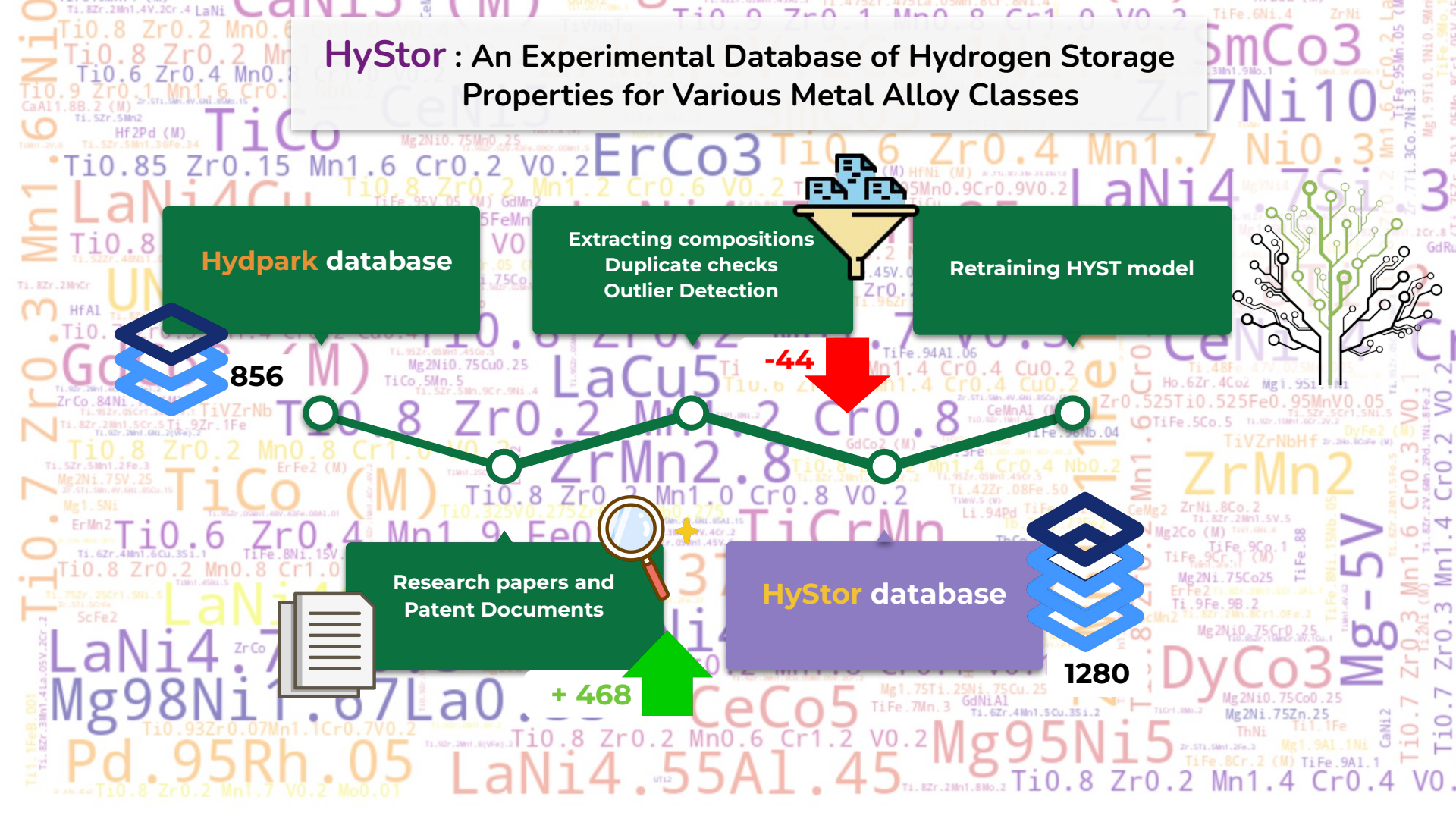
HyStor : An Experimental Database of Hydrogen Storage Properties for Various Metal Alloy Classes



- HyStor is an updated experimental database for metal alloys hydrogen storage properties as a function of absorption temperature.

- HyStor comprises of 1280 metal alloys spread across 10 different alloy classes and represents 54 elements.

- New alloy classes of LEA, MEA, and HEA are also included in the updated database.

- ML model HYST trained on HyStor performs better on benchmarking different alloy classes in comparison to existing HydPark database.

- The updates to alloy classes and element representations are also discussed.

HyStor: An Experimental Database of Hydrogen Storage Properties for Various Metal Alloy Classes

Nikhil Wilson[a], Ashwini Verma[a,b], Piyush Ranjan Maharana[a], Ameeya Bhusan Sahoo[a], Kavita Joshi*[a,b]

[a]*Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pashan, Pune-411008, India.*
[b]*Academy of Scientific and Innovative Research (AcSIR),Ghaziabad- 201002, India.*

Abstract

In this work, we introduce the HyStor database, consisting of 1280 metal alloys along with their hydrogen storage capacities (H2wt%) as a function of absorption temperature. Given the lack of updates in the existing open access HydPark database since 2002, we sourced compositions from recent research articles and various patent documents, resulting in a total of 468 compositions. The addition is reflected in the data across all existing classes of alloy compositions and low entropy alloys (LEA), medium entropy alloys (MEA), and high entropy alloys (HEA) have been included newly. This has broadened the scope of the database to encompass the latest materials of interest for hydrogen storage. HyStor contains representation of 54 elements, with a temperature range of 200-800K, and H2wt% range of 0.1-7.19. To ensure data quality, we conducted thorough checks for duplicate entries, erroneous data, and conflicting compositions within the database. Furthermore, we conducted multiple tests to identify potential outlier compositions by benchmarking the database against the pretrained HYST model on HydPark data. After eliminating these potential outliers, we successfully improved the error metrics of the HYST model, reducing the Mean Absolute Error (MAE) from 0.32 to 0.28 and increasing the R2 score from 0.78 to 0.82. We also tested individual classes and observed that the performance of the HYST model has increased for majority of the classes.

Keywords: Solid-State Hydrogen Storage; Metal Hydrides; Databases; Machine Learning

1. Introduction

For centuries, the discovery of new materials has relied on a combination of intuition and experimentation, or sometimes sheer chance. Now, as technology advances faster than ever, there is a demand for a more systematic and efficient approach. This is where machine learning (ML) discoveries steps in, a powerful statistical method combined with computation for unlocking unprecedented insights, accelerating discovery, and guiding the design of advanced materials with targeted properties.[1] At its core, machine learning leverages algorithms and statistical techniques to analyze datasets, identify

Email address: k.joshi@ncl.res.in (Kavita Joshi*)

patterns, and make predictions. However, the efficacy of ML models hinges crucially on data available for training and testing. The quality and diversity of the data directly impacts the accuracy of ML predictions.[2] Therefore, comprehensive and well-curated datasets plays a pivotal role in determining the success of ML applications in material research.

With the advent of powerful computing, the data available in the scientific domain is expanding exponentially. Leveraging computational simulations, researchers have generated a wealth of material property data and most of the data has been archived in publicly accessible computational databases.[3, 4, 5, 6, 7, 8] This available wealth of data has helped in developing ML models for prediction of material properties like, formation energy [9, 10], crystal structure [11], and catalytic activity [12] among others. The accuracy levels achieved by these ML models are close to the ab-initio calculations, marking significant progress in computational materials science research.[13] For example, with advances in machine learning potentials, it is possible to quickly evaluate the energies and forces in a system with close to ab-initio accuracy with reduced computational cost.[14, 15] In our previous work, we integrated deep learning and reinforcement learning with DFT calculations for cluster structure prediction. This approach facilitated the accelerated search for ground state structures of gallium clusters.[16, 17] While integrating computational methods with machine learning (ML) has demonstrably accelerated workflows in many areas, computational calculations themselves have limitations. Due to these limitations, accurately predicting advanced material properties, such as heat transport, stress-strain relationships, magnetic behavior, reaction dynamics, etc remains a challenge. [18] This necessitates the continued dependence on time-consuming and costly experimental route to measure material properties.

When attempting to apply machine learning in conjunction with experimental data, a notable scarcity of data arises. Few of the large open access experimental data-sets in material sciences are available for compositions, structures and phases like Inorganic Crystal Structure Database (ICSD), Crystallographic Open Database (COD), and High Throughput Experimental Materials (HTEM).[19, 20, 21] Material properties can exhibit significant variations depending on the synthesis route, temperature, and pressure used during the experiment. This variability highlights the importance of comprehensive data collection, including the specific experimental conditions. Unfortunately, the complete details of these experiments are often only documented in published scientific literature, not readily accessible in easily searchable databases. The possible solution for building large and diverse experimental materials data-sets is to extract information from these published research articles and patent documents. Traditionally, manual extraction is used to gather information from literature by meticulously reading each document to identify relevant material properties. However, this method is labor-intensive and time-consuming. Recent advancements in natural language processing (NLP) and web scraping offer faster alternatives for text mining. These techniques enable the extraction of materials synthesis

conditions and parameters from large corpus of articles. For instance, ChemDataExtractor is an NLP pipeline to automatically extract chemical information from scientific documents.[22] BatteryBERT, a pre-trained language model, utilises NLP in text mining battery research articles.[23] Although these methods automate the extraction process, they are also associated with drawbacks like handling of linguistic nuances, manual annotations for creation of training set, and generalizability. Therefore constructing an experimental database to train a model for targeted material property prediction is a complex task. Ensuring the accuracy and consistency of the extracted data is crucial for effective model training. This involves rigorous measures to identify and address potential errors or inconsistencies within the data.

