JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY

DOCTORAL THESIS

Elucidation of physical similarity concepts by machine learning approach

Author:
NGUYEN Duong Nguyen

Supervisor:
Assoc. Prof. DAM Hieu Chi

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Dam laboratory
School of Knowledge Science

September 24, 2019

Declaration of Authorship

I, NGUYEN Duong Nguyen, declare that this thesis titled, "Elucidation of physical similarity concepts by machine learning approach" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this Institute.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this Institute or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date _____

"The thing that doesn't fit is the thing that most interesting."

Richard Feynman

JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY

Abstract

Knowledge Management
School of Knowledge Science

Doctor of Philosophy

Elucidation of physical similarity concepts by machine learning approach

by NGUYEN Duong Nguyen

Canonical similarity measurements in machine learning determine relationships among data objects over the description space to achieve efficient inferences. In contrast, in human recognition and materials science, the similarity between two objects depends on whether they follow a common mechanism/function. In this thesis, we show the use of existing machine learning model as well as developing methods to determine similarity–dissimilarity among data objects with respect to physical properties.

Acknowledgements

I would like to express my sincere thanks to my advisor, Professor Dam Hieu Chi, for allowing me to participate and help me find happiness in doing scientific research. Please let me emphasize again that, the happiness of doing scientific research is the most valuable that I have earned here. It was the harmony between several layers of emotions and experiences.

It was the moment when he encouraged me to revise twenty-nine times for my very first publication since the previous ones stand still had "something" not in perfectly smooth. After that, the very first "similarity" terminology was born and drives all my thesis now. It was the lesson of not compromising with anything we could do better. The emotions could be disappointed, but in the end, it was comprehended then thankful.

professional work,

naive but concise thinking,

chasing and dedicating our mind to the beauty of science,

It was apart of the lessons I have learned from him. Finally, after three years of working under his advise, I might not know yet the answer to what science looks like, but I quite certain about how the happiness of doing scientific research looks like.

Besides that, I would like to send my appreciation to all collaborators in my lab: Mr. P. T. Lam, Mr. N. V. Cuong, Mrs. N. T. Linh, Mr. N. V. Doan, Mr. D. T. Thai, Mr. L. D. Khiet, Mr. H. M. Quyet. I might not finalize my works without constructive and fruitful discussions from them. All the members experienced with me through 150 seminars, conducted in five hours each time. Let's imagine how many beautiful things I have learned from them.

All of my work was partly supported by PRESTO; by the "Materials Research by Information Integration Initiative" (MI^2I) project of the Support Program for Start-Up Innovation Hub, from the Japan Science and Technology Agency (JST); by JSPS KAKENHI Grant-in-Aid for Young Scientists (B) (Grant JP17K14803), Japan; by MEXT as a social and scientific priority issue (Creation of New Functional Devices and High-Performance Materials to Support Next-Generation Industries; CDMSI) to be tackled using a post-K computer.

Lastly, I would like to spend this thesis as thankful to my family, especially my parents. By somehow they might not contribute to detail works, but they taught me about the foundation about how to do the right thing.

...

Contents

List of Figures

List of Tables

List of Symbols

$\mathcal{D}$	data set
$\boldsymbol{x}$	representation vector of data
$\boldsymbol{y}$	target variable representation of data
$\boldsymbol{x}_i$	vector representation of data instances indexed i
$\boldsymbol{y}_i$	value of target variable of data instances indexed i
R^2	Coefficient of determination
MAE	mean absolute error
$\mathcal{N}$	normal distribution function

