the attempt to include CGCNN, a structure-based neural network, as a base model into ECSG (becoming ECSG+C). While this addition fortunately improves predictive performance compared to the ECSG, it is unfortunate that ECSG+C does a poor job at distinguishing stable from unstable polymorphs (as does CGCNN alone). Indeed, more research is needed to improve the ability of machine learning to distinguish the energetics of polymorphs. I wonder if the poor performance is due to the low energy difference between polymorphs of the same composition, which is lower than the resolution of the ECSG+C model.

I also appreciate the inclusion of stability assessments (Supplementary Tables 9 and 10) using not just the MP convex hull but also OQMD and JARVIS. The fact that there is some disagreement on stability assessments between the DFT databases raises an important issue with DFT-based stability assessments: that DFT data on competing phases is often incomplete. This is important to keep in mind when interpreting machine learning predictions of stability.

(Remarks on code availability)
The README file now contains more information to aid users in installing the required Torch libraries, which can vary by GPU machine.

Summary

The authors would like to thank the reviewers for their valuable comments and suggestions for improving this manuscript. Below we provide point-by-point responses to the comments along with corresponding amendments made in the manuscript. The significant changes in the revised manuscript were highlighted in red color.

Answers to Reviewer #1

Reviewer #1 (Remarks to the Author):

Comment 1. *In materials science, predicting the thermodynamic stability of inorganic compounds is crucial. Traditionally, this stability has been assessed through experimental methods or computational approaches like density functional theory (DFT), which are often time-consuming and resource-intensive. Recent advancements in machine learning (ML) present promising solutions to these challenges, as discussed in previous works cited by the authors. This study is motivated by the need for more efficient and accurate methods to predict the thermodynamic stability of inorganic compounds. However, based on the key findings and model performances, reviewer did not find the results to be particularly ground-breaking or exciting. Therefore, I do not suggest this paper for publication in Nature Communications.*

Authors' Response: Thank you for your thorough review. We acknowledge your concern regarding the impact of our results. We would like to highlight the following contributions that position our model as a step forward in the field of computational materials science:

(i) **Incorporating Electronic-Level Features for Deeper Insights**: Previous machine learning models typically focus on features derived from elemental properties, such as atomic radii or ionization potentials. However, these models rarely explore electron-level features, limiting their ability to capture key electron-related attributes. Our method leverages electron configuration (EC) to encode a detailed view of electron distribution, providing valuable insights into essential properties such as ionization energy, electronegativity, and bond valence. These properties play a fundamental role in quantitatively assessing chemical behavior and material stability [1]. The Electron Configuration Convolutional Neural Network (ECCNN) introduced in this study efficiently encodes electronic structures and extracts electronic-level features, offering an innovative way to predict material properties at a finer granularity.

(ii) **Mitigating Inductive Bias through Multidomain Integration**: Many models suffer from inductive bias by relying on domain-specific knowledge. To address this, we integrate insights from three complementary domains: interatomic interactions, atomic properties, and electronic configuration. This multidomain approach reduces inductive bias and enhances the generalization capabilities of our ECSG model, especially in data-scarce scenarios.

(iii) **Exploring New Chemical Spaces and Discovering Novel Materials**: ECSG’s ability to predict stability in unexplored chemical spaces was validated through case studies. We excluded halide perovskites, lithium-containing compounds, and transition metal oxides from the training set to treat them as unseen data. In screening 35 double perovskite oxides, our model confirmed 25 stable candidates using DFT calculations, which demonstrates ECSG’s superior capability to explore new materials.

(iv) **Enhanced Sample Efficiency**: ECSG exhibits excellent sample efficiency, achieving an AUC of 0.800 on The Materials Project (MP) database using only 10% of the training data. In contrast, comparison models such as Roost [2] and CrabNet [3] required 70% of the training data to reach the same performance level. This efficiency makes ECSG particularly effective in domains where data availability is limited.

In addition, after carefully considering all reviewers' feedback, we made several significant revisions to strengthen the manuscript:

(i) To further validate the reliability of the proposed ECSG across multiple application scenarios, we conducted a new case study on 2D materials, focusing on identifying wide bandgap semiconductor candidates. The results show that ECSG can achieve good performance in predicting the stability of wide bandgap 2D materials, which can promote the development of such materials.

(ii) In the case study on double perovskite oxides, we added DFT validation for 35 materials selected by the comparative model [4] to further illustrate the effectiveness of our approach. The results show that our method significantly outperforms the comparison models.

(iii) We have provided additional technical details on our methods, including the base models, combination steps, key terminology, and clearer figures, to help readers better understand our approach.

(iv) To further enhance prediction accuracy, we introduced structure-based models into ECSG. The results showed that incorporating structural information can improve the model's performance. As a result, we offer two input options: in most cases, where the structure is unknown, predictions can still be made using only compositional data. Relevant updates are available on our GitHub repository (https://github.com/Haozou-csu/ECSG).

References

[1] Tofanelli, M. A., & Ackerson, C. J. (2012). Superatom electron configuration predicts thermal stability of Au25 (SR) 18 nanoclusters. Journal of the American Chemical Society, 134(41), 16937-16940.

[2] Goodall, R. E. A. & Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nat. Commun. 11, 6280 (2020).

[3] Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. npj Comput. Mater. 7, 77 (2021).

[4] Talapatra A, Uberuaga BP, Stanek CR, Pilania G. A Machine Learning Approach for the Prediction of Formability and Thermodynamic Stability of Single and Double Perovskite Oxides. Chem. Mater. 33, 845-858 (2021).

Comment 2. *The primary objective of this research is to develop a machine learning framework that accurately predicts the thermodynamic stability of inorganic compounds using electron configuration information. However, the authors need to provide more detailed explanations of the Electron Configuration Convolutional Neural Network (ECCNN), including diagrams to aid understanding, and enhance the clarity of figures illustrating the model architectures.*

Authors' Response: We appreciate your constructive feedback regarding the description of the Electron Configuration Convolutional Neural Network (ECCNN) model architecture. In response, we have expanded the section detailing the ECCNN architecture within the 'Methods' subsection.

Here are the key clarifications and enhancements made:

(1) **Model Architecture Overview**: As shown in Fig. R1 (Fig. 1(b) in the revised manuscript), the ECCNN model consists of two convolutional layers specifically designed to extract features from input electron configuration (EC) matrices. These matrices are represented as a 3D tensor with a shape of 118×168×8, where: 168 represents the length of the EC vector, reflecting the electron configurations of various elements. 118 indicates the number of distinct element types considered in our study. 8 channels correspond to the number of atoms of each element, following a binary conversion process. For example, if elements A and B in a chemical formula consist of 2 and 10 atoms, respectively, their values in the 8 channels would be represented as: Element A: 01000000; Element B: 01010000. In this representation, a value of 1 in a channel indicates that the EC vector of the corresponding element is copied into that channel.

The input then undergoes two convolutional operations, each with 64 filters of size 5×5. The second convolution is followed by a batch normalization (BN) operation and 2×2 max pooling. The extracted features are flattened into a one-dimensional vector, which is then fed into fully connected layers for prediction.

(2) **Enhanced Model Diagram**: We have also revised the model architecture diagram to enhance clarity and readability. The updated diagram (Fig. R1) now includes comprehensive labels for each layer—such as the convolutional layers, pooling layers, and fully connected layers—and annotations to elucidate the key operations performed at each stage of the model. This enhancement aims to provide a better understanding of data flow within the network and the functional role of each component in ECCNN architecture.