In the field of solid-state hydrogen storage in metal hydrides, machine learning has proven effective in predicting material's crucial properties like hydrogen storage capacity[24, 25], enthalpy of hydride formation[26, 25], equilibrium plateau pressure[26] etc. Most of these models have been trained on HydPark database[27] and some are trained on manually extracted data from published literature[24]. Till date, HydPark is the only available database on metal hydrides, consisting of metal compositions and their hydrogen storage properties. However, this database's last update dates back to 2002 and suffers from numerous missing entries. Most of the ML models trained on HydPark have been employed to interpret composition – hydrogen storage property relationships.[24, 26] Recently, some models have been utilized to construct predictive models and propose potential families of alloys for hydrogen storage.[28, 25] Hattrick-Simpers et. al. developed an ML model and identified 6110 potential alloys as hydrogen compressors.[28] In our previous work, we have also used HydPark data to build models to predict hydrogen storage capacity as a function of temperature and enthalpy of hydride formation. Our models prediction identified 6480 compositions that meet the stationary storage target properties established by the US - Department of Energy. We also observed discrepancy between predicted and experimentally measured H2wt% at lower temperatures for Mg-based compositions. This mismatch likely stems from a bias in the training data. Specifically, the training set contained a dominance of Mg-based compositions with lower temperatures that also had lower H2wt%. Consequently, the model learned this correlation and became biased towards predicting lower H2wt% for unseen Mg-compositions at lower temperatures.[25] Therefore, the performance of these models exhibits a strong dependence on the available data. Thus, to expedite the search for materials suitable for hydrogen storage, a large and diverse experimental data-set is imperative.

In this work, we present a systematically updated HydPark database, named as HyStor comprising metal hydrides' hydrogen storage capacity as a function of absorption temperature, totaling 1280 data points. We utilized recent journal articles and multiple patent documents, to develop this database. Before appending the data from these sources, we checked these compositions for duplicate and conflicting entries. Our final dataset consists of compositions from ten different classes of metal

alloys AB2, AB, A2B, AB5, Mg, MIC, SS, LEA, MEA, and HEA. Notably, LEA, MEA, and HEA class of compositions were absent in the HydPark database. We benchmarked the performance of the machine learning (ML) model trained using HyStor against our previously trained model, HYST on HydPark[25]. Additionally, we have developed a framework to evaluate the reliability of data before updating the database. The retraining of the HYST model on HyStor has resulted in an improved error metrics with MAE of 0.28 and R2 score of 0.82 in comparison to the previous benchmarks of HYST on HydPark with MAE of 0.33 and R2 score of 0.80. We also compared the performace of HYST model trained on HydPark and HyStor database and observed a good deal of improvement in the error metrics using HyStor on individual alloy classes.

2. Results and Discussions

2.1. HyStor database

HyStor is an experimental database which comprises of 1280 metal alloys, with their H2wt% as a function of absorption temperature. The data recorded experimentally is spread across 10 different classes of metal alloys: A2B, AB, AB2, AB5, MIC, SS, Mg, LEA, MEA, and HEA with a representation of 54 elements as shown in Figure 1. The total number of compositions and the top eight frequently occurring elements in each class of the metal alloys are also noted in the figure.

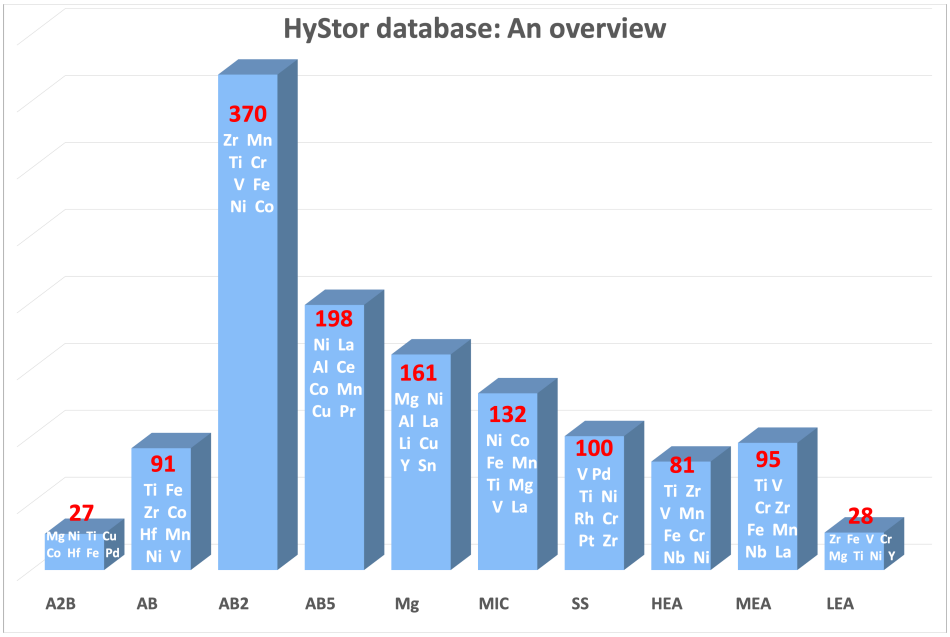


Figure 1: Represents the overall data distribution in different classes of the HyStor database. Each class is also indicated with the first eight most frequent elements present in that class and also the number of compositions in that class.

The process of extracting data from research articles and patent documents contributed 468 compositions. These updates are evident across all alloy classes. Moreover, the inclusion of compositions

from recent years into the database reflects the ongoing research on metal alloys, providing a comprehensive view of developments in this field. We will explore these observations and the performance of the ML model on these alloy classes after retraining with the new data in an upcoming section of this research article.

2.2. Data curation and pre-processing

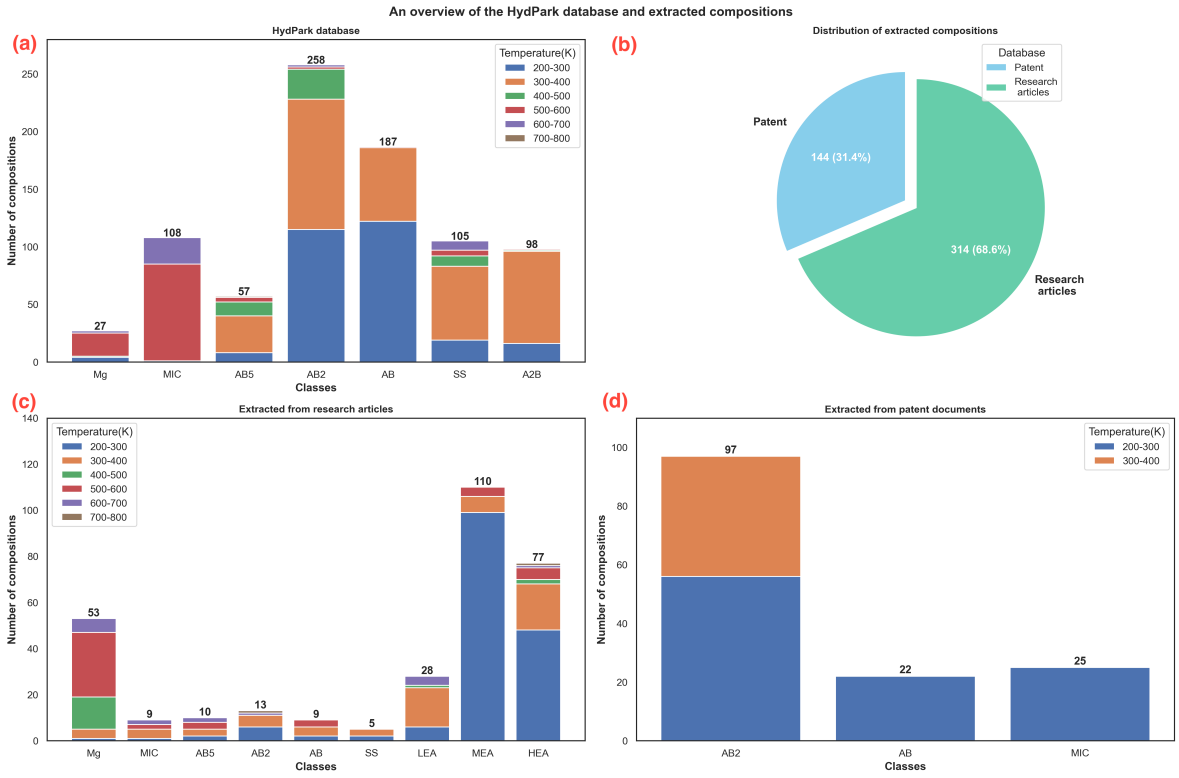


Figure 2: Represents the HydPark database and compositions extracted from research articles and patent documents distributed in different classes. (a) represents the overview of the compositions present in the HydPark database, (b) represents the overall distribution of extracted compositions from research articles and patents, (c) represents the extracted compositions from research articles and (d) represents the extracted compositions from multiple patent documents. Each class is also indicated with its absorption temperature ranges.