Dedicated to my…

Chapter 1

Introduction

1.1 Breakthrough discoveries in Materials science

Materials science is the oldest form of engineering, applied science and concurrently appears with human live. The studying about the composition of Nature eventually become defining point in Three-age history system with Stone Age, which began at 500,000 years BC 1.5, Bronze Age (2,000 years BC) and Iron Age (700 BC). The discovery of Cooper or Iron significantly effect to human society with trading development; weapon growth or urban civilization evolution etc,. In the first revolution, Bronze, a ductile alloys helps people much convenient in producing daily utilities, weapons, coins comparing with stone. At the time, people discovered both Cooper, Tin and Arsenic and Bronze is only one of various alloys were made by difference in combinations. The other with almost the same functionality could not bring something new to social. However, Iron with significant different properties as: easier to forge; hardness control by carbon composition and magnetic property become center of the second revolution. From this historic story, people earn the very first lesson that a breakthrough discovery is to find new things with difference in **functionality viewpoint**.

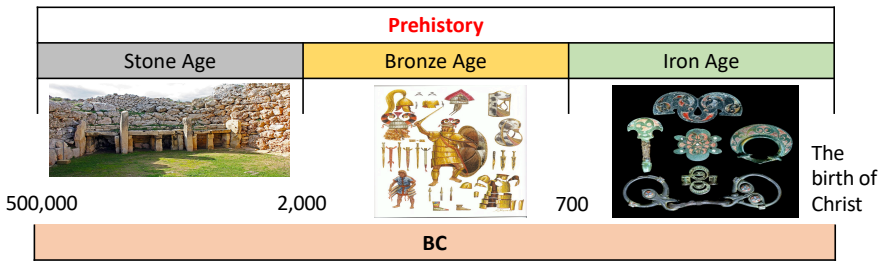


FIGURE 1.1: Prehistoric Timeline. (Source: historyinteractive.co.uk)

In the modern era, concerning to breakthrough discoveries in Science, Mendeleev periodic table should be the most notable name. Its own story, Figure 1.2 gives us useful information about the struggling the way to understand the nature. In here, I list up only some significant landmarks follows the time line. In 1829, Johann Wolfgang Döbereiner observed that there is a number of the fundamental elements that could belong to distinct groups. He called these groups as triads Döbereiner, 1829 and each triad is dissimilar to the other in considering to its chemical behavior in Nature. Lithium, sodium, and potassium, for example, were grouped together in a triad as soft, reactive metals. Döbereiner also claimed that, by ordering in **atomic weight**, there is a very simple rule throught out these triads. If taking the mass of