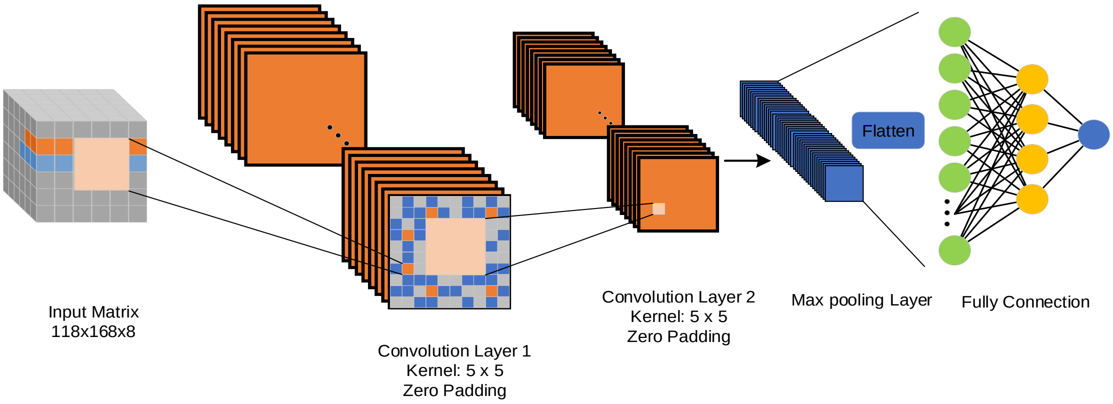

Fig. R1. The architecture of ECCNN. A typical ECCNN model contains an input layer, two convolution layers, a max pooling layer, and several fully connected layers.

Comment 3. *While the authors conducted several case studies using this model, they do not effectively address the research challenges in these materials, and the reviewer cannot discern the necessity of this model for solving the stated problems.*

Authors' Response: Thanks for your comments. In the revised manuscript, we provided a more detailed discussion of how ECSG addresses the research challenges in the presented materials and added a new case study on two-dimensional (2D) materials in the subsection 'Case Studies' under the section 'Results'.

In the subsection 'Prediction in unknown space', we aimed to test the ECSG model's ability to predict thermodynamic stability in unexplored chemical spaces. Predicting the thermodynamic stability of materials in unknown space is crucial because it is very similar to the material discovery in the real world. However, the challenge lies in how to make accurate predictions in the absence of information about the same type of materials. To this end, we excluded certain specific types of materials (such as lithium-containing oxides and transition metal oxides) from the training set and tested these materials. Because the ECSG model combines theoretical knowledge from multiple fields and has strong generalization capabilities, its prediction results are superior to other models even in completely unknown spaces, showing significant advantages in exploring new materials.

In the subsection ‘Case studies’, the case study on double perovskite oxides was designed to simulate the real material discovery process. The composition space of materials is often huge, and it is a challenge to screen out candidate materials that meet the conditions in such a huge composition space, as stable materials only account for a tiny part. We demonstrated ECSG's ability to efficiently explore vast composition spaces by incorporating electronic-level information alongside atomic-based features. By using ECSG, 35 candidate materials were selected from a composition space of more than 4 million. After DFT verification, 25 of them are in line with expectations, proving the reliability of our method. However, of the 35 perovskites screened by the comparison model Tala, only two are confirmed to be stable by DFT.

To further validate the reliability of the proposed ECSG across multiple application scenarios, we conducted a new case study on 2D materials, focusing on identifying wide bandgap semiconductor candidates. Since the discovery of graphene in 2004, 2D materials have gained prominence in materials science due to their unique properties and potential applications in electronics and optoelectronics. Despite graphene’s excellent performance, its small bandgap limits its use in semiconductors. As a result, researchers are focusing on 2D semiconductors with wide bandgaps (>2.0 eV), which hold promise for use under blue and ultraviolet light, crucial for new optoelectronic devices [1, 2]. However, designing materials that meet performance requirements but lack thermodynamic stability would hinder their practical use, leading to wasted resources and time. Thus, it is imperative to ensure both performance and thermodynamic stability for practical applications.

Table R1 The performance of ECSG and comparison models testing in the 2DMatpeida.

No.	Model	ACC	Precision	Recall	F1	NPV	AUC	AUPR
1	ECSG	**0.737**	**0.763**	0.775	**0.769**	0.701	**0.790**	0.786
2	Roost	0.687	0.730	0.714	0.718	0.644	0.752	0.769
3	CrabNet	0.691	0.671	**0.892**	0.766	**0.755**	0.752	0.768
4	RF	0.711	0.760	0.714	0.737	0.655	0.778	**0.807**
5	Adaboost	0.703	0.730	0.755	0.742	0.666	0.759	0.760
6	Magpie	0.706	0.750	0.721	0.735	0.654	0.777	0.790
7	Meredig	0.705	0.735	0.748	0.741	0.665	0.763	0.770
8	ElemNet	0.649	0.716	0.639	0.664	0.604	0.719	0.718
9	ATCNN	0.664	0.724	0.655	0.686	0.604	0.733	0.744
10	ECCNN	0.606	0.678	0.578	0.623	0.540	0.649	0.696

We used the ECSG, which incorporates information on electron configuration, to find potential 2D semiconductor candidate materials with wide bandgap to meet the needs of future technological development. To initiate this investigation, we extracted the composition and stability information of the materials from the C2DB database and trained an ECSG model to predict the thermodynamic stability of 2D materials [3]. Subsequently, we combined the large language model (LLM) DARWIN-7B [4], which has been proven to perform well in predicting experimental bandgaps, to screen 2D materials with bandgaps greater than 2.0 eV. During the screening process, we tested materials from the 2Dmatpedia database, which contains 4,743 materials [2]. As shown in Table R1 (Supplementary Table 8), ECSG obtains the best or the second-best results in all metrics.

Next, we conducted a statistical analysis of the samples predicted by ECSG to be positive and the samples predicted by DARWIN-7B to have bandgap greater than 0 eV. As shown in Fig. R2 (Fig. 6 in the revised manuscript), a total of 393 2D materials with bandgap greater than 2.0 eV are found. After verifying the stability using labels from 2dMatpedia, 313 of them are found to meet the stability requirements. The stability prediction accuracy of ECSG for these materials is 79.6%. It demonstrates that ECSG can be used to screen out materials that meet specific practical applications, avoiding the ineffective screening of experimentally unstable materials and significantly improving the screening efficiency.

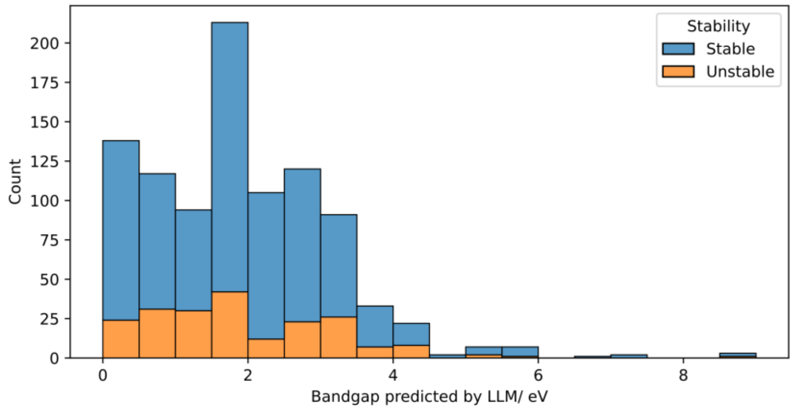

Fig. R2 The bandgap histogram of samples that are predicted to be thermodynamically stable. The legends of stable or unstable indicate the true labels in the 2DMatpedia database. The bandgap values were predicted by LLM DARWIN-7B.

References

[1] Ryu B, Wang L, Pu H, Chan MK, Chen J. Understanding, discovery, and synthesis of 2D materials enabled by machine learning. Chem. Soc. Rev. 51, 1899-1925 (2022).

[2] Zhou J, et al. 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. Sci. Data 6, 86 (2019).

[3] Gjerding MN, et al. Recent progress of the computational 2D materials database (C2DB). 2D Mater. 8, 044002 (2021).

[4] Xie, T., Wan, Y., Huang, W., Zhou, Y., Liu, Y., Linghu, Q., ... & Hoex, B. (2023). Large language models as master key: unlocking the secrets of materials science with GPT. arXiv preprint arXiv:2304.02213.