As a starting point, we pre-processed the HydPark database which contains 2,722 compositions and their hydrogen storage properties. As our target is prediction of H2wt% which is a function of absorption temperature, we require absorption temperature and corresponding H2wt% which was available only for 856 compositions. A brief overview of the HydPark database distributed in different classes and temperature regions is depicted in figure 2(a) Data curated from the research articles are 314 compositions and this data is spread across different alloy classes as shown in figure 2(c). The 144 compositions extracted from patent documents represents AB2, AB, and MIC classes of compositions as illustrated in figure 2(d). As seen from figure 2 (c) and (d), the extracted compositions also contribute to different temperature regions from 200-800K . The 20% of extracted compositions also represent higher temperature regions of 400-800K. Considering that the HydPark database comprises 23% of

compositions in this high-temperature range, the addition of these 20% compositions will increase the distribution in high temperature training data and will maintain a balance across different temperature regions. As H2wt% depends upon absorption temperature, the addition of compositions from different temperature regions will improve the predictive accuracy in those regions. The addition of new classes of compositions, such as LEA, MEA, and HEA, has significantly enriched the dataset with diverse materials. These alloys exhibit remarkable properties, including exceptional mechanical strength, high temperature stability, and show promising hydrogen storage performance at ambient conditions.

Ensuring the quality of data within a database is vital for accurate analysis and informed decision-making, particularly when designing new experiments. On inspecting the database, we notice few issues of data quality:

- Duplicated entries where two data points are seen with the same composition, temperature, and hydrogen storage capacity.
- Conflicting compositions where data-points with the same composition and temperature were recorded with a different hydrogen weight percent.
- Erroneous compositions which were recorded with the wrong hydrogen storage capacity in the HydPark database.

Composition	Temperature (K)	H2wt%	Cases
Mg1.5Ni	573	2.6	Duplicate composition
Mg1.5Ni	573	2.6	
Ti.98Zr.02Mn1.5V.43Fe.09Cr.05	293	1.9	Conflicting composition
Ti.98Zr.02Mn1.5V.43Fe.09Cr.05	293	1.6	
ZrNi1.4Mn.5V.5	303	1.6	Erroneous data
ZrNi1.4Mn.5V.5	303	1.5	Correct data

Table 1: Few representative cases of data quality checks performed on the HydPark and the extracted data points

During data quality checks, 17 pairs of duplicate compositions were identified. One of the examples of duplicate entry of Mg1.5Ni is shown in table 1. These duplicate entries are mostly due to the presence of same data point tagged in two different alloy classes such as in the case of Mg1.5Ni tagged in both AB and Mg alloy classes. In such cases we retain one of these compositions in our database. We also found three other pairs of compositions with the same temperatures but different H2wt% values. Upon investigation, we discovered that these differences were due to variations in pressure values as recorded in the original research articles. We also found three other pairs of compositions with varied synthesis routes, resulting in different H2wt% values. In cases where compositions exhibited differences in pressures and synthesis routes, only one composition is retained with the highest H2wt% value. Two erroneous entries from the existing Hydpark database were identified, and corrections were made based on information from their original research articles.

Dataset	Number of composition
Pre-processed HydPark database	856
Compositions extracted from research articles and patent documents	468
Compositions removed through dataquality tests	25
Compositions remaining in the HydPark dataset	842
Compositions remaining in the extracted dataset	457
Final data-points after data quality tests	1299

Table 2: Number of compositions in the existing database, extracted data-points and compositions identified through data quality checks

After the data quality checks, the number of compositions that are to be eliminated are 25. Table 2 indicates the number of compositions at each stage of the data quality checks. Number of compositions removed from the HydPark database are 14, while 11 compositions were removed from the extracted dataset. This results in 842 compositions remaining in HydPark and 457 compositions remaining in extracted dataset as noted in table 2. The final number of compositions after data quality tests in the HyStor database is 1299.

Having ensured the quality of the updated data, we can now refine the ML model. To achieve this, we will analyze how the individual data compositions affect the model's performance. This essentially involves the evaluation of each data point to understand its impact on the model's accuracy. Through this exercise, we aim to identify those compositions which are consistently predicted with higher error margins and investigate them in depth. This will allow us to take a judicious call on appending new extracted data from literature.

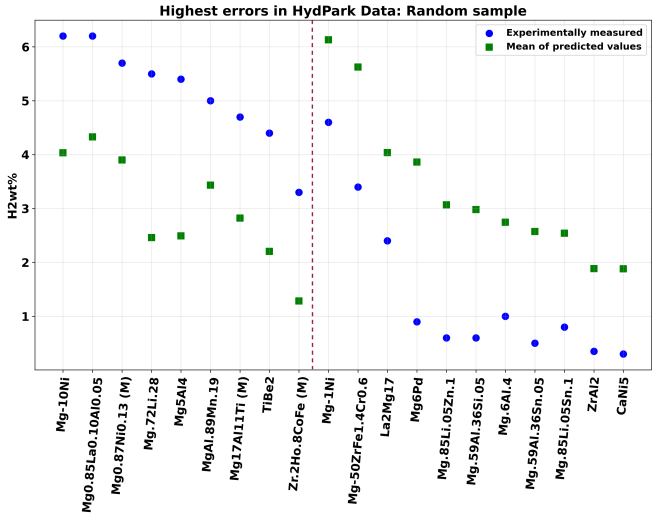
3. Anomaly and outlier detections:

In this section, we will conduct several tests to identify compositions that consistently exhibit poor performance during model predictions. Our objective is to detect anomalous compositions and potential outlier compositions through these tests. We conduct these tests initially on the existing HydPark database to eliminate potential outliers, and then proceed to conduct these tests on the extracted dataset.

3.1. 90/10 random data split on HydPark database

In this test, we randomly select 10% of the existing HydPark database as a hold-out dataset and use the remaining 90% for training the HYST model. We repeat this process 1000 times, shuffling the data each time, and calculate the error bars associated with individual compositions whenever they are present in the test set.

In figure 3, compositions are sorted based on two regions of ML predicted values. The region on the left represents compositions that are underpredicted by the model, whereas the region on the right are compositions that are overpredicted by the model. The accompanying table displays the frequency of these compositions occurring in the test set generated through random shuffling. Some compositions,



Composition	Frequency	Temperature(K)	Material_class
Mg-10Ni	99	573	Mg
Mg0.85La0.10Al0.05	87	573	Mg
Mg0.87Ni0.13 (M)	95	598	Mg
Mg.72Li.28	96	623	Mg
Mg5Al4	107	599	MIC
MgAl.89Mn.19	95	623	MIC
Mg17Al11Ti (M)	105	577	Mg
TiBe2	98	295	AB2
Zr.2Ho.8CoFe (M)	124	300	AB2
Mg-1Ni	107	573	Mg
Mg-50ZrFe1.4Cr0.6	103	573	Mg
La2Mg17	89	373	MIC
Mg6Pd	100	433	MIC
Mg.85Li.05Zn.1	110	583	Mg
Mg.59Al.36Si.05	99	583	Mg
Mg.6Al.4	93	583	Mg
Mg.59Al.36Sn.05	104	583	Mg
Mg.85Li.05Sn.1	87	583	Mg
ZrAl2	93	293	AB2
CaNi5	85	295	AB5

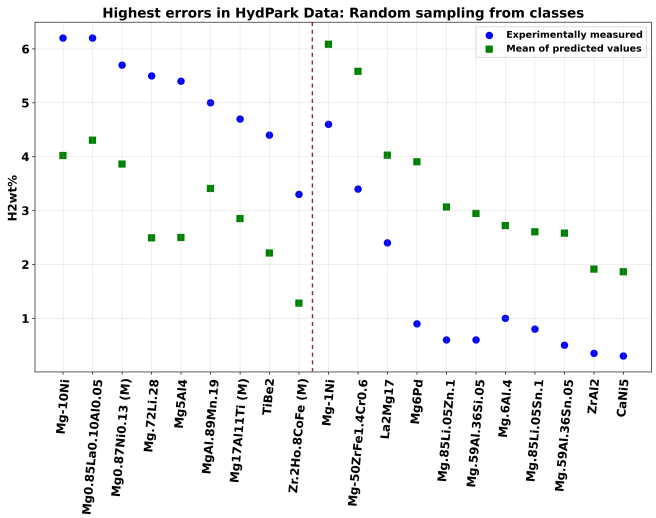
Figure 3: This figure illustrates the compositions that consistently performed poorly across the 1000 shuffle runs. The table within the figure displays the frequency of occurrence of each composition throughout the entire shuffle run. The compositions are distributed from left to right based on the ML model predictions. The left region represents compositions that are underpredicted by the model, while the right side corresponds to compositions where the model overpredicts.

such as TiBe2, CaNi5, and CaAl1.8B0.2(M), whose constituent elements Be, Ca, and B have limited representation in our database. Without enough representative compositions with these elements in our training data, the model has a higher probability of predicting with bigger error margins. Upon investigating the remaining compositions, we encountered 11 compositions whose original source material was inaccessible or not found during our search. The aim was to crosscheck their originally recorded values in research articles, since these compositions experimental values are not representative of similar compositions in our database. For example, compositions like Mg0.85Li0.05Zn0.1 at 583K was measured with an H2wt% of 0.6. 75% of compositions where Mg has a fraction of 0.80 and above in our database possess an H2wt% of 3 and above. Few other compositions with such irregular values are also identified, like Mg0.59Al0.36Si0.5, Mg0.59Al0.36Sn0.05, Mg0.85La0.1Al0.05, and Mg0.85Li0.05Sn0.1, all with a significant fraction of Mg and recorded at high temperatures. Before eliminating these compositions, we perform the next test to identify potential outliers by sampling from classes present in our databases. This is to confirm the presence of these potential outlier compositions through a different sampling method.