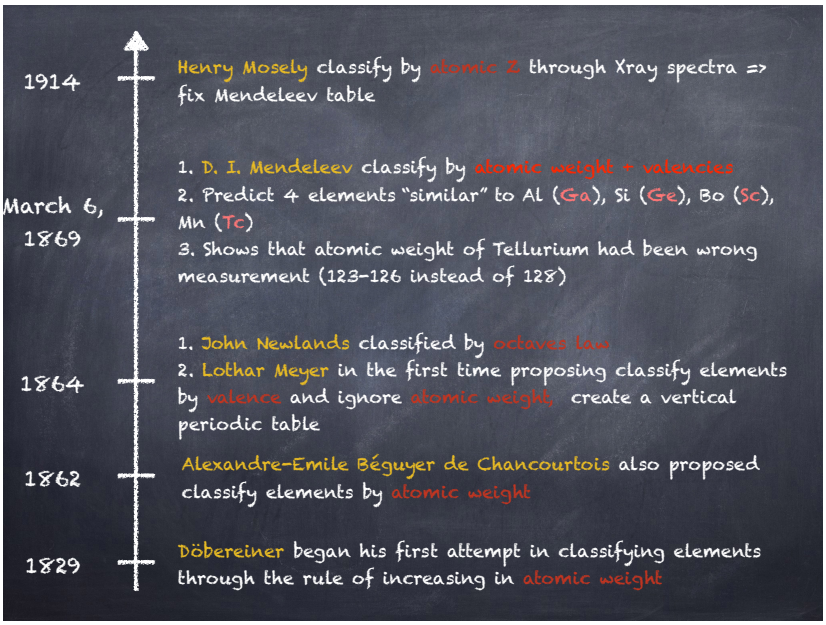


FIGURE 1.2: Notable points in history of the periodic table

the second member in every triads, we found that it is roughly equal to the average of the first element and the third element. In 1862, researcher A. E. B. de Chancourtois published an his idea about the very first formation of the periodic table "Table of the natural classification of elements, called the "telluric helix"" by taking a same sorting rule of **atomic weight** of all elements. and became the first person noticed about the periodicity of the elements. After two years, other scientists proposed different properties to classify element such as John Newlands Newlands, 8 August 1865 with **octaves law** and Julius Lothar Meyer Meyer, 1864 with **valence value** rather than atomic weight its self.

Three years after proposed version of Newlands and Meyer, on March 1869, Dmitri Mendeleev published his version D., 1869 of periodic table with classification by both **atomic weight** and chemical property shown through **valence value**. The most notable work of Dimitri Mendeleev is to predict four missing four elements which is similar to Aluminium, Silicon Boron. He predicted possible elements which are all lighter than the rare-earth element family, boron-likely (later known as Eb, under Boron, 5), aluminium-likely (later known as Ea, under Al, 13), manganese-likely (later known as Em, under Mn, 25), and silicon-likely (later known as Es, under Si, 14), proved to be good predictors of the properties of Scandium (21), Gallium (31), Technetium (43), and Germanium (32) respectively.

In fact, the periodic table we use today is not exactly the Mendeleev version. The major revision was done by Henry Moseley, Moseley, 1914 in 1914, only one year before his death at Gallipoli, with his classification by **atomic number, Z**, rather than weight by analyzing relationship of X-ray wavelength of an element and its atomic number. He corrected the Mendeleev version by placing argon before potassium in spite of the fact that weight of Argon (39.9) is slightly bigger than the Potassium's atomic weight (39.1). His method introduce a new order with the existed periodic table and show more agreement with the chemical behaviors of these elements. In fact, Argon is a noble gas and potassium is an Alkali-metal, two groups of elements even closely in pair of atomic weights but significantly different in chemical reaction. By doing the same method, Henry Moseley swithced the possitions of Cobalt before Nickel and explain that Tellurium should be placed before Iodine, without the need