Comment 4. *Additionally, upon checking the code, the reviewer found that it still requires debugging to run the demo. Therefore, the authors need to address these issues.*

Authors' Response: We appreciate your valuable suggestions regarding code compatibility and debugging. We have taken steps to enhance the compatibility of our code across various hardware configurations, including Linux and Windows, to resolve potential issues that may arise during execution.

We found that most of the debugging situations are caused by the package *torch_scatter* because this package needs to call the *CUDA* backend to implement. Although it improves the training speed, it has strict requirements for the environment and the *PyTorch* and *CUDA* versions. Only a few fixed combinations of *PyTorch*, *CUDA*, and *torch_scatter* versions are allowed to call *torch_scatter.* To improve the compatibility of our code, we implemented the functions in *torch_scatter* with custom functions based on *PyTorch*. If the issues that need to be debugged are related to *torch_scatter*, users can uninstall this package. After uninstallation, our custom function will be called.

We believe these improvements will enhance user experience and provide a smoother execution of the code. We recognize that code that needs debugging may be a compatibility issue, so we have improved the compatibility of our code. If you have any questions or encounter any issues, please open an issue on GitHub (https://github.com/Haozou-csu/ECSG) or contact us at jxwang@mail.csu.edu.cn.

Reviewer #1 (Remarks on code availability):

Comment 5. *The code can be run after debugging. However, the results of the paper cannot be replicated as the database was not provided in the GitHub repository.*

Authors' Response: We appreciate your constructive feedback. For the reproducibility of our results, we have made several enhancements regarding data accessibility. In addition to providing the original download URLs for the databases, we have included preprocessed database files—specifically, MP_data.csv, JARVIS_data.csv, and OQMD_data.csv—in our GitHub repository. To reproduce our results, users can simply specify the path to the relevant database file when executing the code. We have included detailed instructions in the README file of the repository to guide users through this process, ensuring that they can successfully replicate our findings.

Answers to Reviewer #2

Reviewer #2 (Remarks to the Author):

Comment 1. *Overall, the discussion is well-structured and clear. The authors have done a good job explaining the motivation behind their work, the challenges in the field, and how their proposed method addresses these challenges. However, it could benefit from more technical details and context to support the claims made.*

Authors' Response: We appreciate your recognition of the value of our work and your insightful comments. We have addressed each of your comments individually, providing additional technical details to support our claims.

Comment 2. *Moreover, my feeling is that a graphical representation of the used metrics would help the general audience to better understand their relevance.*

Authors' Response: We appreciate your constructive feedback. To enhance the understanding of our performance metrics for a broader audience, we have incorporated radar charts and ROC curves into the revised manuscript. As shown in Fig. R3 (Fig. 2 in the revised manuscript), these visualizations provide a clear comparative analysis of the performance metrics across various models.

We hope these improvements meet your expectations and provide clearer insights into our work.

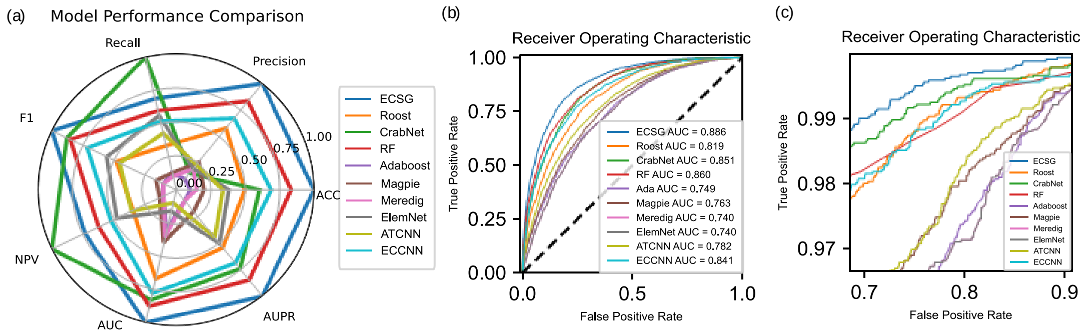

Fig. R3 (a) The radar plot of different models in terms of seven metrics, **(b)** AUC curves, and **(c)** zoomed AUC curves of different models.

Comment 3.

Minor editorial suggestions

Fig 1.: Nodes should be in the layer above the connections.

Fig 7: the 2 and 3 of Al2O3 collide with the surrounding box.

Consider refraining from using the term "impressive" in the assessment of the models all over the manuscript.

Authors' Response: Thank you for your helpful suggestions. We have replaced Fig. 1 and Fig. 7 with updated versions that correct any previous inaccuracies. We have also addressed your concern regarding the use of the term "impressive." We have replaced it with more precise and objective language throughout the manuscript.

Scientific technical questions and remarks:

Comment 4. *The authors mainly use AUC to assess the models` performance, however, there is no addressing of how they avoided the known problem of poor classification performance, where AUC incorporates irrelevant areas. Furthermore, relying solely on the outcome of the AUC would result in a lack of precision and negative predictive value information due to the fact that AUC focuses on sensitivity and specificity but does not provide information regarding precision or negative predictive value. For the ablation study section it would be more representative to present the outcome of the other tested metrics mentioned within the manuscript, similar to the presentation of performance in the prediction in unknown space section. Could you please elaborate on that in the manuscript? （section ablation study and section comparison）*

Authors’ Response: Thank you for your insightful comments. To provide a more comprehensive assessment of our model’ s performance, we have incorporated additional metrics such as precision, recall, and negative predictive value (NPV). The metrics—including accuracy (ACC), precision, recall, F1-score, NPV, area under the curve (AUC), and area under the precision-recall curve (AUPR)—are detailed in Fig. R3.

In the ablation experiment, we applied these metrics to evaluate the impact of various model components on performance. This thorough comparison allows us to assess each model's contributions more accurately. The updated results are provided in Table R2 (now Table 1 in the revised manuscript) for better clarity and alignment with our discussion.

Table R2 The performance of combining different base models in ECSG in the MP database. M1, M2, and M3 represent the base models ECCNN, Roost, and Magpie respectively.

Model	ACC	Precision	Recall	F1	NPV	AUC	AUPR
M1	0.766	0.727	0.669	0.697	0.788	0.842	0.770
M2	0.741	0.712	0.603	0.652	0.758	0.820	0.740
M3	0.702	0.662	0.532	0.590	0.721	0.766	0.670
M1+M2	0.805	0.776	0.725	0.750	0.823	0.883	0.828
M1+M3	0.779	0.733	0.711	0.722	0.809	0.861	0.800
M2+M3	0.799	0.761	0.729	0.745	0.822	0.873	0.813
M1+M2+M3	**0.807**	**0.778**	**0.728**	**0.752**	**0.824**	**0.886**	**0.834**

Comment 5. *Related to the point that ECSG outperforms other models in terms of classification metrics and sample efficiency. However, it would be beneficial to include more specific details about these base-level models for comparison. For example, further details about how the nature and hypothesis used in construction of the tested models is influencing the classification metrics and sample efficiency. Could you please comment on this?*

Authors’ Response: We appreciate your thoughtful feedback. In the revised manuscript, we have added a comprehensive description of the base-level models and their influence on stability classification performance and sample efficiency.

ECSG contains three base-level models: Magpie, Roost and ECSG. Each base-level model contributes to the feature construction and prediction processes within the ECSG framework:

Magpie: This model emphasizes the importance of including statistical features derived from various elemental properties, such as atomic number, atomic mass, and atomic radius [1]. The statistical features encompass mean, mean absolute deviation, range, minimum, maximum, and mode. This broad range of properties captures the diversity among materials, providing sufficient information for accurately predicting their thermodynamic properties.

Roost: Roost conceptualizes the chemical formula as a complete graph of elements, employing graph neural networks to learn the relationships and message-passing processes among atoms [2]. By incorporating an attention mechanism, Roost effectively captures the interatomic interactions that play a critical role in determining the thermodynamic stability of materials.