3.2. 20% class sampling test on HydPark dataset

In this test, we randomly select 20% of compositions from each class of the HydPark dataset and assign them as an out of the bag dataset. The rest of the database will be used for training the model. This test helps us to create a dataset representative of compositions from each class. We conduct this test 1000 times to identify compositions which are anomalous in the training data.

The compositions seen in Figure 4 identified through this test are identical to those from previous



Composition	Frequency	Temperature(K)	Material_class
Mg-10Ni	201	573	Mg
Mg0.85La0.10Al0.05	193	573	Mg
Mg0.87Ni0.13 (M)	203	598	Mg
Mg.72Li.28	184	623	Mg
Mg5Al4	192	599	MIC
MgAl.89Mn.19	207	623	MIC
Mg17Al11Ti (M)	212	577	Mg
TiBe2	173	295	AB2
Zr.2Ho.8CoFe (M)	206	300	AB2
Mg-1Ni	201	573	Mg
Mg-50ZrFe1.4Cr0.6	194	573	Mg
La2Mg17	192	373	MIC
Mg6Pd	194	433	MIC
Mg.85Li.05Zn.1	209	583	Mg
Mg.59Al.36Si.05	197	583	Mg
Mg.6Al.4	190	583	Mg
Mg.85Li.05Sn.1	184	583	Mg
Mg.59Al.36Sn.05	190	583	Mg
ZrAl2	186	293	AB2
CaNi5	207	295	AB5

Figure 4: This figure illustrates the compositions that were selected from various classes of the extracted compositions and which consistently performed poorly across the 1000 shuffle runs. The table within the figure displays the frequency of occurrence of each composition throughout the entire shuffle run. The compositions are distributed from left to right based on the ML model predictions. The left region represents compositions that are underpredicted by the model, while the right side corresponds to compositions where the model overpredicts.

tests as depicted in figure 3. This indicates that these same compositions are predicted with higher errors after sampling the test set in a different manner. Since we confirm the presence of these compositions through both these tests, we proceed to eliminate the outlier compositions. Three compositions identified with lack of representation in the database are removed. Further we proceed to remove the 11 compositions whose original source article was inaccessible or not found during our search. In total we remove 14 compositions which are listed in SI-Table 1. Six compositions whose H2wt% are accurately recorded as identified from their research articles are retained in the database. These six compositions Mg0.72Li0.28, Mg6Pd, Mg-50ZrFe1.4Cr0.6, Mg17Al11Ti, Mg.6Al.4, and Zr.2Ho.8CoFe which mainly belong to the Mg class provide a good deal of variability to our database. Mg6Pd and Mg.6Al.4 although comprising a high content of Mg posess a lower H2wt% value at higher temperatures and consequently the model over predicts their H2wt% values. The other four compositions posess a higher H2wt% value which is underpredicted by the model. These six compositions are also recorded with high temperature and therefore retaining them will maintain the distribution of high temperature training data.

Dataset	Number of composition	MAE	R2 score
HydPark database	842	0.31	0.78
Deleting outlier compositions	828	0.28	0.82

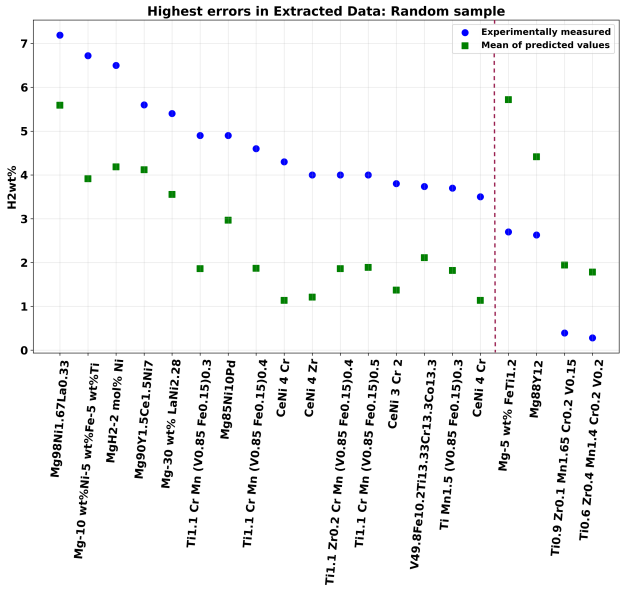
Table 3: Improvements in the error metrics on exclusion of outlier compositions

In table 3, we note the error metrics of the HYST model trained on the HydPark database and outlier removed HydPark database. We see improvements in the error metrics with MAE decreasing from 0.30 to 0.27 and R2 score increasing from 0.78 to 0.82. From this point, we proceed with further

analysis by excluding the identified outlier compositions. This is to refine the database, preparing it for subsequent testing and evaluation. Our objective is to assess the validity of newly extracted composition and determine their suitability for inclusion in the database update process. This approach ensures a more robust model for testing and validating new compositions.

3.3. 90/10 random data split on extracted compositions

For this test, we use the extracted compositions as a sample space and identify those compositions which performs with higher error predictions by the HYST model. From the sample space, we randomly select 10% of compositions and use the HYST model trained on the above finalized dataset for evaluation. We run this experiment 1000 times to acquire the error metrics for the individual composition.



Composition	Frequency	Temperature(K)	Material_class
Mg98Ni1.67La0.33	96	598	Mg
Mg-10 wt%Ni-5 wt%Fe-5 wt%Ti	90	573	Mg
MgH2-2 mol% Ni	98	423	Mg
Mg90Y1.5Ce1.5Ni7	81	573	Mg
Mg-30 wt% LaNi2.28	99	553	Mg
Ti1.1 Cr Mn (V0.85 Fe0.15)0.3	99	283	AB2
Mg85Ni10Pd	89	473	Mg
Ti1.1 Cr Mn (V0.85 Fe0.15)0.4	93	283	MIC
CeNi 4 Cr	95	566	AB5
CeNi 4 Zr	100	566	AB5
Ti1.1 Zr0.2 Cr Mn (V0.85 Fe0.15)0.4	105	283	MIC
Ti1.1 Cr Mn (V0.85 Fe0.15)0.5	104	283	MIC
CeNi 3 Cr 2	106	606	AB5
V49.8Fe10.2Ti13.33Cr13.3Co13.3	103	298	MEA
Ti Mn1.5 (V0.85 Fe0.15)0.3	120	283	MIC
CeNi 4 Cr	100	606	AB5
Mg-5 wt% FeTi1.2	91	673	Mg
Mg88Y12	94	373	Mg
Ti0.9 Zr0.1 Mn1.65 Cr0.2 V0.15	95	293	AB2
Ti0.6 Zr0.4 Mn1.4 Cr0.2 V0.2	100	296	AB2

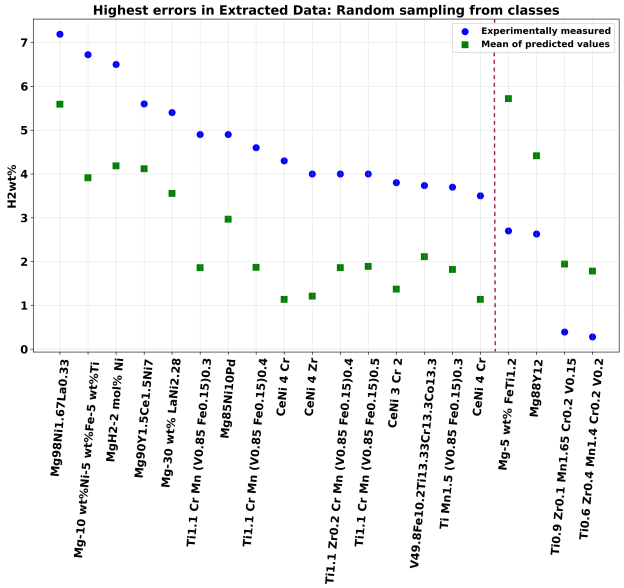
Figure 5: This figure illustrates the compositions that were selected from the extracted data and which consistently performed poorly across the 1000 shuffle runs. The table within the figure displays the frequency of occurrence of each composition throughout the entire shuffle run. The compositions are distributed from left to right based on the ML model predictions. The left region represents compositions that are underpredicted by the model, while the right side corresponds to compositions where the model overpredicts.