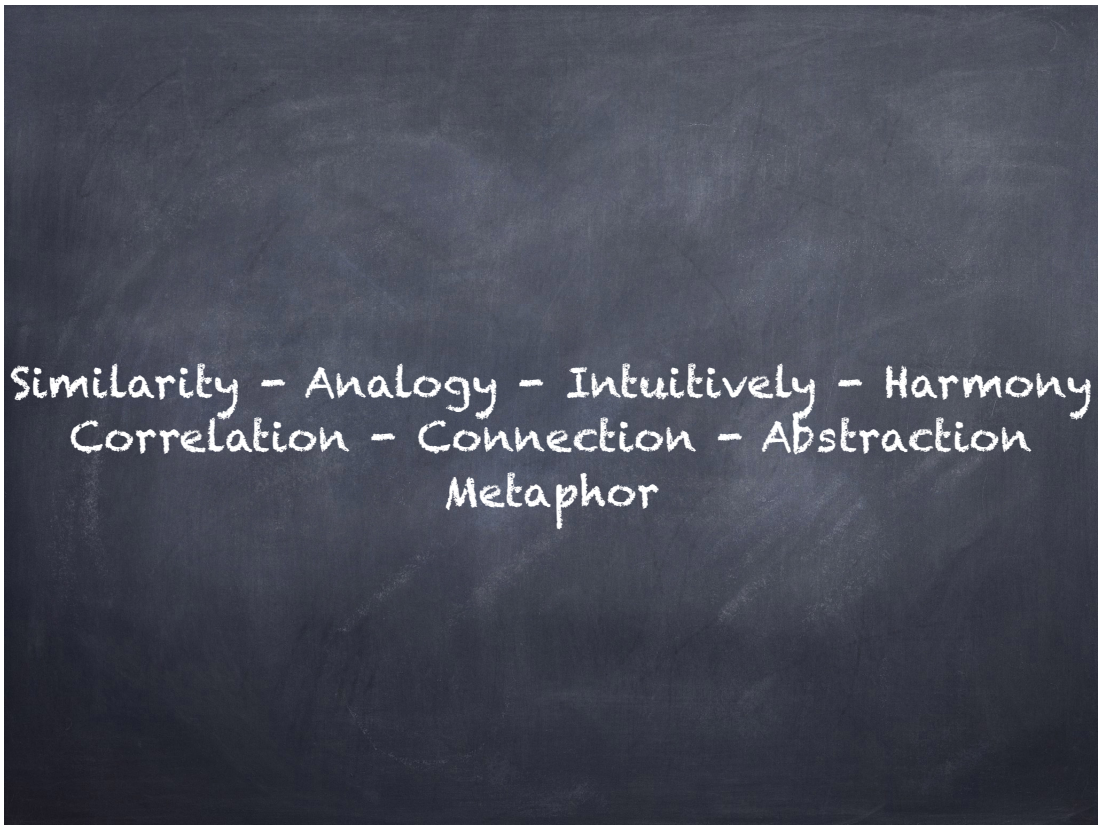


FIGURE 1.3: The way creativity is driven, and so the world does.

of reconsidering the order of their experimental atomic weights.

The story of the breakthrough discovery, periodic table again shows us about the **similarity concepts among elements relying on commons in chemical properties**, rather than its weight, the description values.

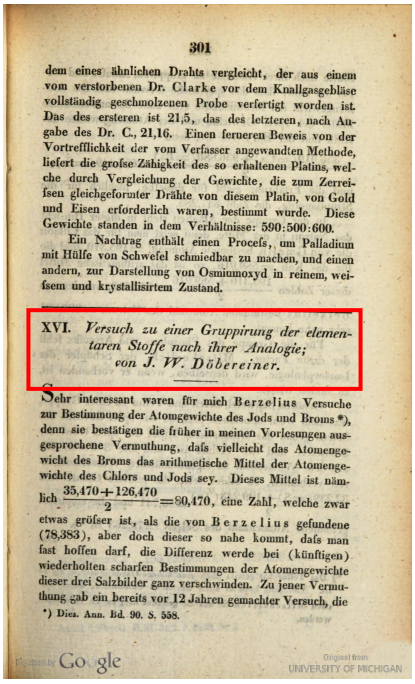
1.1.1 Message

- Breakthrough discoveries must be objects with dramatically changes in functional viewpoint.
- People prefer grouping similar objects under a given property viewpoint, rather than its description.

1.2 Similarity, Intuitively, Harmony

Analogy is the core of cognition Breakthrough discoveries are important but they have already existed. Awareness about the way the discoveries made by or the origin of human cognition drive the process of creating discovery might help us go far more away, i.e to warming up, providing guidelines for future discovery. Focusing on this meta-knowledge, the essence of this section, or in general, the essence of my thesis is the similarity.

Correlation is equivalent to Similarity Stepping back a little into the previous example of the breakthrough finding of the periodic table. One of the very first clues in finding a periodic table made by J. W. Döbereiner with his paper about: "An attempt to group elementary substances according to their analogies". Once again,



301

dem eines ähnlichen Drahts vergleicht, der aus einem vom verstorbenen Dr. Clarke vor dem Knallgasgebläse vollständig geschmolzenen Probe verfertigt worden ist. Das des ersteren ist 21,5, das des letzteren, nach Angabe des Dr. C., 21,16. Einen ferneren Beweis von der Vortrefflichkeit der vom Verfasser angewandten Methode, liefert die grofse Zähigkeit des so erhaltenen Platins, welche durch Vergleichung der Gewichte, die zum Zerreifsen gleichgeformter Drähte von diesem Platin, von Gold und Eisen erforderlich waren, bestimmt wurde. Diese Gewichte standen in dem Verhältnisse: 590:500:600.

