Process Book: Visualizing Error in Materials Informatics

Steven K. Kauwe · Yucheng Yang

Abstract We present a website for visualizing the error associated with machine learning predictions in materials science.

Keywords Machine Learning · Band Gap · Density Functional Theory · Ensemble Learning · Inorganic Solids

1 Project Proposal

1.1 Background and Motivation

This project comes from the research of Steven Kauwe who investigates the use of materials informatics for predicting material properties as a Ph.D student in the Department of Materials Science and Engineering. The current state of research in that area has focused on showing proof of concept machine learning using various data types. Because materials science is inherently chemistry based, researchers have found it is useful to represent chemical systems based on their constituent elements. This is the foundation for what we will refer to as the composition-based feature vector (CBFV). The CBFV is a common method of vectorizing chemistry space, allowing a representation of physical composition to be parsed into traditional machine learning frameworks. Because of the relatively new interest in this area, some fundamental ideas have yet to be implemented in an open and accessible way. In particular, to our knowledge there exists no central tool for researchers to evaluate the effect constituent elements have in prediction error. This project aims to provide a proof of concept tool for just such composition based error exploration.

Steven Kauwe Yucheng Yang
Department of Computer Science, University of Utah, Salt Lake City
E-mail: u0742731@utah.edu, u0878896@utah.edu

1.2 Project Objective

This project looks to answer the question whetherr certain chemistries are more prone to error in machine learnign models that are build from compositional information. In particular, we would like to know if this error is caused by the chemistry itself of by the lack or representation in the dataset. Using a variety of visual techinqes, we hope to give the user the ability to answer the following questions:

1. What elements have the most error associated with them?
2. Are non-real compositions less accurate that real compositions for this data?
3. How does the relationship between elemental properties affect the emergent properties of the molecule
4. Are there patterns in error relative to feature values?
5. Do compounds with similar elements cluster close to each other in feature space, and predictions?

1.3 Data Acquisition

Data is collected from the computational materials property repositories: Automatic - Flow for Materials Discovery (AFLOW), and The Materials Project (MP) who do high throughput density functional calculations (DFT) to compute material properties as a part of the Materials Genome Initiative. The AFLOW data was retrieved directly from their search tool after selecting the required electronic property, Band Gap aflowlib.org. The

MP data was obtained via a third party, Citrine Informatics, at citrination.com: data set #150675. The data is then vectorized via composition and used in a machine learning scheme that saves the cross-validation predictions and trains a model for future predictions. The error associated with the cross-validation results comprise the interactive aspect of the project.

1.4 Data Processing

The data may have duplicate entries and compositions that do not represent real compounds. The duplicates values can be solved by averaging all reported values. The non-real entries will not be rejected from the data set, but will instead be marked to let researchers know how this effects model error. To do this, we will need to cross-checked all AFLOW and MP data against a crystal structure database. This will be done with the Pearson Crystal Database (PCD).

We expect to derive a predicted value of band gap from our data. Error metrics for all cross-validation predictions will also be obtained during model creation. Before use online, we will need to classify the data entries by the constituent elements and whether we could confirm if a real structure is reported for the composition. The compositions will also be vectorized in a composition-based feature vector (CBFV) which is used in the machine learning. This will start as a vector with approximately 250 elements, however principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) will reduce this down to 2-5 dimensions for human interaction.

Data processing will be implemented in python 3 using the pandas and sk-learn libraries. All data will be processed offline then compactly represented as CSVs or JSONs.

2 Visualization Design

This data will mostly be displayed using scatter plots, distribution plots, trees, and a interactive periodic table. The markers will typically represent individual chemical compounds.

The user should be able to interact with a periodic table to select compounds of interest. These, in turn, will be highlighted in the other views to allow the user to quickly and intuitively explore the interaction of composition with error. The user could interact with a distribution plot to filter for values they would like to consider. It could also be possible for the user to put in their own composition and generate a prediction.

There will be three pages: overview, elements, and search. Each page will independently operate, providing a specific use for the user. The overview will show overall trends in the data and allow for high level exploration. The elements is designed to let users to see specific information associated with the individual elements and the formulae they are a part of. The search page allows users to see previous searches and generate new predictions.

This three page design is made to be captivating while simple to interact with. It gives the user enough choices to spend significant time on the webpage, but also limits the available information to prevent a feeling of being overwhelmed. The representation of data is standard for most publications in the field and is appropriate.

2.0.1 Must-Have Features

Completion of the overview page would constitute success. This is the main aspect of the project and provides the user with the most functional information.

2.0.2 Optional Features

The element and search pages aim to present our more detail in the data set. These are less complicated and serve a secondary purpose of which is helpful for evaluating individual formula, but not critical for gauging the error in general.

3 Current Data Structure

3.1 Periodic Data Table

The file ptable.csv (see Figure 1) is the data that we load to create the periodic table view (ptable view). Each row of this file describes an element. The attributes in this table contains symbol, row, column, data, name, count, average residual, average predicted and average actual. Row and column attributes provide the information to allow us to format the periodic table. Symbol and name attributes provide the information that we use for generating classes for each element. These allow us to easily communicate across view. The other attributes are made available for creating heat maps, their values are used to generate the color scale, which dictates the color each element block.