ECCNN: This model further advances the analysis by focusing on the electron configuration of atoms. Since the electron configuration directly influences the total energy of the material, it is a crucial input for calculating thermodynamic properties using density functional theory (DFT) [3]. ECCNN constructs the EC vector, generates the EC feature matrix, and utilizes convolutional neural networks (CNNs) to extract complex high-dimensional features related to the electron structure, enhancing the prediction of thermodynamic stability.

The features provided by these three base-level models are both relevant and complementary, significantly enriching the input information for the ECSG model. The richness of this information is a key factor influencing sample efficiency. Specifically, Magpie supplies element-based physical properties, Roost captures interatomic interactions, and ECCNN delves into the electron configuration level, collectively enhancing the model's ability to make accurate predictions with limited data.

We added these technical details about base-level models and how they affect prediction performance in the subsection ‘Model development’.

Reference

[1] Ward L, Agrawal A, Choudhary A, Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput. Mater. 2, 16028 (2016).

[2] Goodall REA, Lee AA. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nat. Commun. 11, 6280 (2020).

[3] Tofanelli, M. A., & Ackerson, C. J. (2012). Superatom electron configuration predicts thermal stability of Au25 (SR) 18 nanoclusters. Journal of the American Chemical Society, 134(41), 16937-16940.

Comment 6. *The authors explain that their method expands the parameter space and reduces the error between predictions and ground truth. However, the explanation could benefit from more technical details. How exactly does the method combine multiple models? How will a different method of combining the different models affect the propagation of errors? Moreover, how does it ensure the complementarity of different base-level models? Could you please add a paragraph to make this clearer?*

Authors’ Response: Thanks for your insightful comments. We have added further technical details on the combination method used in ECSG and discussed how different approaches to combining models affect error propagation and complementarity. Below, we address these points in detail:

(1) *How does ECSG combine multiple models?*

The ECSG framework uses stacked generalization (SG) [1], a technique that builds an optimal weighted combination of predictions from multiple base-level models. In our study, the base models include ECCNN, Roost, and Magpie, denoted as f_1, f_2, and f_3, respectively. A multi-response linear regression (MLR) model, acting as the meta-level model, assigns non-negative weights (ω_1,

ω_2, ω_3) to the outputs of the base models. The final ECSG output is a linear combination:

$$F(y_1, y_2, y_3) = \omega_1 y_1 + \omega_2 y_2 + \omega_3 y_3 + \varepsilon \ ,$$

where ω_1, ω_2, ω_3, and ε are learnable parameters.

To train the meta-model, we applied a five-fold cross-validation on the training data. As shown in Fig. R4 (Fig. 9 in the revised manuscript), each base model is trained on four folds and evaluated on the remaining fold, repeating the process five times to generate predictions across all folds. These predictions are concatenated to form inputs for the meta-level model, which is then trained to learn the optimal weights. This combination method ensures that the final ECSG model captures diverse patterns from the base-level models.

We have added the above combination steps in section 'Methods'

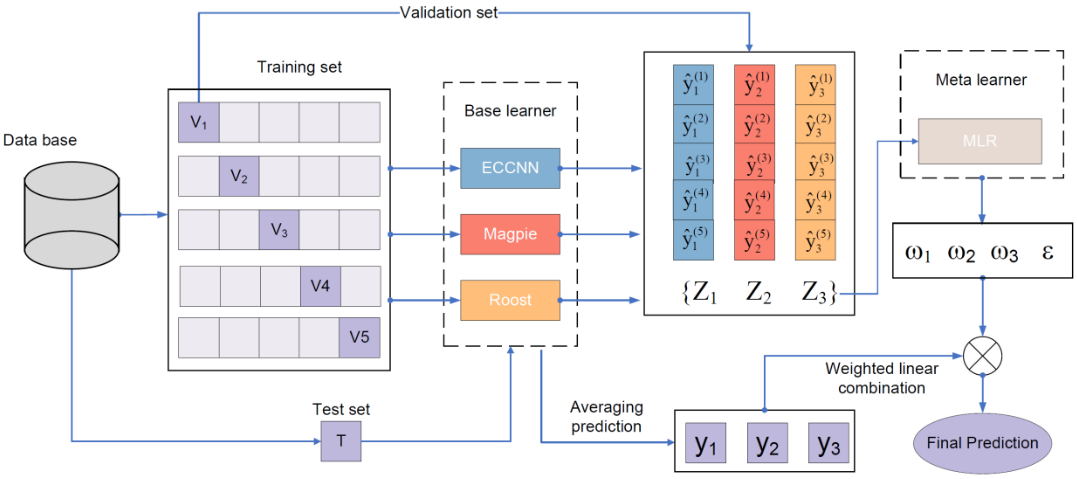
Validation set
Training set
Data base
V1
V2
V3
V4
V5
Base learner
ECCNN
Magpie
Roost
$\hat{y}_1^{(1)}$ $\hat{y}_2^{(1)}$ $\hat{y}_3^{(1)}$
$\hat{y}_1^{(2)}$ $\hat{y}_2^{(2)}$ $\hat{y}_3^{(2)}$
$\hat{y}_1^{(3)}$ $\hat{y}_2^{(3)}$ $\hat{y}_3^{(3)}$
$\hat{y}_1^{(4)}$ $\hat{y}_2^{(4)}$ $\hat{y}_3^{(4)}$
$\hat{y}_1^{(5)}$ $\hat{y}_2^{(5)}$ $\hat{y}_3^{(5)}$
$\{Z_1 \ Z_2 \ Z_3\}$
Meta learner
MLR
ω_1 ω_2 ω_3 ε
Weighted linear combination
Test set
T
Averaging prediction
y_1 y_2 y_3
Final Prediction

Fig. R4 The combination process of base-level models in ECSG.

(2) *How do different combination methods impact error propagation?*

In addition to SG, we explored other common ensemble methods, including averaging and voting [2]. As shown in Table R3 (Supplementary Table 7), SG produced the best performance across all metrics. While averaging applies fixed weights to base-model outputs, it can amplify errors from underperforming models if weights are not properly assigned. Voting, on the other hand, assigns equal importance to all models, limiting the ensemble's flexibility. In contrast, SG dynamically learns optimal weights, which helps minimize error propagation by assigning more weight to better-performing models. This adaptive weighting mechanism is particularly beneficial when base models exhibit varying strengths across different samples.

Table R3 The performance of integrating the three base models using different combination methods.

Methods	ACC	Precision	Recall	F1	NPV	AUC	AUPR
SG	**0.807**	**0.778**	**0.728**	**0.752**	**0.824**	**0.886**	**0.834**
Averaging	0.788	0.772	0.673	0.719	0.797	0.865	0.804
Voting	0.763	0.726	0.663	0.693	0.785	0.747	0.762

We have added the above results of different combination methods in Section 'Results'.

(3) *How does ECSG ensure complementarity among base-level models?*

We ensured complementarity by selecting base models from distinct knowledge domains: Magpie focuses on elemental properties, Roost captures interatomic interactions using attention mechanisms, and ECCNN analyzes electron configurations at a finer level. Each model contributes unique insights, enriching the ensemble’ s overall predictive power.

To validate complementarity, we performed error correlation analysis and assessed the entropy distribution of each model’ s predictions. As shown in Fig. R5(a) (Supplementary Fig. 3(a)), the Pearson correlation coefficients between model errors range from 0.37 to 0.49, indicating weak correlations and minimal redundancy among models. The entropy distributions in Fig. R5(b) (Supplementary Fig. 3(b)) further illustrate their complementary nature: Roost is concentrated in low-entropy regions, Magpie in high-entropy regions, and ECCNN maintains a balanced distribution. This diversity in predictions enhances ECSG’ s ability to generalize across different datasets and material types.