As observed from Figure 5, the four compositions of the Ti1.1CrMn and one compositions of TiMn1.5 family doped with V0.85Fe0.15 exhibit high H2wt% ranges of 3.5-4 compared to the average value of 1.5 wt% for these families in our dataset. When analyzing the extracted data, $Ti1.1CrMn(V0.85Fe0.15)_x$ where x= 0.2, 0.3, and 0.4 have H2wt% values of 1.4, 4.9, and 4.6 all measured at 283K. Since HYST model utilises the compositions normalised fractions to calculate its features, a slight variation of 0.02 for V and 0.04 for Fe in the composition fraction only results in a similar feature set. The complete feature set of these three compositions are depicted in SI-figure 1. Since our model does not take into consideration the structural and experimental features of a composition, the changes that may have been observed with these slight variations of V0.85Fe0.15 is not provided to the model. Thus we

consider the Ti1.1CrMn doped with (V0.85Fe0.15) families listed with high errors as potential outliers to models learning. We also note the CeNi5 compositions doped with Cr and Zr with a high H2wt% of 3.5 and above. Other CeNi5 and its doped compositions in our training data are limited to compositions with low temperature ranges of 293-323K. The CeNi5 doped with Cr and Zr flagged as potential outliers are all measured at very high temperature range of 566-606K. Since these compositions provide the information of CeNi5 and its dopants at high temperatures, we retain them in our database. We further test the extracted compositions by sampling them randomly from the alloy classes in our database.

3.4. 20% class sampling test on extracted compositions

For this test, we use the extracted compositions as a sample space and randomly select 20% from each class and create a hold out test set. This test set is then used to benchmark the HYST model.



Composition	Frequency	Temperature(K)	Material_class
Mg98Ni1.67La0.33	172	598	Mg
Mg-10 wt%Ni-5 wt%Fe-5 wt%Ti	170	573	Mg
MgH2-2 mol% Ni	177	423	Mg
Mg90Y1.5Ce1.5Ni7	177	573	Mg
Mg-30 wt% LaNi2.28	172	553	Mg
Ti1.1 Cr Mn (V0.85 Fe0.15)0.3	199	283	AB2
Mg85Ni10Pd	162	473	Mg
Ti1.1 Cr Mn (V0.85 Fe0.15)0.4	199	283	MIC
CeNi 4 Cr	101	566	AB5
CeNi 4 Zr	120	566	AB5
Ti1.1 Zr0.2 Cr Mn (V0.85 Fe0.15)0.4	216	283	MIC
Ti1.1 Cr Mn (V0.85 Fe0.15)0.5	214	283	MIC
CeNi 3 Cr 2	141	606	AB5
V49.8Fe10.2Ti13.33Cr13.3Co13.3	210	298	MEA
Ti Mn1.5 (V0.85 Fe0.15)0.3	185	283	MIC
CeNi 4 Cr	107	606	AB5
Mg-5 wt% FeTi1.2	193	673	Mg
Mg88Y12	196	373	Mg
Ti0.9 Zr0.1 Mn1.65 Cr0.2 V0.15	188	293	AB2
Ti0.6 Zr0.4 Mn1.4 Cr0.2 V0.2	182	296	AB2

Figure 6: This figure illustrates the compositions that were selected from various classes of the HydPark database and which consistently performed poorly across the 1000 shuffle runs. The table within the figure displays the frequency of occurrence of each composition throughout the entire shuffle run. The compositions are distributed from left to right based on the ML model predictions. The left region represents compositions that are underpredicted by the model, while the right side corresponds to compositions where the model overpredicts.

The sampling of compositions from different classes of alloys from the extracted database has also resulted in the exact same compositions that were analyzed during the previous experiment. The error metrics are evaluated in higher frequency during these runs but the error metrics remain relatively same for these compositions. To proceed with the further analysis, we eliminate the Ti1.1CrMn and TiMn1.5 doped with (V0.85Fe0.15). The list of compositions eliminated through these tests are listed in SI-table 2. The other compositions belonging to CeNi4 doped with Cr and Zr, Mg based compositions, AB2, and MEA are retained. Although these compositions gave high error margins, we choose to retain them as these compositions provide variability to our data. To proceed with further analysis,

Dataset	Number of composition	MAE	R2 score
HyStor database	1299	0.31	0.78
Deleting outlier	1280	0.28	0.82

Table 4: Error metrics on exclusion of outlier compositions

we remove the outlier compositions from the extracted data and finalise the HyStor database.

As seen from table 4, the error metrics of the HYST model trained on HyStor has improved on exclusion of all the potential outliers identified from the HydPark dataset and extracted compositions. The final overall error metrics of HYST on HyStor database with 1280 compositions is MAE of 0.28 and R2 score of 0.82. To further benchmark the performance of the HyStor trained model, we sample individual classes in the dataset and compare the performance of HYST trained on HydPark and HyStor database.

4. Alloy class error benchmarks using HydPark and HyStor

In this section, we utilize the HYST model trained on HydPark and HyStor database and use it to benchmark individual classes that are present in our database. To create the test set for a class, we utilize the HyStor database and sample 20% of the data, which we keep aside as an out-of-bag dataset. Subsequently, we train the HYST model by excluding these test set compositions from both the HydPark and HyStor databases. Finally, we calculate the error metrics for each class over 100 iterations based on the trained model.

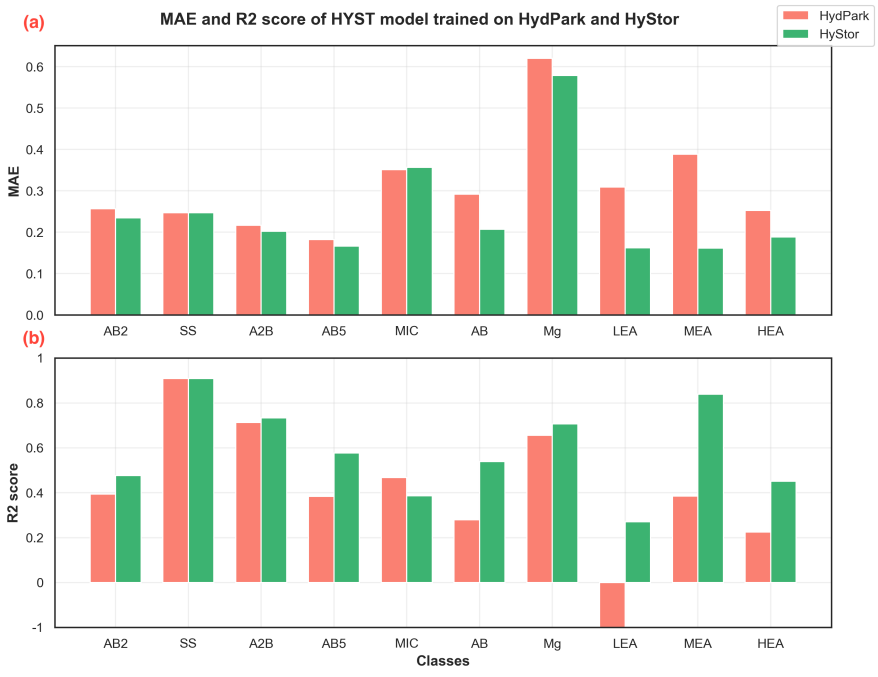


Figure 7: (a) represents the MAE and (b) represents the R2 score of individual classes benchmarked using the HYST model trained on HydPark and HyStor database. In figure 7(b), the y-axis is scaled down between -1 to 0 to depict the negative R2 score of -1.0 achieved by the HydPark model on prediction of LEA class.

Figure 7 (a) depicts the MAE of each class when predicted with HydPark trained and HyStor trained model. The MAE for classes like AB2, A2B, AB5, AB, Mg, LEA, MEA, and HEA have reduced on utilizing the model trained on HyStor. This signifies that the addition of compositions to individual classes has enhanced the models performance for these classes. In figure 7 (b), we see that for these same classes, the R2 score has also increased which also depicts the improved fitting of the model on these test classes. Specifically the MAE and R2 score of LEA, MEA, and HEA class of alloys have seen a significant improvement using the HyStor model. This is attributed to the fact that the compositions of these alloy classes were missing in the HydPark database. Alloy classes of SS and MIC have shown a similar performance in the error metrics using models trained using both HydPark and HyStor database. The SS class has only seen an increase of 5 compositions and this can be attributed to the similar error metrics by the model. The MIC classes although added with 26 compositions, exhibit similar error metrics.

5. Comparative analysis of HydPark and HyStor database

In this section, we conduct a comparative analysis of the compositions, classes and elements present in the HydPark and the HyStor database.