3.2 Predictions Data Table

The file experimental_predictions.csv (see Figure 2) is the data that we load to create the element's residual

	A	B	C	D	E	F	G	H	I
1	symbol	row	column	data	name	count	average_residual	average_predicted	average_actual
2	S	3	16	1	Sulfur	1582	-0.018450124	2.029286807	2.047736932
3	C	2	14	1	Carbon	711	-0.017141874	2.121395889	2.138537763
4	Se	4	16	1	Selenium	687	-0.022107451	2.071090942	2.093198393
5	O	2	16	1	Oxygen	544	-0.036612011	2.227308763	2.263920773
6	B	2	13	1	Boron	519	-0.019434629	2.167436367	2.186870997
7	P	3	15	1	Phosphorus	473	-0.0488924	1.950587856	1.999480256
8	Te	5	16	1	Tellurium	401	-0.04280867	1.850754382	1.893563052
9	I	5	17	1	Iodine	397	0.000306056	2.027968025	2.027661969
10	Ga	4	13	1	Gallium	313	-0.006491058	2.07397963	2.080470688
11	In	5	13	1	Indium	289	0.023309962	2.047781665	2.024471703
12	Sb	5	15	1	Antimony	266	-0.01736794	1.844183754	1.861551694
13	N	2	15	1	Nitrogen	238	-0.045934682	2.29876152	2.344696201
14	Ge	4	14	1	Germanium	218	0.065067597	2.070349413	2.005281816
15	Cu	4	11	1	Copper	217	-0.000401033	2.169833656	2.170234688
16	As	4	15	1	Arsenic	204	-0.093110865	1.868990186	1.96210105
17	Cd	5	12	1	Cadmium	191	-0.064922956	1.955095054	2.02001801
18	Ba	6	2	1	Barium	190	-0.021764965	2.373987895	2.39575286
19	Bi	6	15	1	Bismuth	190	-0.036584624	1.930019388	1.966604012
20	Sn	5	14	1	Tin	184	-0.077443006	2.088748831	2.166191837
21	K	4	1	1	Potassium	177	0.050816866	2.356437768	2.305620901
22	Pb	6	14	1	Lead	175	-0.039275034	1.793590866	1.832865901
23	H	1	1	1	Hydrogen	173	-0.075598245	2.303625155	2.3792234
24	Cs	6	1	1	Cesium	165	0.007199795	2.235948793	2.228748999
25	Zn	4	12	1	Zinc	158	-0.02406119	1.970638491	1.99469968
26	Ag	5	11	1	Silver	124	-0.067895739	2.052740591	2.120636331

Fig. 1 Periodic Data Table

view and actual vs prediction view. Each row of this file describes a formula. The attributes in this table contains formula, actual, predicted, elements and residual. We classify tables of each elements based on the elements attributes. Actual and predicted attributes provide the information to allow us to format actual vs prediction view. Residual attribute provides the information to allow us to format residual view.

	A	B	C	D	E
1	formula	actual	predicted	elements	residual
2	DyIn3S6	2.43	1.615021285	Dy In S	0.814978715
3	K4Ba2Nb4S22	0.29	0.957665397	K Ba Nb S	-0.667665397
4	K2Ho4Cu4S9	1.926667	1.78392936	K Ho Cu S	0.142737307
5	Pb0.9Mn0.1Te	0.278833	0.254484975	Pb Mn Te	0.024348358
6	RbInSnS4	2.96	3.57878207	Rb In Sn S	-0.61878207
7	PbGa2GeSe6	1.31	1.087782872	Pb Ga Ge Se	0.222217128
8	EuYb2S4	4.1	4.0814794	Eu Yb S	0.0185206
9	Hg0.01Cd0.99S	3.67	5.616579075	Hg Cd Se	-1.946579075
10	VO2	3.26	2.886051502	V O	0.373948498
11	InSb0.2As0.8	2.48	2.085664494	In Sb As	0.394335506
12	EuIn2S4	1	1.355621908	Eu In S	-0.355621908
13	CdIn30Te46	2.16	1.869877649	Cd In Te	0.290122351
14	Cu3VS4	1.81	1.198602672	Cu V S	0.611397328
15	NaGe3P3	2.64	2.495712255	Na Ge P	0.144287745
16	CsDy9Cd4Se18	2.3	1.974589013	Cs Dy Cd Se	0.325410987
17	Hg0.77Cd0.23T	1.725	1.869572861	Hg Cd Te	-0.144572861
18	In4S5	1.18	0.9668109	In S	0.2131891
19	BaGa2Se4	3.4	2.978371702	Ba Ga Se	0.421628298
20	Ce2Sn3Se9	4.22	4.041873144	Ce Sn Se	0.178126856
21	Ag2O	0.123333	0.12148003	Ag O	0.001853303
22	CuIn5S8	0.7	0.840249935	Cu In S	-0.140249935
23	BaGa4S7	1.83	1.673360331	Ba Ga S	0.156639669
24	CsCd4In5Se12	0.43	0.352246184	Cs Cd In Se	0.077753816
25	Cd4GdB3O10	1.08	0.853661206	Cd Gd B O	0.226338794
26	Ca2Si	2.2	1.983242966	Ca Si	0.216757034
27	Cs2ZnGe3Se8	1.9	1.756622402	Cs Zn Ge Se	0.143377598
28	Rb1.45Pb3.1Sb	3.7	3.932230091	Rb Pb Sb Se	-0.232230091
29	AgIn9Te14	1.95	2.116740627	Ag In Te	-0.166740627
30	Rb3CdB5O10	2	1.584567136	Rb Cd B O	0.415432864