To ensure complementarity, we have selected domain knowledge from different scales: interatomic interactions, atomic properties, and EC. Models based on different domain knowledge have their own advantages. More specifically, the input of Magpie is the statistics of various element properties, Roost considers the interaction between atoms through the attention mechanism, and ECCNN goes a step further to the level of the electron configuration of atoms. We validated this through error correlation analysis and entropy distribution of each model. The error correlation matrix is a common tool to measure the correlation between the prediction errors of multiple models on the same data set. The smaller the correlation, the stronger the complementarity. The entropy distribution is also used to analyze the difference in uncertainty of the models in prediction. If one model has high uncertainty on some samples and the other model does not, it indicates that they are complementary, as shown in Fig. R5 (Supplementary Fig. 3).

We provided the complementarity of different base-level models in Supplementary Note 2.

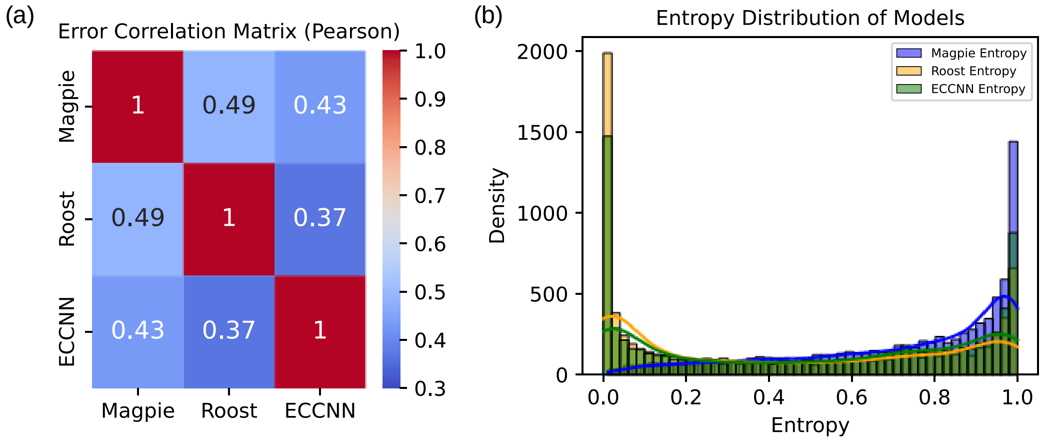

Fig. R5 (a) Error correlation matrix of three base-level models. (b) Entropy Distribution of three base-level models.

We included these details in the revised manuscript under the “Methods” and “Results” sections to clarify the model combination process and demonstrate the benefits of using SG.

References

[1] Martinez-Gil J. A comprehensive review of stacking methods for semantic similarity

measurement. Mach. Learn. Appl. 10, 100423 (2022).

[2] Zhou, W. et al. Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. Nat. Commun. 12, 1259 (2021).

Comment 7. *Starting from line 640 "Our method also offers advantages in integrating heterogeneous data in materials science and engineering. Data in these fields often exhibit heterogeneity, encompassing numerical table data, spectra data, and image data." This is a significant contribution. However, there is no description or details of how the proposed method integrated the heterogeneous data within any of the tested cases or models. How does the model handle different types of data? How does it preserve data integrity?*

Authors' Response: We appreciate your insightful comments and suggestions. In the revised manuscript, we have added a detailed discussion on how the ECSG model handles heterogeneous data.

ECSG leverages stacked generalization (SG), a highly flexible framework, to combine predictions from multiple models. This flexibility enables the seamless integration of new models designed to handle different data types, such as numerical data, spectral data, and image data.

When new base models are introduced, the meta-level model dynamically adjusts the weights for all base models during training. For example, the original ECSG model combines predictions from three base models (ECCNN, Roost, and Magpie), with the meta-level model calculating the weighted prediction: $\hat{y} = \omega_1\hat{y}_1 + \omega_2\hat{y}_2 + \omega_3\hat{y}_3 + \varepsilon$, where $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$ are predictions from the initial models, ω_1, ω_2, ω_3 are their corresponding weights, and ε is the interception. To incorporate additional data types, such as spectral data, we can introduce a new base model, K_dos_fea [1], which extracts features from spectra using one-dimensional convolution. This model generates a new prediction $\hat{y}_4$. Upon adding the spectral model, the meta-level model updates the weighted combination as follows: $\hat{y} = \omega_1^{'}\hat{y}_1 + \omega_2^{'}\hat{y}_2 + \omega_3^{'}\hat{y}_3 + \omega_4^{'}\hat{y}_4 + \varepsilon'$.

Similarly, new base models using CNNs can be introduced to handle image data, with the meta-model dynamically adjusting the weights for all sub-models. This modular approach ensures that ECSG can efficiently integrate diverse data sources, assigning appropriate weights to each type for accurate predictions.

Each base-level model processes only its respective data type, and the training of these models remains independent. Since the inputs are not fused or converted, data integrity is maintained throughout the process, avoiding potential information loss.

We have included the discussion of integrating heterogeneous data in Supplementary Note 3.

References

[1] Fung V, Hu G, Ganesh P, Sumpter BG. Machine learned features from density of states for accurate adsorption energy prediction. Nat. Commun. 12, 88 (2021).

Comment 8. *For future work, the authors mention plans to apply ECSG to other material properties. It would be an interesting element in the final paragraph of the manuscript to address the following question: What specific properties will you investigate? What challenges do you anticipate?*

Authors’ Response: We appreciate your thoughtful suggestion regarding future work. In the revised manuscript, we have added a paragraph in the 'Discussion' section that highlights the potential applications of the ECSG model and the challenges.

We plan to extend the ECSG model to predict several critical material properties, including bandgap, Young's modulus, and alloy hardness. These properties are essential for a wide range of applications in materials science and engineering. For instance, the bandgap is a key determinant of electrical conductivity, making it a crucial parameter in semiconductor and photovoltaic technologies.

However, we recognize several challenges in expanding ECSG’s scope to these properties. While compositional data provides a robust foundation, it primarily captures the ratios of elements within a material, often neglecting the spatial arrangement of these elements. This limitation becomes significant when predicting properties sensitive to crystal symmetry and doping levels, such as the bandgap. In these cases, compositional data alone may lack the physical and chemical context needed to capture subtle variations, hindering prediction accuracy.

Alloys can form multiple phases, including solid solutions, intermetallic compounds, and amorphous phases. Each phase exhibits distinct chemical compositions and interactions that determine the alloy's overall properties. A key challenge in applying ECSG to alloys lies in the need to predefine possible phases and their corresponding compositions to ensure accurate predictions.

Moving forward, we plan to address these challenges by developing structure-based models and phase composition prediction models. We will integrate these models into ECSG to capture both compositional and structural aspects, enhancing the framework's ability to predict complex properties accurately. This ensembling approach will allow ECSG to bridge the gap between compositional and spatial information, making it more effective across diverse material systems.

Answers to Reviewer #3

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.
Authors' Response: We sincerely thank you for the thorough reading of our manuscript and for providing valuable comments and insights.

Answers to Reviewer #4

Reviewer #4 (Remarks to the Author):

The authors present a new machine learning framework, called ECSG (Electron Configuration models with Stacked Generalization), for the prediction of materials stability, in which three composition-based sub-models Magpie, Roost, and a new architecture ECCNN (Electron Configuration Convolutional Neural Network) are combined by stacked generalization into a super-learner. The idea behind the ECSG framework is to limit the bias inherent in sub-models' constructions of the composition-property relationship by unifying the models into a single framework. The authors demonstrate the ability of ECSG to outperform each of the three sub-models as well as several other composition-based machine learning models in predicting the stability of inorganic compounds from three large DFT databases (MP, OQMD, and JARVIS). Additionally, they demonstrate the ability of ECSG to predict new stable compounds in the largely unexplored family of double perovskite oxides. The ECSG framework is clearly impressive and useful, although it appears to be only marginally better than some models and is missing consideration of crystal structure. The manuscript is well written, and I believe it is suitable for publication. I have some questions and comments about the work, which are below.