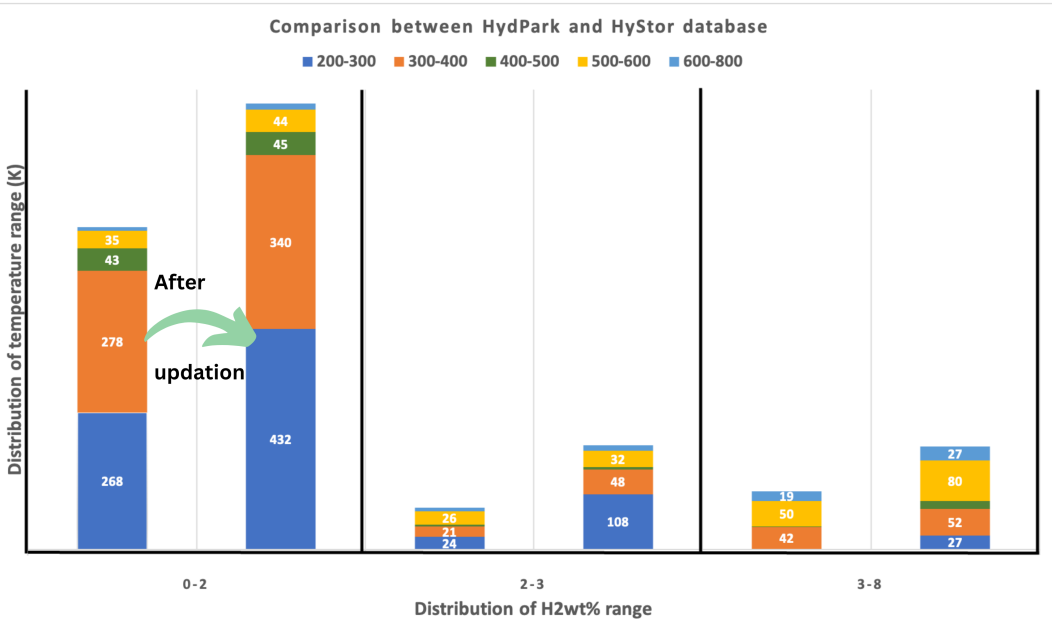


Figure 8: Distribution of compositions in both the Hydpark and HyStor databases. Compositions are divided according to their H2wt% distribution. Each distribution is further divided through their absorption temperature ranges. The initial bar in each window represents the distribution of compositions within the HydPark database whereas the second bar represents the HyStor database representation.

Figure 8 represents the distribution of compositions in both the HydPark and HyStor databases in different H2wt% ranges. Majority of the compositions fall within the lower H2wt% range of 0-2 wt%. A large increase in compositions are observed across the 200-300K range. This is because the updates

has included a significant number of compositions from the families of TiCrMn, TiFe, TiMn, and HEA whose absorption temperature is near room temperature. The majority of these compositions are recorded with 0-2 wt%, and they include compositions with lanthanides such as La, Pr, Nd, Gd, Tb, Dy, Ho, Er, as well as other heavy elements like Pd and U. The combined temperature range of 400-800K exhibits the highest number of compositions with H2wt% of 4 and above, predominantly attributed to Mg-based compositions. The number of compositions with a higher H2wt% of 3 and above with a lower temperature range of 200-300K was negligible in the HydPark database. The addition of new compositions has increased the representation of higher H2wt% compositions in the lower temperature region of 200-300K.

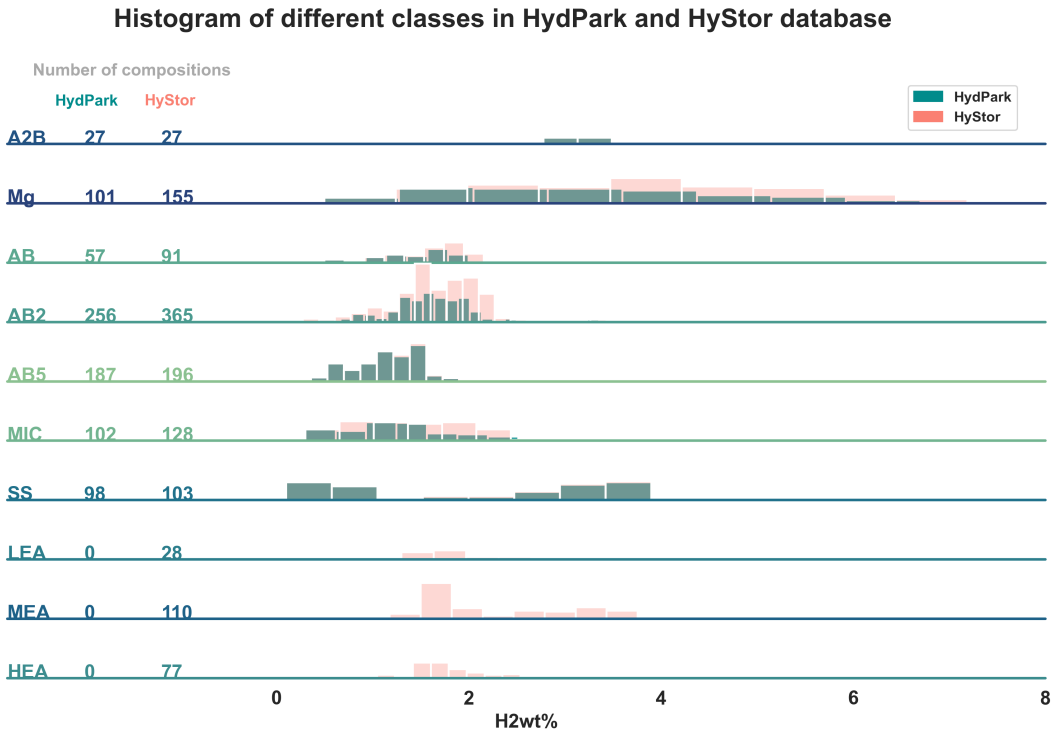
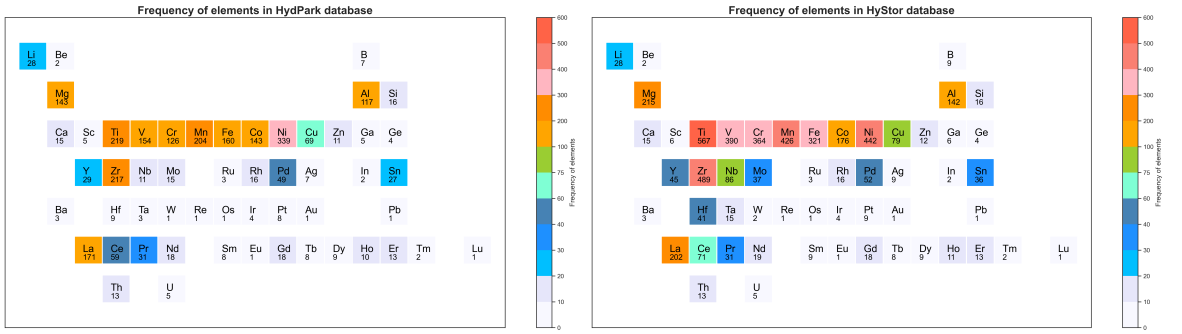


Figure 9: This figure represents the histogram distribution of H2wt% of various classes in the HydPark and HyStor database.

In figure 9, we can see the histogram distribution of H2wt% across different regions of the ten material classes. A corresponding class and its number of compositions in a particular dataset are also labeled in figure 9. There is a significant addition to the classes of AB, AB2, Mg and LEA, MEA, and HEA are newly added classes to the database. Each of the A2B, AB, AB2, and AB5 classes represents compositions where the A part belongs to a hydriding element, and the B part belongs to a non-hydriding element. There is an observed trend of decreasing H2wt% in these classes with an increase in the concentration of the B part of the composition. The A2B class comprises compositions whose hydriding part belongs to elements like Mg, Ti, and Hf. The higher H2wt% ranges are occupied by the Mg based compositions while the lower ranges of this class belong to Hf

and Ti based compositions. The AB class in the database mainly comprises compositions where Ti serves as the hydriding part, while later transition metals such as Fe, Mn, Co, Ni, and Cr occupy the non-hydriding part. These compositions exhibit an H2wt% range between 0.5 and 2, with the majority operating in the temperature range of 300-400K. The addition comprises mainly of TiFe and its various dopants to the database. These compositions also has increased representation in the higher H2wt% region of this alloy class. The AB2 class belongs to compositions whose A part is occupied by Ti, Zr or a mix of both Ti and Zr with varying fractions. The B part is occupied by elements like V, Mn, Fe, Ni, Cr, and Si. Few notable compositions in this class have H2wt% of 2.6 - 3.6 and these contain V in their B part. These compositions also operate at around 300-400K temperatures. The number of compositions updated in this class of alloy is the highest. Majority of these added compositions belong to TiCrMn family of alloys. This addition has also increased representation in the higher regions of this class thereby improving the overall distribution in this alloy class. The AB5 class comprises of compositions with lanthanides like La, Ce, Sm, Er, Nd, and Tb occupying the A part and later transition metals occupying the B part. They operate at low temperatures of 298 -310K, and their H2wt% range from 0.3-2 wt%. The Mg-based compositions are distributed across the entire H2wt% range, spanning from 0.5 to 7 wt%. Significantly, Mg-based compositions stand out for having H2wt% greater than 4 when compared to other compositions. Nearly all of these compositions are also associated with elevated temperatures, specifically 550K and above. The addition of Mg based compositions are mainly seen in the higher H2wt% regions. MIC or miscellaneous inter-metallic compounds contain different combinations of A and B part with elements like Mg, V, Ti, La, and Zr occupying the hydriding part. These composition basically follow different range of temperatures and have also varied H2wt% range of 0.3-6 wt%. SS class has two major regions within their class. The region in the range of 0.1-2 corresponds to compositions consisting of Pd as the hydriding part whereas the region in the range of 3-4 H2wt% range belongs to compositions where V acts as the hydriding element. LEA class of compositions exhibit an H2wt% range between 0.5 and 2, with the majority operating in the temperature range of 298-373K. These compositions mainly comprises of three elements in an equiproportion fraction and is seen with elements like Ti, V, Cr, Fe, Zr, and Mn. Few compositions also comprise of Mg in their mixture. MEA compositions in HyStor database exhibit a range of H2wt% value of 1.5-2 and 2-3.4 wt%. These compositions operate at lower ranges of temperature of 273K to 373K. HEA compositions in the database are highly occupied with compositions whose H2wt% lie in the range of 1-2 wt%. The elements used in the HEA predominantly belong to transition metals. Few compositions also contain Mg and Al.