Fig. 2 Predictions Data Table

3.3 Histogram Table

The files in hist_data folder (see Figure 3) are the data that we load to create the element's residual view. Each row of this file describes a range of residual. The attributes in this table contains $\text{bin}_count, bin_freq, bin_rangeandbin_numb$

	A	B	C	D
1	bin_count	bin_freq	bin_range	bin_number
2	14	0.21875	(-0.142, 0.0165]	13
3	13	0.203125	(0.0165, 0.175]	14
4	9	0.140625	(-0.301, -0.142]	12
5	8	0.125	(0.175, 0.333]	15
6	5	0.078125	(-0.459, -0.301]	11
7	3	0.046875	(0.333, 0.492]	16
8	3	0.046875	(-0.618, -0.459]	10
9	2	0.03125	(0.809, 0.968]	19
10	2	0.03125	(0.492, 0.651]	8
11	2	0.03125	(-0.935, -0.776]	17
12	1	0.015625	(0.651, 0.809]	7
13	1	0.015625	(-1.093, -0.935]	0
14	1	0.015625	(-1.727, -1.569]	18
15	0	0	(-2.044, -1.886]	6
16	0	0	(-1.886, -1.727]	5
17	0	0	(-0.776, -0.618]	4
18	0	0	(-1.569, -1.41]	3
19	0	0	(-1.41, -1.252]	2
20	0	0	(-1.252, -1.093]	1
21	0	0	(-2.203, -2.044]	9

Fig. 3 Histogram Table

3.4 Element Data Table

The files in element_data folder (see Figure 4) are the data that we create to load line graph view. Each row of this file describes a formula. The attributes in this table contains formula, actual, predicted , element and residual. actual and predicted attributes provide the information to allow us to format line graph view.

4 Discussion of Feedback from Peer Review

4.0.1 Comments from Peers (response in Cyan)

1. You will need a good explanation of how the data is to be used. This could be a tool tip, or a text box.
   - This is a great idea. I think the text box will be the most appealing for users, as it will be easy to see and doesn't depend on mouse location to give information.
2. If you include machine learning features in your figures, you will need to show how those are derived.

	A	B	C	D	E
1	formula	actual	predicted	elements	residual
2	AgAlTe2	0.035	0.1279492	Ag Al Te	-0.0929492
3	AlN	6.89	6.033343363	Al N	0.856656637
4	BaAl4Se7	4.76	4.709375625	Ba Al Se	0.050624375
5	In0.6Al0.4P	0.365	0.749499896	In Al P	-0.384499896
6	Ga0.5Al0.5As	2.23	2.080964205	Ga Al As	0.149035795
7	In0.99Al0.01P	2.64	3.065795527	In Al P	-0.425795527
8	Li6Al2Te8O22	0.236666667	0.365359098	Li Al Te O	-0.128692431
9	Cs2Al2B2O7	1.56	1.519802546	Cs Al B O	0.040197454
10	Ga0.01Al0.99P	1.1	0.585675597	Ga Al P	0.514324403
11	Ba8Al10B12O41	4.76	4.232035163	Ba Al B O	0.527964837
12	AlVTe2O8	3.78	4.370909274	Al V Te O	-0.590909274
13	CuAlSe2	0.24	0.407407173	Cu Al Se	-0.167407173
14	Ga0.25Al0.75As	1.86	2.676783984	Ga Al As	-0.816783984
15	NaAlGeS4	1.03	1.595339969	Na Al Ge S	-0.565339969
16	In0.8Al0.2P	6.443333333	6.262782046	In Al P	0.180551288
17	Ga0.85Al0.15As	2	1.826215518	Ga Al As	0.173784482
18	Al2S3	2	2.021122155	Al S	-0.021122155
19	AgAlO2	0.2695	0.468579571	Ag Al O	-0.199079571
20	Ga0.7Al0.3As	2.54	2.349350999	Ga Al As	0.190649001
21	Cd0.01Te0.01Al0.	3.68	3.693911035	Cd Te Al Sb	-0.013911035
22	Li4.5Al0.5Te06	0.783333333	0.784428831	Li Al Te O	-0.001095497
23	Ga0.4Al0.6As	2.79	3.64249763	Ga Al As	-0.85249763