Comment 1. T*he authors use only composition-based models rather than structure-based ones. Their justification is that composition-based models are advantageous over structure-based models in that it is easier to acquire compositional information compared to detailed structure data. However, DFT databases (MP, OQMD, and JARVIS) contain crystal structures for every entry. Why not include this structure information into the ECSG model? This would enable us to compare ECSG with existing structure-based models such as CGCNN (Xie and Grossman, Phys. Rev. Lett. 120, 145301, 2018). Also, if the model is composition-based, would ECSG give us the same stability for two different crystal structures having the same chemical composition? If so, how would we know that Na2WNiO6, one of the stable perovskite oxides recommended by ECSG, does not have a polymorph that is lower in energy than the perovskite structure?*

Authors' Response: Thank you for your insightful comments. We chose compositional information as input because structural data is often agnostic when exploring new materials, and relying on composition offers a more practical and resource-efficient approach. While databases such as the Materials Project (MP) provide extensive structural information, such data is often unavailable or challenging to acquire for uncharacterized materials. Structural characterization typically requires complex experimental techniques, such as X-ray diffraction or electron microscopy, or computationally demanding methods like Density Functional Theory (DFT). These methods are time-consuming, costly, and require significant expertise and specialized equipment. In contrast, compositional data can be easily obtained by sampling the compositional space, making it well-suited for high-throughput screening and accelerating the discovery of new materials.

Since ECSG is primarily composition-based, it assigns identical predictions to materials with the same chemical composition, such as different polymorphs of a perovskite oxide. Recognizing this limitation, we explored the integration of structure-based models into ECSG to assess how structural data could enhance performance. Specifically, we incorporated the Crystal Graph Convolutional Neural Network (CGCNN) [1] as a base-level model and developed a hybrid model, ECSG+C, which combines predictions from both composition-only ECSG and CGCNN. The final prediction is obtained through SG combination of the base models' outputs.

To assess the impact of incorporating structural data, we downloaded 125,451 structural datasets from the MP, referred to as the MP-structure dataset, covering 89,204 unique compositions. We split these datasets into training and test sets in an 8:2 ratio, ensuring the split was based on composition. The test set included 1,471 samples with polymorphs, from which we paired stable and unstable materials with identical compositions, resulting in 1,038 polymorph pairs. We then evaluated the models' ability to differentiate between these pairs, represented by ACC_M in Table R4 (Table 2 in the revised manuscript).

Table R4 Performance of ECSG after integrating CGCNN on the MP-structure database. ECSG+C represents the model after integrating CGCNN into ECSG, and ACC_M denotes the accuracy in correctly distinguishing polymorphs.

	Accuracy	Precision	Recall	F1	NPV	AUC	AUPR	ACC_ M
ECSG	0.826	0.719	0.557	0.628	0.853	0.879	0.721	0
CGCNN	0.835	0.738	0.578	0.648	0.860	0.899	0.746	**0.193**
ECSG+C	**0.844**	**0.753**	**0.607**	**0.672**	**0.869**	**0.905**	**0.769**	0.121

The inclusion of structural information through CGCNN improves the overall predictive performance of ECSG+C compared to composition-only ECSG. However, distinguishing polymorphs remains a challenge for all models. The CGCNN model correctly identified 19.3% of polymorph pairs (ACC_M = 0.193), whereas ECSG+C achieved 12.1% (ACC_M = 0.121), indicating a slight reduction in polymorph differentiation after structural integration.

These results highlight the trade-off between enhanced general predictive accuracy and the challenge of differentiating polymorphs. While ECSG+C outperforms both ECSG and CGCNN in most metrics, further research is needed to improve sensitivity to polymorph differences without compromising accuracy.

We have added a new subsection, ‘Integration of Structural Information,’ in the “Results” section to discuss these findings. Additionally, our GitHub repository has been updated with code to enable users to predict thermodynamic stability using CIF files, allowing ECSG to incorporate structural models where such data is available

References

[1] Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. Phys Rev Lett 120, 145301 (2018).

Comment 2. *The “TRUE” perovskite oxides in Supplementary Table 6 are predicted to lie on the convex hull of the Materials Project database, correct? I wonder how many of these compounds would also lie on the convex hull of the OQMD and JARVIS databases, as these databases have different compounds in them, although many compounds are identical between the databases.*

Authors’ Response: Thank you for your valuable inquiry regarding the classification of the perovskite oxides listed in Supplementary Table 6. Yes, the term "TRUE" denotes the perovskite oxides predicted to lie on the convex hull of the Materials Project (MP) database, indicating their thermodynamic stability according to the MP dataset.

To evaluate the consistency of our predictions across multiple datasets, we analyzed how many of the stable double perovskite oxides identified by ECSG also lie on the convex hulls of the Open Quantum Materials Database (OQMD) and JARVIS databases. While these databases share many compounds, they also contain unique entries that can influence stability predictions.

As shown in Table R5 (Supplementary Table 9), our analysis indicates that 26 out of the 35 stable double perovskite oxides identified by ECSG are confirmed to lie on the convex hull of the OQMD database, while 32 compounds are stable according to the JARVIS database. In comparison, there are 25 out of 35 stable compounds in the MP database.

These results demonstrate the robustness of our model across multiple datasets, underscoring the consistency of ECSG’ s predictions even when different databases are used. This cross-database analysis highlights the reliability of our approach and confirms that ECSG consistently identifies stable compounds across different datasets. We appreciate your thoughtful question and hope this clarification enhances the understanding of our findings.

Table R5. VASP calculation results for stable perovskite oxides recommended by ECSG. The unit of total energy is eV per unit cell, and the unit of formation energy is eV per atom. Stability predictions are based on the convex hulls of the MP, OQMD, and JARVIS databases.

NO	Composition	Toal Energy	Formation Energy	Stability on MP	Stability on OQMD	Stability on JARVIS
1	Na2WNiO6	-65.853	-2.299	TRUE	TRUE	TRUE
2	Na2MnTbO6	-62.898	-2.498	TRUE	TRUE	TRUE
3	Ba2SmWO6	-79.120	-3.611	TRUE	TRUE	TRUE
4	YbPrMnNiO6	-72.760	-3.000	TRUE	TRUE	TRUE
5	PrGdV2O6	-84.592	-2.607	FALSE	FALSE	FALSE
6	TmInMn2O6	-73.733	-2.668	TRUE	TRUE	TRUE
7	LuPmCo2O6	-73.406	-2.842	TRUE	TRUE	TRUE
8	LaGdNi2O6	-71.186	-1.912	FALSE	FALSE	TRUE
9	YPdWCrO6	-82.612	-2.686	TRUE	TRUE	TRUE
10	YBaNbCoO6	-79.564	-3.247	TRUE	TRUE	TRUE
11	CuNdVCoO6	-71.906	-2.535	TRUE	TRUE	TRUE
12	NaGeFeCoO6	-59.822	-1.682	FALSE	TRUE	TRUE
13	NaMnWAlO6	-77.405	-2.873	TRUE	TRUE	TRUE
14	ZrLiWVO6	-82.252	-2.826	TRUE	TRUE	TRUE
15	HoMgMnMoO6	-75.066	-2.738	FALSE	FALSE	TRUE
16	DyRbYMoO6	-77.132	-3.273	TRUE	TRUE	TRUE
17	KScWSrO6	-72.552	-2.896	FALSE	FALSE	FALSE
18	HoPrMnCoO6	-78.255	-3.112	TRUE	TRUE	TRUE
19	HoMnMgMoO6	-75.956	-2.827	TRUE	TRUE	TRUE
20	HoYbCrTiO6	-82.337	-3.717	TRUE	TRUE	TRUE
21	LiDyYbMoO6	-71.594	-3.119	TRUE	TRUE	TRUE
22	LiYbMoDyO6	-73.111	-3.271	TRUE	TRUE	TRUE
23	YRbMoDyO6	-77.251	-3.285	TRUE	TRUE	TRUE
24	CrGdWNbO6	-86.851	-2.770	FALSE	FALSE	TRUE
25	NaTlFeMnO6	-60.931	-1.813	TRUE	TRUE	TRUE
26	WGdCrNbO6	-82.945	-1.466	FALSE	FALSE	FALSE
27	SrKWScO6	-79.139	-3.555	TRUE	TRUE	TRUE
28	SrYbFeBiO6	-63.481	-2.640	TRUE	TRUE	TRUE
29	LuAgFeMnO6	-70.250	-2.377	TRUE	TRUE	TRUE
30	KScMnTiO6	-77.666	-2.872	FALSE	TRUE	TRUE
31	KGdCuVO6	-67.731	-1.786	FALSE	FALSE	TRUE
32	KGdWMgO6	-75.396	-2.415	FALSE	FALSE	TRUE
33	MoSrWCrO6	-81.562	-2.491	TRUE	FALSE	TRUE
34	NdCsCrZnO6	-64.256	-2.618	TRUE	TRUE	TRUE
35	ScCaMnVO6	-80.052	-3.198	TRUE	TRUE	TRUE