Ein Nachtrag enthält einen Procefs, um Palladium mit Hülfe von Schwefel schmiedbar zu machen, und einen andern, zur Darstellung von Osmiumoxyd in reinem, weifsem und krystallisirtem Zustand.

XVI. *Versuch zu einer Gruppirung der elementaren Stoffe nach ihrer Analogie; von J. W. Döbereiner.*

Sehr interessant waren für mich Berzelius Versuche zur Bestimmung der Atomgewichte des Jods und Broms *), denn sie bestätigen die früher in meinen Vorlesungen ausgesprochene Vermuthung, dafs vielleicht das Atomengewicht des Broms das arithmetische Mittel der Atomengewichte des Chlors und Jods sey. Dieses Mittel ist nämlich $\frac{35{,}470+126{,}470}{2}=80{,}470$, eine Zahl, welche zwar etwas gröfser ist, als die von Berzelius gefundene (78,383), aber doch dieser so nahe kommt, dafs man fast hoffen darf, die Differenz werde bei (künftigen) wiederholten scharfen Bestimmungen der Atomengewichte dieser drei Salzbilder ganz verschwinden. Zu jener Vermuthung gab ein bereits vor 12 Jahren gemachter Versuch, die

*) Dies. Ann. Bd. 90. S. 558.

"An attempt to group elementary substances according to their analogies".
Johann Wolfgang Döbereiner,
Annalen der Physik und Chemie.
2nd series (in German). **15**: 301–307, 1829

FIGURE 1.4: The J. W. Döbereiner paper in 1829 which shows his identifies about Triads correlation

the analogies. The analogies in his paper mentioned about the Triads Döbereiner, 1829 or Figure 1.4 which are shown later in the same group: Alkali-forming elements (Lithium, Sodium, Potassium), Salt-forming elements (Calcium, Strontium, Barium), Acid-forming elements (Chlorine, Bromine, Iodine), Transition metal elements (Iron, Cobalt, Nickel). He found these groups by trivial observations: the mass of the middle element is an average of the two others. By his analysis, these correlations maybe imply some hidden knowledge about periodicity among elements in Nature.

Intuitively is equivalent to Similarity In Science, especially in studying Physics, the word Intuitively is absolutely common if any researchers talk about his/she very initial idea coming up. Intuitively is somehow equivalent to Similarity. They all describe how the researcher map from complex, abstract, debatable phenomena to a simple one, in their mind. If the simple one is easy to transfer the key idea to other people, then the Similarity is the key to create the simple one.

Harmony is equivalent to Similarity In music, harmony are considered as an exhibition in which a number individual sounds of different music instruments or person are composed. In other meaning, harmony could be a composing of many super positions of sounds. The composition should be made by integrating Similar individuals by predefined hidden rules rather than a random mixing. In scientific collaboration, the harmony in knowledge creation happens in the seminar room, in casual discussion in the bar, smoking area, etc or "bar" terminology in knowledge creation theory. In these mediums, people share a common problem, seeking for the same purpose. The harmony atmosphere happens as the key and scientific results are the consequences.

Metaphor, the art of interdisciplinary, is equivalent to Similarity

The metaphor normally is figure of speech in literature, poem; pictures in fine-art photography; players's action in stages etc which all try to map from visualizable objects to abstract object. Artists working on these fields encode messages into their

creation by using the similarity language. For that similarity language, people in different field could earn the same essence of the story. Further coherence and resonance from that could be developed.

Above aspects related to the similarity is to show how important of similarity effects to human cognition. In the next section, we show a new power era to extract similarity resources named Data mining.