Fig. 4 Element Data Table

   - We will include an info box that will have links to relevant publications, and web pages if the user wants to learn about the specifics of this data.
3. You need to ensure that the elements are very clearly marked in each data view.
   - This is a great idea. It will probably be useful to have figure captions similar to a publication.
4. It will be important to label charts and figures. Make sure that the user can understand each figure from the labeling.
   - We agree labeling will be important. However we do not want to present to much test to the user.
5. You need to add a full description on what the tool can do. Perhaps a question mark that has a tool tip on hover. Regardless, you need to provide A full description of what this tool uses.
   - We will provide a link in the general info box that fully describes how the data was generated, and gives commentary (if there is any) on specific trends that are interesting.
6. You should include a introduction to the issue for those who are not in materials science
   - We will try to have a link in generate info box "Click here if new to materials informative" (or something like that)
7. You need to make sure that the heat map is clearly explained and useful. It is the first point of contact for the user.
   - This is extremely important. We have added a legend that explains the cut-ff used for the different displayed colors. We will also try to add buttons that allows the user to explicit select a material property (and inherently show what property is displayed).
8. Focus on story telling. Perhaps you can include a small tutorial that shows the user how it works. You could also put the tool in a particular state at start up.
   - We are not quite ready to handle this step yet. This is a tool that is designed to help with the data analysis. Once the tool is more mature, we can look for an interesting story and tailor things in a way that highlights the fact.
9. Add a way for people to see selected compounds on the main page.
   - This is a great feature. We will implement this once the main data views are properly communicating with each other.
10. Add brush or hover to highlight compounds and make specific space to display results.
   - This brings up a good point about displaying data. Currently we have a too tip, but that can hide the data and is not always visible. We will consider adding a box superficially dedicated to displaying values of interest
11. One second page. Make it so that there is explicit explanation of how the combining elements will work
   - Hahaha the second page will probably not be added as a part of this project

   .
12. You can probably drop the TSNE and the histogram if there is not enough time.
   - The TSNE should be a relatively simple figure to generate. Interaction with the TSNE is likely to provides an interesting aspect of "the story". This comes from exploration of similar (though less functional) tool provided by Citrine Informatics.
13. If you can only show one thing on the heat map, it should be the residual.
   - This is an intersting comment. It makes a lot of sense to show residual. On the other hand, showing the relative frequency of data makes interaction with the data easier.
14. You can probably get away with not having the brushing. Instead show a list of all the formula that have the currently selected elements.
   - This will almost certainly be the solution we implement.
15. All of it is too aggressive. Probably just a part of the first page is enough.
   - Agreed.
16. The visualization techniques are boiler plate. The way that it works is useful, but necessarily novel. Not very innovative. Maybe have to reduce the amount of data that can be explored.

- The original data has been dropped for a subsection of data (44,000 to 2,500) that only includes examples with experimental band gap information.

17. If you have too much data, you can only plot things that are clicked on.
    - This a good idea, and has been added to the current iteration of the act_vs_pred data view.

5 Project Status: Mid-semester update

The ptable view was the first view implemented. This is the main tool for interacting with the data (see Figure 5). This provides the user with the ability to investigate the data on an element-by-element basis. This form of the ptable view allows the user to hover and click on individual elements. On hover, the act_vs_pred view displayes formulae with those elements. On click, the data is updated to only show formulae with that element (as seen in Figure 6). We have currently implemented "or" logic (graph shows formulae with "Al" or "Ag") if two or more elements are clicked. This means that you only add new formulae with each subsequent element click. We hope to eventually add "and" logic as well, allowing the user to drill down into subsets of material types.

5.1 Periodic Table

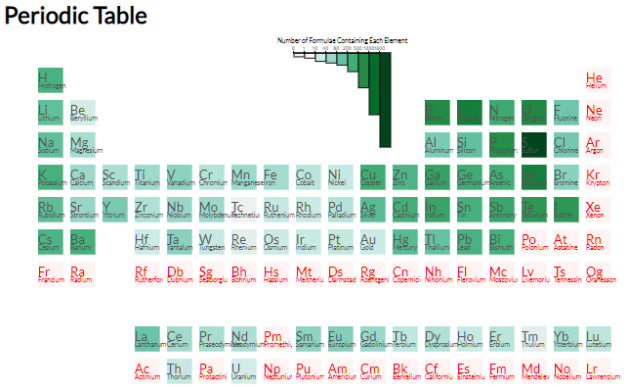

Fig. 5 Version 1 of the ptable view. Colors are not decided, but the heat map and element layout work well for interaction.

There is currently no interactivity with heat map specifically. The heat map is also limited to showing the relative frequency of formulae that contain the specific constituent element (eg. Ag is a part of 10-40 elements). Currently, the ptable does communicate with the act_vs_pred data view. This allows us to load the proper element data on selection making for a smooth and fast user experience. The current color scheme on ptable is a place holder until we get a better overall feel for the web page. The green squares all have representative data. The red squares indicate that there are no examples of that element in the data set for the user to interact with.

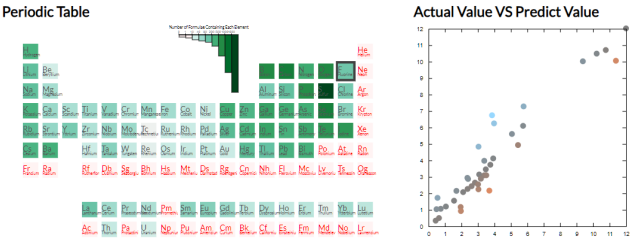

Fig. 6 When elements are clicked on, they are classed as 'selected'. This allows the other views to read to appropriate data and update to only show formulae specified by the user.