Comment 3. *Personally, I would find it helpful if I had some idea of how the AUC and other scores translate to the actual numbers of compounds predicted to be stable or not. Without this I do not have a sense of which scores are good or bad. For example, in Table 1, the AUC score of ECSG is*

0.887 whereas for RF it is 0.862, so it appears to me that ECSG is only marginally better than RF.

Authors’ Response: Thank you for your valuable suggestions regarding the inclusion of model comparison metrics. To provide a more detailed understanding of model performance, we have included confusion matrices for ECSG and other models, as shown in Fig. R6. These matrices offer a breakdown of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), allowing for a more granular analysis of prediction accuracy. For example, as presented in Table 1, ECSG achieves an AUC score of 0.887, while RF has an AUC score of 0.862. The confusion matrix reveals that ECSG accurately identifies 2,489 stable compounds as true positives and 4,397 unstable compounds as true negatives, compared to RF’s 2,383 true positives and 4,291 true negatives.

Notably, ECSG identifies 4.45% more stable compounds than RF. Furthermore, the number of false positives predicted by RF (787 samples) is 15.6% higher than that of ECSG (681 samples). These results underscore ECSG’ s improved ability to minimize false positives, thereby enhancing its reliability in practical applications.

We have added the confusion matrix in Supplementary Fig. 2.

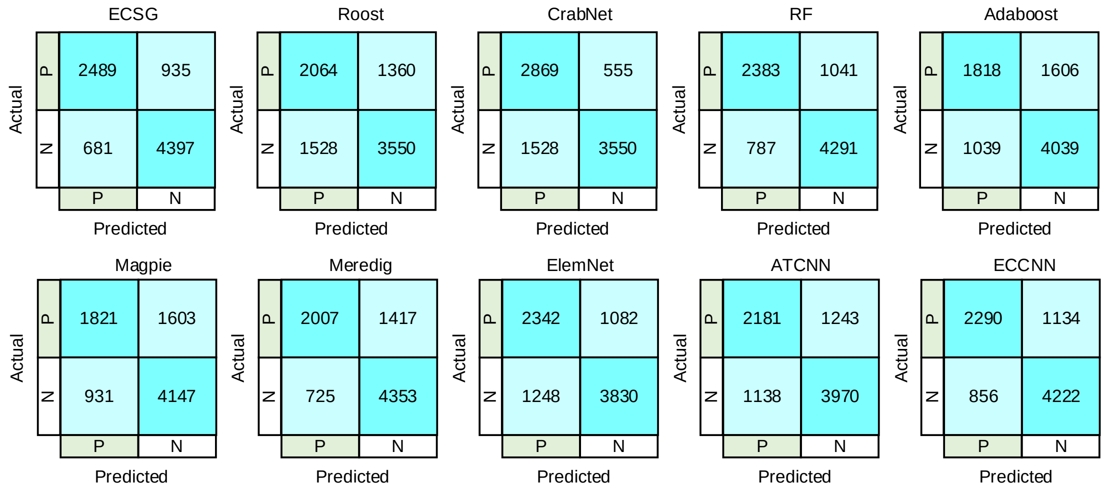

Fig. R6 Confusion matrices for ECSG and other composition-based models. 'P' and 'N' indicate the total number of positive and negative samples, respectively. 'Actual' denotes the true class labels, while 'Predicted' represents the classifications made by the models.

Comment 4. *This may be nitpicky, but how exactly is ΔHd defined? If it is the “energy above the hull”, then it can never be negative, but the “TRUE” DFT values in Supplementary Table 6 are negative. I think in this case ΔHd is the energy above the convex hull of compounds in the Materials Project at that moment in time.*

Authors’ Response: Thank you for your valuable comments. In this study, the decomposition energy (ΔH_d) is defined as the total energy difference between a given compound and its competing compounds in a specific chemical space. It reflects the magnitude of (in)stability with respect to phase separation [1].

The concept of ΔH_d differs from the commonly used “energy above the hull.” To determine the ΔH_d of a compound, the compound must be excluded when constructing the energy convex hull of the system in which it resides. The ΔH_d then corresponds to the distance from the compound’ s energy to the hull.

As illustrated in Fig. R7 (Supplementary Fig. 1), the ΔH_d of A_4B is calculated as the distance from the convex hull to A_4B. If the compound lies above the energy hull, ΔH_d will be larger than zero, corresponding to the commonly reported "energy above the hull". However, for compounds that lie on the convex hull, such as AB_3, the ΔH_d represents the distance from AB_3 to a hypothetical convex hull constructed without AB_3 (indicated by the dashed line in Fig. R7). In this scenario, the ΔH_d value becomes negative, quantifying the compound' s stability. This negative value offers valuable insights into the uncertainty of stability assessments and guides the rational design of synthesis pathways [2].

The detailed definition and discussion of ΔH_d are provided in Supplementary Note 1.

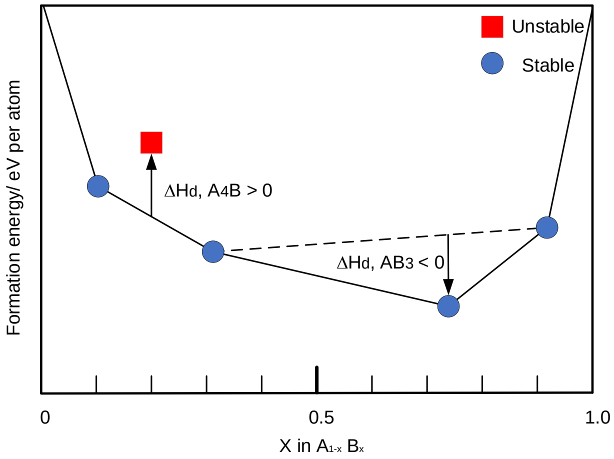

Fig. R7 Illustration of the definition of ΔHd by constructing a convex hull.

References

[1] Bartel, C. J. et al. A critical examination of compound stability predictions from machine-learned formation energies. npj Comput. Mater. 6, 97 (2020).

[2] Bartel, C. J. Review of computational approaches to predict the thermodynamic stability of inorganic solids. *J. Mater. Sci.* **57**, 10475-10498 (2022).

Comment 5. *Does the "stability probability" outputted by ECSG translate to the real probability that the compound is stable? For example, should I expect 20% of compounds with probability of 0.2 to be stable? If not, then I do not see the value of the distribution in Figure 5a, except that the compounds with "probability" greater than 0.9 represent a tiny fraction of all candidates. It is impressive that 25 of the 35 predicted-stable perovskite oxides were confirmed to be stable by DFT, but I have to wonder whether other composition-based models would perform nearly as well as ECSG in this case.*

Authors' Response: Thank you for your insightful question. The "stability probability" generated by ECSG should be interpreted as a confidence score rather than a direct probability in the frequentist sense. While this confidence score ranges from 0 to 1, it quantifies the model's certainty in its prediction. A higher score reflects greater confidence in the prediction outcome.

The primary purpose of Figure 5(a) is to rank candidate materials based on their stability scores and facilitate the selection of a suitable subset for further experimental verification. We prioritize candidates with scores greater than 0.9 to focus on those with the highest predicted stability, thereby increasing the likelihood of identifying promising materials for future research and validation.