Figure 10 (a) represents the frequency of the different elements present in the HydPark database whereas figure 10(b) represents the element distribution in the HyStor database. HydPark database reflects the elements which were studied before the year 2002 for the purpose of hydrogen storage. The



(a) (b)

Figure 10: (a) and (b) shows the element representation before and after addition of new points. (a) represents the element representation in the initial dataset of 856 compositions all selected from HydPark database after preprocessing. (b) represents the HyStor database which comprises 1280 compositions element representation.

new and updated HyStor database has included compositions post 2002 and various patent documents and signifies the usage of elements which have been on focus in recent years. According to Figures 10 (a) and (b), a few 3d and 4d transition series elements have experienced a significant increase in numbers following the addition of new compositions to the database. In particular, elements like Ti, V, Cr, Mn, Fe, Ni, Y, Zr, Hf, Nb, and Mo have witnessed an almost twofold increase in their numbers. Literature studies reveal that the early transition metals occupy the hydriding or A part of the alloy, while the later transition metals serve as the non-hydriding or B part of the alloy. Ti, V, and Zr are well-known for their hydriding capabilities, making them subjects of significant interest in metal hydride research, as evidenced by both the original Hydpark and updated databases. Transition metals are seen to form metal hydride alloys with room temperature operability, thus making them a practical choice for the purpose of hydrogen storage. Additionally, elements like Nb, Mo, and Hf, which were infrequently used before 2002, have also shown an increase in utilization, according to the updated database. However, metals from other transition series, such as Cu, Pd, Sn, and Pt, exhibit relatively limited increase in numbers, suggesting that they have not been extensively explored or utilized in recent research, as indicated by the updated database. Furthermore, there is an observed increase in the usage of Mg based compositions on the updated database. Conversely, other hydriding alkali and alkaline elements like Li and Ca show little to no increase. Li and Ca metals possess higher hydriding capacities but forms irreversible hydrides making them impractical for recharging processes. This indicates the lack of increase in these elements even after the updates to the database. Overall the number of elements present in both the databases is 54. The representation in HyStor database has increased the frequency of 18 elements significantly and has specifically improved the representation of transition metals and Mg.

6. Conclusions

In this work, we present the HyStor database, comprising 1280 data-points of metal alloys with their hydrogen storage capacity as a function of absorption temperature. To address the lack of updates in the existing HydPark database, we sourced data from recent research articles and patent documents, resulting in the extraction of 468 compositions with their H2wt% as a function of absorption temperature. Through rigorous data preprocessing and quality checks, we identified and resolved issues such as duplication within the databases, erroneous entries, conflicting values in the database which resulted from varying pressure and different synthesis routes as identified from the original research article. Additionally, each composition underwent rigorous anomaly and outlier detection tests to optimize the database for the ML model HYST, used for prediction of H2wt%. Utilizing this fully optimized dataset, we were able to reduce the Mean Absolute Error (MAE) from 0.32 to 0.28 and increase the R2 score from 0.78 to 0.82 for the HYST model. We also tested the HYST model's performance on individual alloy classes and observed improvements in the error metrics for 8 out of the 10 classes in comparison to the existing HYST model trained on HydPark data. The rest 2 alloy classes are seen with similar error metrics to that of existing model.

Furthermore, we discussed various updates introduced in our work specific to different regions of H2wt% and absorption temperatures, different alloy classes, and element representation. We also introduced new alloy classes such as LEA, MEA, and HEA to HyStor. The addition of compositions from patent documents resulted in addition of practical and commercially tested alloy compositions to our database.

CRediT authorship contribution statement

Nikhil Wilson: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Review & Editing, & Visualization. Ashwini Verma: Software, Formal analysis, Methodology, Investigation, Data Curation, Writing - Original Draft. Piyush Ranjan Maharana: Formal analysis, Investigation, Writing - Review & Editing and Visualization. Ameeya Bhusan Sahoo: Data Curation & Formal analysis. Kavita Joshi: Conceptualization, Writing - Review & Editing, Visualization, Supervision, Project administration, and Funding acquisition.

Conflicts of interest

"There are no conflicts to declare".

Acknowledgements

Authors acknowledge CSIR for financial support through grant HCP-44-05 (HEART). AV thanks DST for INSPIRE fellowship.

References

[1] R. Batra, L. Song, R. Ramprasad, Emerging materials intelligence ecosystems propelled by machine learning, Nature Reviews Materials 6 (8) (2021) 655–678.

[2] L. Zhu, J. Zhou, Z. Sun, Materials data toward machine learning: advances and challenges, The Journal of Physical Chemistry Letters 13 (18) (2022) 3965–3977.

[3] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al., Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL materials 1 (1) (2013).

[4] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, et al., Open catalyst 2020 (oc20) dataset and community challenges, Acs Catalysis 11 (10) (2021) 6059–6072.

[5] M. Scheidgen, L. Himanen, A. N. Ladines, D. Sikter, M. Nakhaee, ÃdÃ¡m Fekete, T. Chang, A. Golparvar, J. A. MÃ¡rquez, S. Brockhauser, S. BrÃ¼ckner, L. M. Ghiringhelli, F. Dietrich, D. Lehmberg, T. Denell, A. Albino, H. NÃ¤sstrÃ¶m, S. Shabih, F. Dobener, M. KÃ¼hbach, R. Mozumder, J. F. Rudzinski, N. Daelman, J. M. Pizarro, M. Kuban, C. Salazar, P. OndraÄka, H.-J. Bungartz, C. Draxl, Nomad: A distributed web-based platform for managing materials science research data, Journal of Open Source Software 8 (90) (2023) 5388. doi:10.21105/joss.05388.
URL https://doi.org/10.21105/joss.05388

[6] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd), Jom 65 (2013) 1501–1509.

[7] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, et al., Materials cloud, a platform for open computational science, Scientific data 7 (1) (2020) 299.

[8] S. S. Borysov, R. M. Geilhufe, A. V. Balatsky, Organic materials database: An open-access online database for data mining, PloS one 12 (2) (2017) e0171501.

[9] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, Schnet–a deep learning architecture for molecules and materials, The Journal of Chemical Physics 148 (24) (2018).

[10] K. Meng, C. Huang, Y. Wang, Y. Zhang, S. Li, Z. Fang, H. Wang, S. Wei, S. Sun, Bnm-cdgnn: Batch normalization multilayer perceptron crystal distance graph neural network for excellent-performance crystal property prediction, Journal of Chemical Information and Modeling 63 (19) (2023) 6043–6052.

[11] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, G. Ceder, Predicting crystal structures with data mining of quantum calculations, Physical review letters 91 (13) (2003) 135503.

[12] P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, Machine learning for computational heterogeneous catalysis, ChemCatChem 11 (16) (2019) 3581–3601.

[13] G. R. Schleder, A. C. M. Padilha, A. Reily Rocha, G. M. Dalpian, A. Fazzio, Ab initio simulations and materials chemistry in the age of big data, Journal of chemical information and modeling 60 (2) (2019) 452–459.

[14] C. Chen, S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nature Computational Science 2 (11) (2022) 718–728.

[15] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, G. Csányi, Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, Advances in Neural Information Processing Systems 35 (2022) 11423–11436.

[16] R. Modee, S. Agarwal, A. Verma, K. Joshi, U. D. Priyakumar, Dart: deep learning enabled topological interaction model for energy prediction of metal clusters and its application in identifying unique low energy isomers, Physical Chemistry Chemical Physics 23 (38) (2021) 21995–22003.

[17] R. Modee, A. Verma, K. Joshi, U. D. Priyakumar, Megen-generation of gallium metal clusters using reinforcement learning, Machine Learning: Science and Technology 4 (2) (2023) 025032.

[18] K. Lejaeghere, V. Van Speybroeck, G. Van Oost, S. Cottenier, Error estimates for solid-state density-functional theory predictions: an overview by means of the ground-state elemental crystals, Critical reviews in solid state and materials sciences 39 (1) (2014) 1–24.

[19] M. Hellenbrandt, The inorganic crystal structure database (icsd)—present and future, Crystallography Reviews 10 (1) (2004) 17–22.

[20] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, A. Le Bail, Crystallography open database (cod): an open-access collection of crystal structures and platform for world-wide collaboration, Nucleic Acids Research 40 (D1) (2012) D420–D427. arXiv:https://nar.oxfordjournals.org/content/40/D1/D420.

full.pdf+html, doi:10.1093/nar/gkr900.
URL https://academic.oup.com/nar/article/40/D1/D420/2903497

[21] K. R. Talley, R. White, N. Wunder, M. Eash, M. Schwarting, D. Evenson, J. D. Perkins, W. Tumas, K. Munch, C. Phillips, et al., Research data infrastructure for high-throughput experimental materials science, Patterns 2 (12) (2021).

[22] M. C. Swain, J. M. Cole, Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature, Journal of chemical information and modeling 56 (10) (2016) 1894–1904.

[23] S. Huang, J. M. Cole, Batterybert: A pretrained language model for battery database enhancement, Journal of chemical information and modeling 62 (24) (2022) 6365–6377.