5.2 Actual vs Predicted Values

This view gives the user a soft understanding of error in the data set. The figure will be labeled to indicate that more accurate data lies on the 45° line (which will be implemented soon). The residual error is encoded with color to help the user understand the relative error at a glance. This essentially acts as the main display while the user explores the data (Figure 7). This view dynamically changes as the user interacts with the ptable view. Figure 8 shows the click and highlight function both in use. The click function reduces the displayed data to those containing silicon. The hover element then highlights the subset that also contains phosphorus.

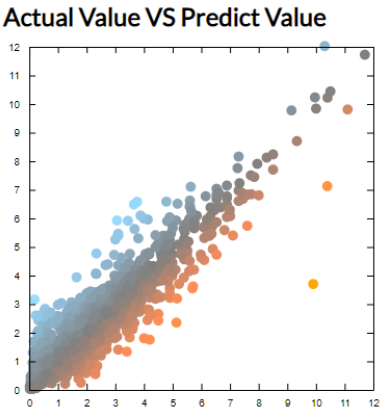

Fig. 7 The actual band gap vs the predicted values are shown here. This allows for quick feed back while the user explores their data.

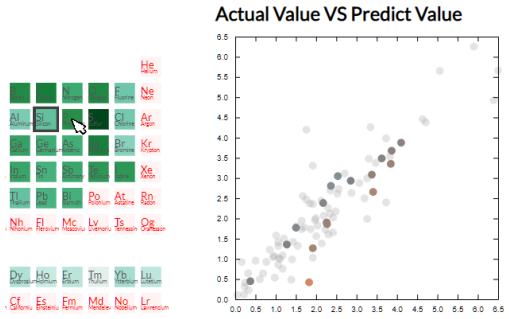

Fig. 8 The "highlight on hover" feature allows the user to preview data without removing data from the current selection. This will likely extend into all data views; however, the exact functionality is still being explored.

6 Project Status: Final Submission

6.1 Periodic Table

The heat map works to well to help user navigate the data set. We would have liked to implement the ability to change the property encoded on the periodic table, time became an issue. As it stands the periodic table view is a nice interactive element. The user is greeted with Figure 9 the table upon entering the site.

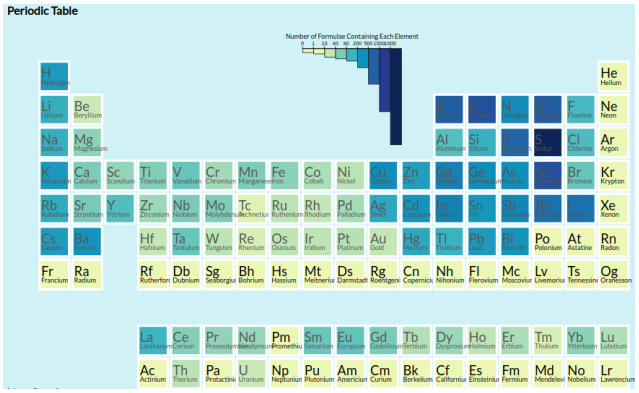

Fig. 9 The improved periodic table view. Colors are updated to consider common color blindness. The element layout works well with the mouse interaction, and gives the user a satisfactory feeling of control.

As the user hovers over the elements in the periodic table, the other views respond. This indicates to there user that the periodic table is interactive. If the user then chooses to click on a single element the periodic table populates a histogram of the residual values for band gap prediction on all compounds associated with the selected element. This gives us Figure 10. If the user decides to click on additional elements, a new histogram is calculated which includes the new information. This allows the user to explore not only the residual associated with element families, but also single elements. As more elements are selected, the size and number of bins dynamically change. Figure 11 demonstrates a user looking at the combined residual of the group 4 metals: titanium, and zirconium.

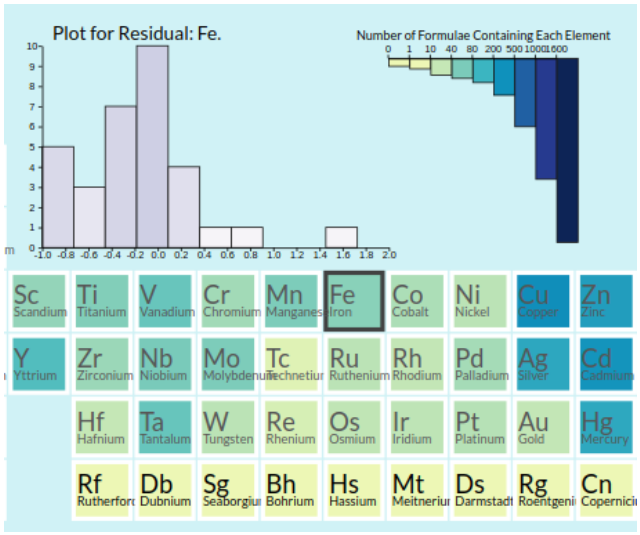

Fig. 10 The residual values for compounds containing iron (Fe) are displayed when iron is selected