To further validate ECSG's performance, we compared it with Tala, a composition-based model developed in Ref. [1]. We used Tala to screen the top 35 double perovskite oxides with the highest predicted stability scores. After performing DFT calculations on these 35 compounds, we computed their total energies and mapped them onto the convex hull of the Materials Project (MP).

As demonstrated in Table R6 (Supplementary Table 10), only two of the 35 compounds identified by Tala are stable based on the MP convex hull. In contrast, ECSG identifies a significantly higher number of stable perovskite oxides. This comparison underscores the superiority of ECSG in accurately predicting stable candidates and highlights its potential for exploring novel double perovskite oxides.

We have incorporated this comparison and the corresponding discussion in the revised manuscript.

References

[1] Talapatra A, Uberuaga BP, Stanek CR, Pilania G. A Machine Learning Approach for the Prediction of Formability and Thermodynamic Stability of Single and Double Perovskite Oxides. Chem. Mater. 33, 845-858 (2021).

Table R6. VASP calculation results for stable perovskite oxides recommended by Tala. The unit of total energy is eV per unit cell, and the unit of formation energy is eV per atom. Stability predictions are based on the convex hulls of the MP, OQMD, and JARVIS databases.

NO	Composition	Toal Energy	Formation Energy	Stability on MP	Stability on OQMD	Stability on JARVIS
1	TeRbReSrO6	-62.108	-1.959	FALSE	FALSE	TRUE
2	NdTaReSrO6	-77.074	-2.206	FALSE	FALSE	FALSE
3	FeAuReSrO6	-63.617	-1.574	FALSE	FALSE	FALSE
4	PbIrReSrO6	-65.133	-1.419	FALSE	FALSE	FALSE
5	NbLaReSrO6	-74.556	-2.113	FALSE	FALSE	FALSE
6	NbLaBaReO6	-72.698	-1.904	FALSE	FALSE	FALSE
7	NdNbReCdO6	-73.250	-2.077	FALSE	FALSE	FALSE
8	PrTaMgReO6	-78.822	-2.388	FALSE	FALSE	FALSE
9	TcAuReSrO6	-64.868	-1.284	FALSE	FALSE	FALSE
10	PrNbMgReO6	-78.459	-2.528	FALSE	FALSE	FALSE
11	MoSmReSrO6	-73.192	-2.244	FALSE	FALSE	FALSE
12	NaTeReSrO6	-61.902	-1.905	FALSE	FALSE	FALSE
13	WNdReSrO6	-76.940	-2.526	FALSE	FALSE	FALSE
14	WPdReSrO6	-67.101	-1.501	FALSE	FALSE	FALSE
15	HfNdMgReO6	-81.463	-2.844	FALSE	FALSE	FALSE
16	ZrNdReSrO6	-77.656	-2.595	FALSE	FALSE	FALSE
17	CrNdReSrO6	-77.741	-2.693	TRUE	FALSE	TRUE
18	NbLaMgReO6	-77.058	-2.372	FALSE	FALSE	FALSE
19	BeAuReSrO6	-58.267	-1.286	FALSE	FALSE	FALSE
20	NiTlReSrO6	-62.661	-1.867	FALSE	FALSE	TRUE
21	NdNbReSrO6	-76.274	-2.301	FALSE	FALSE	FALSE
22	RhPrReCdO6	-68.668	-1.894	FALSE	FALSE	TRUE
23	TeFeReSrO6	-64.136	-1.639	FALSE	FALSE	FALSE
24	RuHgReSrO6	-61.543	-1.357	FALSE	FALSE	FALSE
25	PrNbReSrO6	-75.041	-2.177	FALSE	FALSE	FALSE
26	NdTaMnTiO6	-85.703	-2.943	FALSE	FALSE	FALSE
27	TaTiReSrO6	-70.910	-1.277	FALSE	FALSE	FALSE
28	CsNiReSrO6	-62.788	-2.026	FALSE	FALSE	TRUE
29	RhVReSrO6	-67.798	-1.468	FALSE	FALSE	FALSE
30	RhLaReSrO6	-71.547	-2.088	FALSE	FALSE	FALSE
31	RhPrReSrO6	-72.595	-2.208	FALSE	FALSE	FALSE
32	NbLaReCdO6	-70.940	-1.829	FALSE	FALSE	FALSE
33	TaLaMgReO6	-77.546	-2.245	FALSE	FALSE	FALSE
34	MoPmReSrO6	-78.511	-2.772	TRUE	TRUE	TRUE
35	CrPdReSrO6	-67.525	-1.630	FALSE	FALSE	FALSE

Reviewer #4 (Remarks on code availability):

Comment 6. *I had to install a different set of Torch libraries that were compatible with my GPU machine, but otherwise, I found the instructions in the README file to be straightforward. I did not try to reproduce the numbers in the manuscript, but the code appears to produce stability predictions for hypothetical compositions.*

Authors' Response: Thank you for your review and feedback on our work. We recognize that GPU-compatible Torch libraries may require adjustment depending on the user's hardware and environment. To ensure a smooth user experience across different configurations, we have added detailed instructions in the README file. These instructions address potential compatibility issues across different Torch versions by recommending custom functions to resolve them.

Our code requires Torch version 1.5 or higher since some dependencies are incompatible with earlier versions. For users experiencing issues, we encourage them to open an issue on GitHub (https://github.com/Haozou-csu/ECSG) or contact us directly at jxwang@mail.csu.edu.cn for further assistance.

To facilitate reproducibility, we have included: processed database files and hyperparameters required for training ECSG; pre-trained model files to allow users to validate our results quickly. We hope these updates will make it easier for readers to reproduce our results and experiment with the model.

Answers to Reviewer #4

Reviewer #4 (Remarks to the Author):

I have reviewed the authors' responses to reviewers and changes to the manuscript, and I find them to all be satisfactory. I appreciate the detailed answers to questions and feedback, as well as the additional clarifications and analyses that have been added to the manuscript. For these reasons, I strongly recommended publication of this work to Nature Communications.

Comment 1. *Thank you for the attempt to include CGCNN, a structure-based neural network, as a base model into ECSG (becoming ECSG+C). While this addition fortunately improves predictive performance compared to the ECSG, it is unfortunate that ECSG+C does a poor job at distinguishing stable from unstable polymorphs (as does CGCNN alone). Indeed, more research is needed to improve the ability of machine learning to distinguish the energetics of polymorphs. I wonder if the poor performance is due to the low energy difference between polymorphs of the same composition, which is lower than the resolution of the ECSG+C model.*
Authors' Response: We appreciate your recognition of the value of our work and your insightful comments. We agree that the poor performance in predicting polymorphs likely arises from the limited resolution to effectively distinguish between polymorphs of the same composition. Future research should focus on addressing the challenge of distinguishing the energetics of polymorphs. We have added this point in the manuscript in the section 'Integration of structure information'.

Comment 2. *I also appreciate the inclusion of stability assessments (Supplementary Tables 9 and 10) using not just the MP convex hull but also OQMD and JARVIS. The fact that there is some disagreement on stability assessments between the DFT databases raises an important issue with DFT-based stability assessments: that DFT data on competing phases is often incomplete. This is important to keep in mind when interpreting machine learning predictions of stability.*
Authors' Response: Thank you for your valuable comment regarding the current challenges with DFT-based stability assessments. We fully agree that the incompleteness of the databases poses a significant challenge, as materials previously considered stable may be reassessed as unstable when more competing phases are included. We have included an explanation in the Section "Case Study" when describing the stability based on different databases: It should be noted that the calculated stability results are based on the currently available databases. As these databases continue to expand and evolve, the calculated stability outcomes may be subject to change.

Reviewer #4 (Remarks on code availability):

Comment 3. *The README file now contains more information to aid users in installing the required Torch libraries, which can vary by GPU machine.*
Authors' Response: Thank you.