[24] S. Suwarno, G. Dicky, A. Suyuthi, M. Effendi, W. Witantyo, L. Noerochim, M. Ismail, Machine learning analysis of alloying element effects on hydrogen storage properties of ab2 metal hydrides, International Journal of Hydrogen Energy 47 (23) (2022) 11938–11947.

[25] A. Verma, N. Wilson, K. Joshi, Solid state hydrogen storage: Decoding the path through machine learning, International Journal of Hydrogen Energy 50 (2024) 1518–1528.

[26] M. Witman, S. Ling, D. M. Grant, G. S. Walker, S. Agarwal, V. Stavila, M. D. Allendorf, Extracting an empirical intermetallic hydride design principle from limited data via interpretable machine learning, The Journal of Physical Chemistry Letters 11 (1) (2019) 40–47.

[27] US DOE hydrogen storage database, data obtained from the Hydrogen Materials Advanced Research Consortium (HyMARC) Data Hub at datahub.hymarc.org.

[28] J. R. Hattrick-Simpers, K. Choudhary, C. Corgnale, A simple constrained machine learning model for predicting high-pressure-hydrogen-compressor materials, Molecular Systems Design & Engineering 3 (3) (2018) 509–517.

Supporting Information HyStor: An Experimental Database of Hydrogen Storage Properties for Various Metal Alloy Classes

Nikhil Wilson, Ashwini Verma, Piyush Ranjan Maharana, Ameeya Bhusan Sahoo, and Kavita Joshi*

Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pashan, Pune 411008, India and Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India

(Dated: April 20, 2024)

* k.joshi@ncl.res.in

I. COMPOSITIONS ELIMINATED THROUGH OUTLIER ANALYSIS

Through the tests performed to identify compositions that were potential outliers to the model, these are the compositions eliminated from the HydPark dataset.

Sl.no	Composition	Case	H2wt%	Temperature(K)
1	TiBe2	Limited representation of Be element in database	4.4	295
2	CaNi5	Limited representation of Ca element in database	0.3	295
3	CaAl1.8B.2 (M)	Limited representation of B element in database	0.7	313
4	Mg5Al4	Research article not found	5.4	599
5	Mg.59Al.36Si.05	Research article not found	0.6	583
6	Mg.85Li.05Zn.1	Research article not found	0.6	583
7	Mg-10Ni	Research article not found	6.2	573
8	Mg.59Al.36Sn.05	Research article not found	0.5	583
9	Mg0.85La0.10Al0.05	Research article not found	6.2	573
10	Mg.85Li.05Sn.1	Research article not found	0.8	583
11	MgAl.89Mn.19	Research article not found	5	623
12	ZrAl2	Research article not found	0.35	293
13	La2Mg17	Research article not found	2.4	373
14	Mg0.87Ni0.13 (M)	Research article not found	5.7	598

TABLE I. This table represents the compositions that are eliminated from the HydPark database.

As seen from Table I, the first three compositions eliminated from the HydPark database comprise compositions whose constituent elements, such as Be, Ca, and B, are lacking in the database. This can lead to higher error margins when predicting their compositions. From the rest of the compositions which were found with higher errors, we were unable to cross-check the hydrogen storage properties of 11 compositions listed in Table I from their original research articles due to the inaccessibility of these articles.

The following compositions as seen in Table II are removed on analysis of the outlier extracted compositions. According to Table II, the compositions that are eliminated comprise of Ti1.1CrMn family and TiMn1.5 doped with V0.85 Fe0.15 at varying fractions. As the model is trained on features generated based on compositional fractions , a small variation

Sl.no	Composition	Case	H2wt%	Temperature(K)
1	Ti1.1 Cr Mn (V0.85 Fe0.15)0.3	Outlier data point with high errors	4.9	283
2	Ti1.1 Cr Mn (V0.85 Fe0.15)0.4	Outlier data point with high errors	4.6	283
3	Ti1.1 Cr Mn (V0.85 Fe0.15)0.5	Outlier data point with high errors	4	283
4	Ti1.1 Zr0.2 Cr Mn (V0.85 Fe0.15)0.4	Outlier data point with high errors	4	283
5	Ti Mn1.5 (V0.85 Fe0.15)0.3	Outlier data point with high errors	3.7	283

TABLE II. This table represents the compositions that are eliminated from the extracted compositions

in the fractions of V and Fe between the compositions results in the generation of similar feature sets. A comparison of features generated for three Ti1.1CrMn family of compositions is shown in figure 1

Sl.no	Composition	H2wt%	Temperature(K)
1	Ti1.1 Cr Mn (V0.85 Fe0.15)0.2	1.4	283
2	Ti1.1 Cr Mn (V0.85 Fe0.15)0.3	4.9	283
3	Ti1.1 Cr Mn (V0.85 Fe0.15)0.4	4.6	283

TABLE III. The following compositions are shown as a representative case of similar features generated for Ti1.1CrMn family doped with V0.85 Fe0.15.

As observed in Figure 1, the features generated for these three compositions shown in III are similar, as they depend on the compositional fractions of the constituent elements. Since our model does not incorporate structural and experimental features into its training, it cannot capture changes resulting from variations in V0.85 Fe0.15 . Therefore, we eliminate those compositions with V0.85 Fe0.15 as a dopant and predicted with high error margins as seen in Table II, as they would affect the model's predictions for other compositions in the Ti1.1CrMn family.

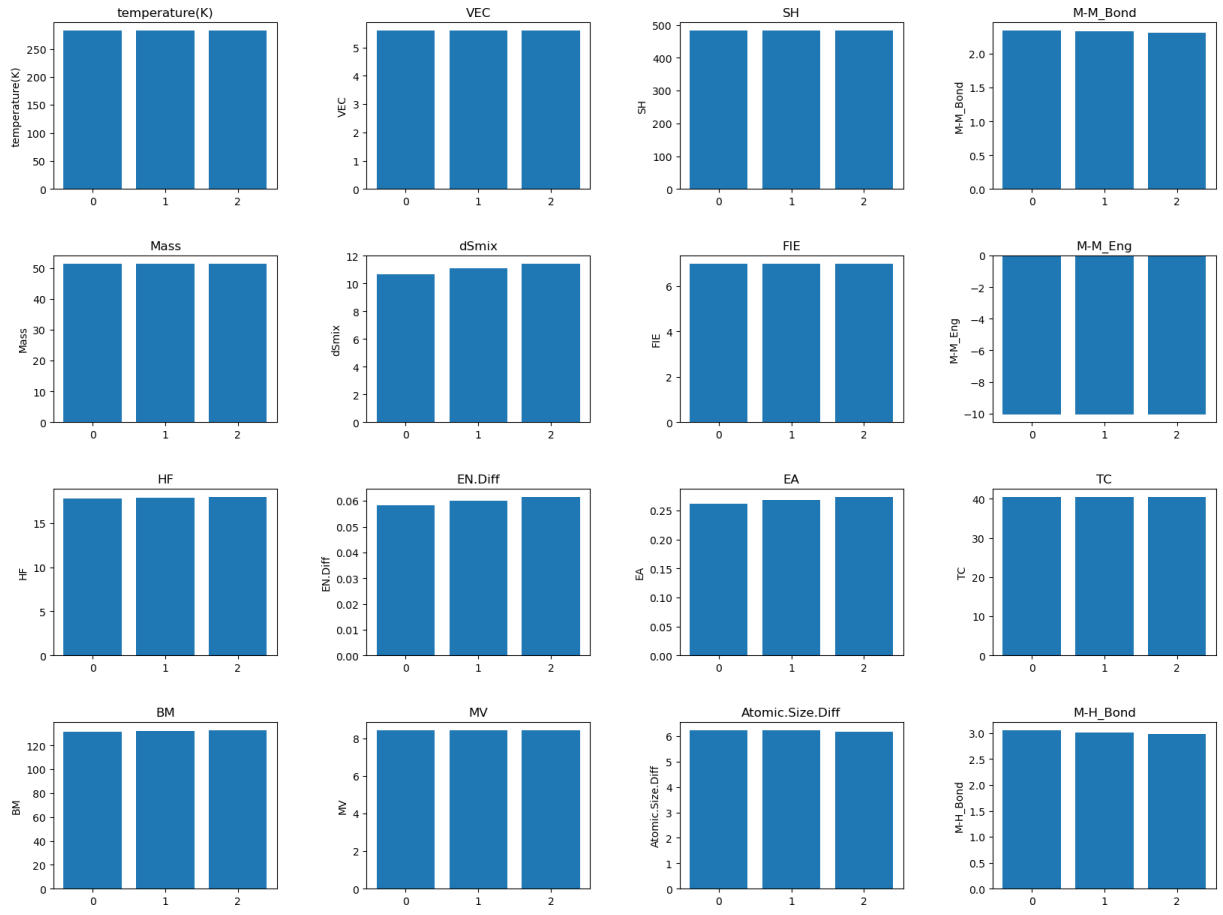


FIG. 1. Represents the features generated for the three compositions as mentioned in Table III. The x axis labels of each figure where represents the three compositions where 0 represents Ti1.1 Cr Mn (V0.85 Fe0.15)0.2, 1 represents Ti1.1 Cr Mn (V0.85 Fe0.15)0.3, and 2 represents Ti1.1 Cr Mn (V0.85 Fe0.15)0.4