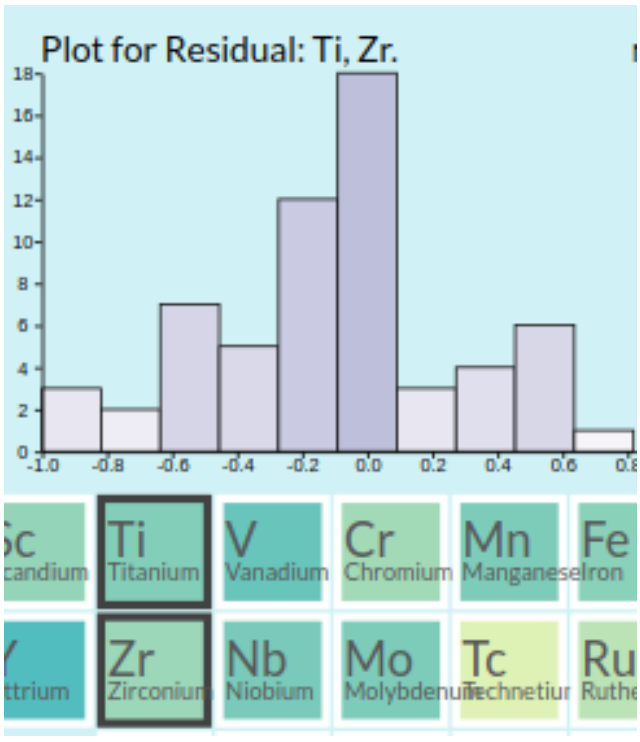

Fig. 11 The residual values for a selection of different elements is shown.

If the user then unselects all clicked elements, the residual plot defaults to a view that shows the residual distribution for the entire data set (Figure 12). You can notice that there is a much darker purple when all residuals are shown. Because the limits of the y-axis dynamically change, the use of color as a redundant encoding of bin size gives the user better intuition when looking and exploring the different distributions.

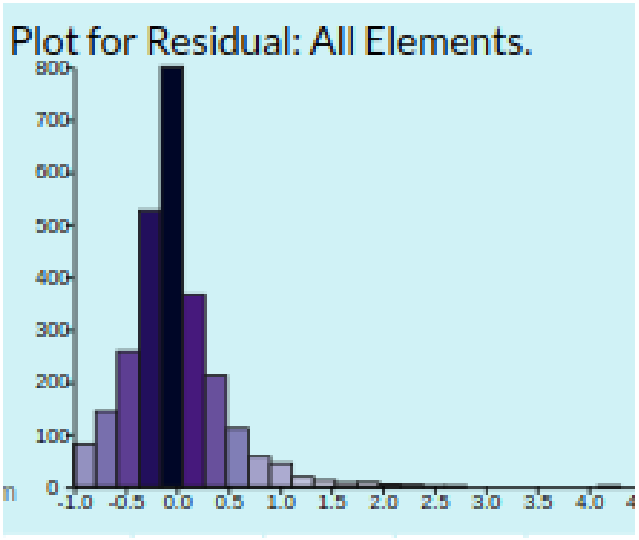

Fig. 12 The residual for all data is displayed if no elements are selected.

6.2 Actual vs Predicted Values

The act_vs_pred view, which contains the plot of actual versus predicted band gap values, has changed the least since the initial implementation. The majority of changes are thematic to match our style choices for the rest of the website (see figure 13). This plot responds to interaction with the periodic table. On hover, entries for the respective element are previewed. On click, the previewed points are added to other selected elements. As seen in Figure 14, this plot provides the user with a quick understanding of the general performance (the range along the x-axis indicates band gap) and the predictive behavior (patterns in the residual can be identified) for the selected elements.

6.3 T-distributed Stochastic Neighbor Embedding

The tsne view is an exciting addition since the mid-semester update. The same chemical information that were used to generate the act_vs_pred view were also used to generate the t-SNE. Figure 15 shows the t-SNE plot, which allows the user to get an intuitive understanding on how the machine learning might have been possible.

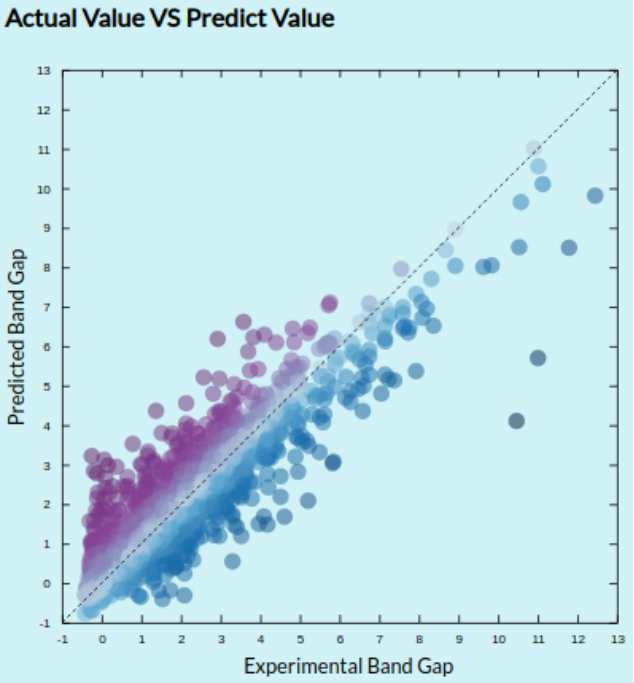

Fig. 13 The actual vs predicted graph underwent a thematic change. This included the addition of axis labels and a guide line for perfect performance.

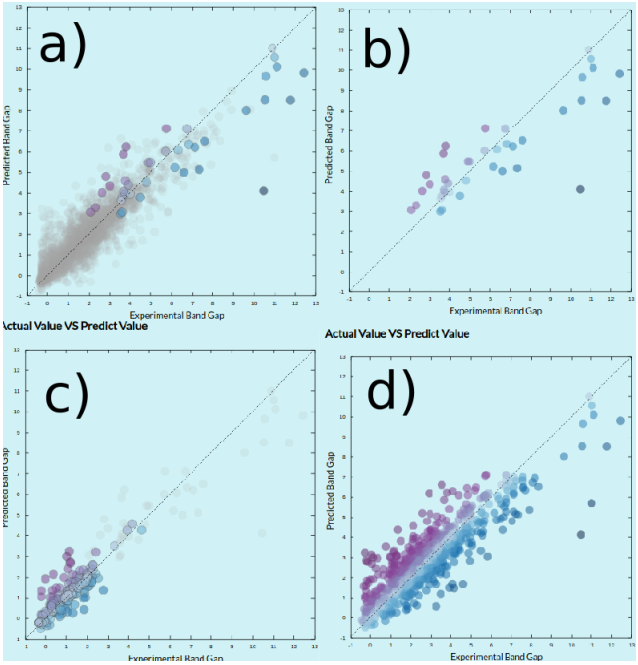

Fig. 14 The interactive elements of the actual vs predicted view can be seen in a) highlight on hover (no selection) b) selection of highlighted element c) highlight on hover (with selection) and c) multiple selections on the same graph.

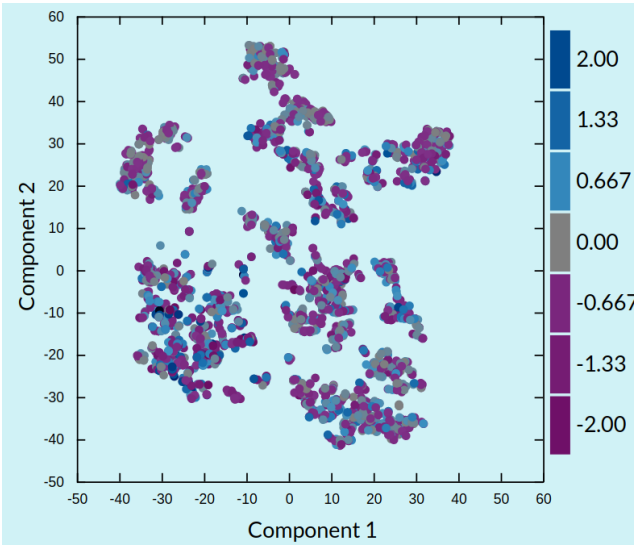

Fig. 15 The t-SNE is a powerful tool for the user as it allows for a more chemistry based exploration of bang gap and residual values from the machine learning algorithm.

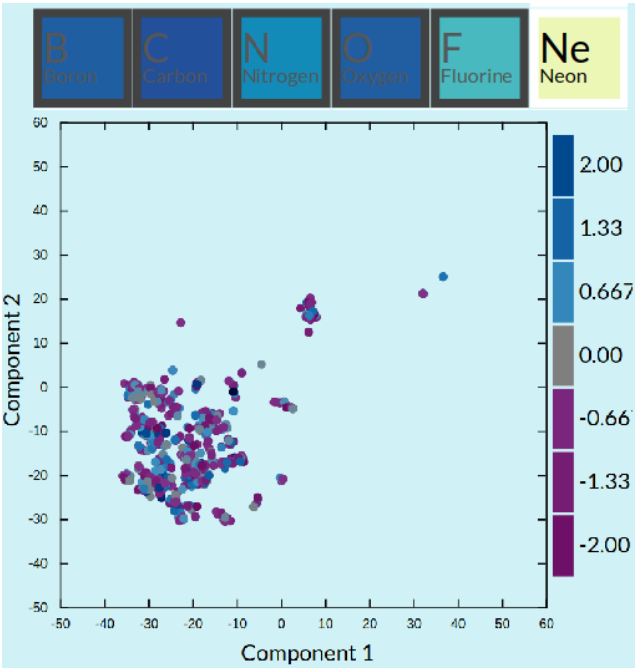

Fig. 16 Grouping of common elements can easily be seen. This is interesting and hints and possible issues with the modeling process. If similar groups are not separated during model training, we may not be learning the necessary relationships to make truly useful discoveries.

By plotting the t-SNE data with interactive elements, we enable the user to explore how chemistry changes between element families and groups. This can be done by interacting with the periodic table to highlight, and select elements, with interesting results shown in Figure 16. To go along with this ability to dynamically change the displayed data, we also give the user the ability to toggle between the machine learning target value, band gap, and the residual value from the learning. This allows for the simultaneous exploration of property relations, but also error analysis (See Figure **??**).

The t-SNE plot does a good job of allowing the user to do "chemistry exploration". To compliment the users ability to explore the chemistry with the periodic table, we also added a brush function to the view (Figure 18). The brush function allows the user to select a group of interest. Upon selection of the desired clusters, bar plots of the average residuals are generated on an element by element basis. The ability to quickly navigate through chemical space is both engaging and insightful. This ability to quickly identify element and high or low error clusters gives great insight into the chemistry and predictive attributes of each cluster (See Figure 19.

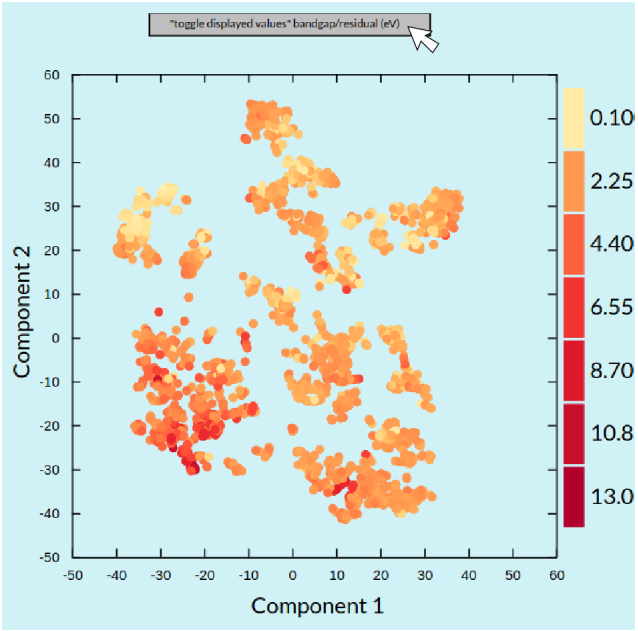

Fig. 17 Band gap and residual values are encoded with color. The user can select either scheme while looking at the data.

7 Project Insight

7.1 Patterns in the data

Overall, elements that were included in this data set have a normal distribution of the residual values. That said there are certain elements that violate this pattern, we were able to identify a few that are interesting. Take Barium for example, this element has an abnormally large spread in variance. This was not something most material scientists would expect a priori. However, looking at the t-SNE, we see that Barium is

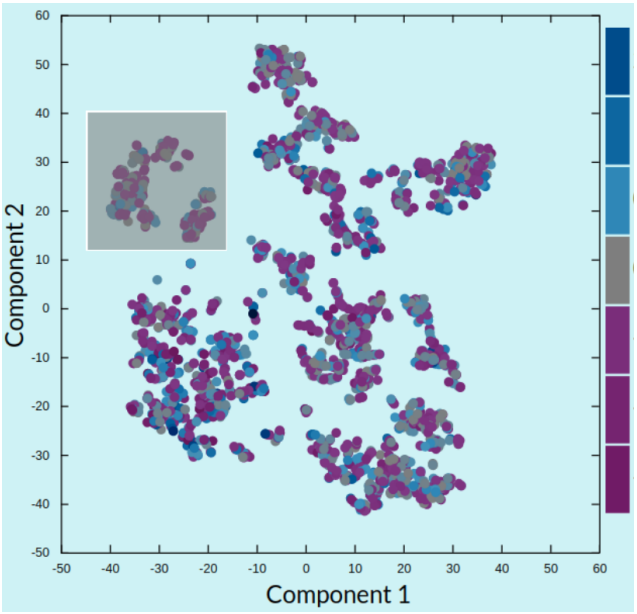

Fig. 18 The ability to brush allows the user to explore individual clusters which they may find interesting.

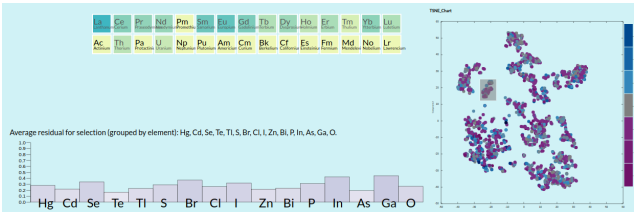

Fig. 19 The average residual error for each element in the cluster is displayed as an aide in developing chemical intuition.

somewhat unique with its lack of clustering. That said, there are examples on non-clustered elements such as manganese. Perhaps there is interesting chemistry that could be done to investigate how clustering and dimensionality reduction corresponds to learnablity of data. Also of note, some elements have more extreme errors. Fluorine and oxygen both exhibit relatively high residuals values, and interestingly are tightly clustered in the same region of the t-SNE graph.

7.2 Work flow improvement

This project was relatively smooth. The majority of the success was due to a well planned schedule that included meetings long before the deadline, and a willingness to do significant work in the early stages. Code review was also effective, and allowed for both team members to work on all parts of the project. Unfortunately, neither team member had historic knowledge of JavaScript. This was a major setback, as simple tasks such as reading multiple "csv" files ended up frustrating progress for many days. Also, the ability to consult with an expert (teaching assistant) as a part of the team is a invaluable resource that we should have taken more advantage of.

7.3 Visualization improvement

The websites overall presentation is hindered by amateur selection of color pallets. The inherently in-balanced element data exacerbated the issue of selecting color scales. Organization of the visualization of the website could also be improved if more time was available. The inclusion of the actual chemical features would also be nice to reference especially in context of the t-SNE.