1.3 Canonical methods in Materials science research

Human living exists on the Earth over 50,000 years and we have had thousands of years looking for scientific questions about Nature. In the field of materials science research, through the effect of sociality, people equipped knowledge, methodologies from other scientific domains as physics, chemistry or computer science. Nowadays, rather than hand-craft pencil, we have a number of utilities to do research, especially the new kind of knowledge and research methods in data science era. Before data science, there are three scientific paradigms that support materials scientists to conduct their research: empirical, theoretical and computational. These fields associated with a lot of connections to people desire to work in data science as brief description in follows.

Empirical science is the first paradigm for people research on the materials topic. In the beginning era, mostly scientific discoveries were made by heuristic and empirical experiment works. People firstly design experiments with assumption about the expected results. Next, they run experiments and write down in detail all input and output for each running time. After collecting a number of results from trials, they look for correlations inside the result and revise the initial assumption and experiments. Finding correlations and revise the initial assumption is the most important part in this era. Indeed, the method mostly depends on insight of each individual researcher. The works described 3, Chapter 5 could be considered as a form of empirical science equipped with modern techniques.

Theoretical science has begun since the 17th century, under the strong development in mathematics with algebra, differential calculus, theoretical physics inherits a strong foundation to develop deductive studying that modeling nature behaviors. The way of modeling with mathematics analysis tools called physical laws with blooming human understanding in classical mechanics, electrodynamics, etc. Starting in this era, people always looking for beauty functions that describe behaviors of materials. The conventional thought about the functional relation among materials connects to the central works in Chapter 3.

Computational science paradigm started in 1950s under the deep understanding in quantum physics and computer science. People could have an additional method to solving scientific question by simulating physical phenomena under the help of numerical calculation. Density functional theory and molecular dynamics are remarkable examples in this era. This paradigm also create a new layer of data, as the viewpoint of data science.

The last paradigm, big-data driven science with machine learning, data mining techniques, people now working in materials science have additional tools to analyze various type of observed or simulation data. By definition, Machine learning provides systems the ability to automatically learn and improve from experience without being explicitly programmed. By constructing these algorithms, Machine learning behave human-like action with outstanding performs in various fields Marr, 2018: image recognition, speech recognition, medical diagnosis, natural language

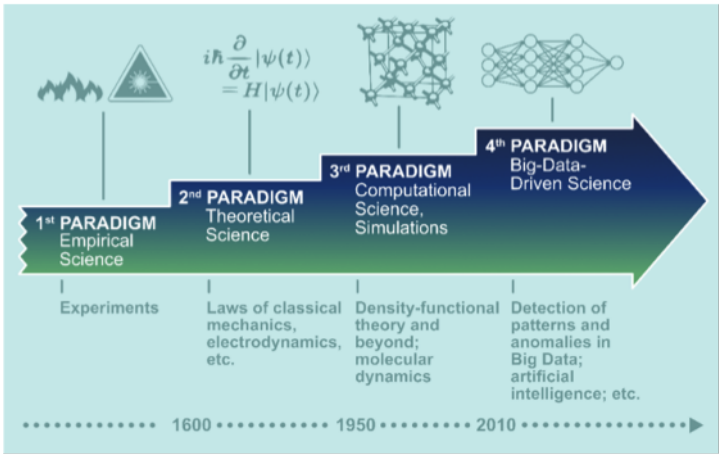


FIGURE 1.5: Four paradigms in materials science development. (Source: https://www.nomad-coe.eu/news/147/39/NOMAD-establishes-new-fourth-paradigm-in-computational-materials-science)

processing, automobile etc,. Learning from experience, or if the learning process inside the machine actually happen by finding analogy as human, it is equivalent to that we have a new source of providing the Similarity.

Nowadays, human knowledge in materials science has a potential chance to exceed a new horizon. However, due to the behavior of machine learning already performed, people have on-going questions: Is there any new kind of knowledge human could not understand? How about the human role in knowledge discovery process? We discuss these topics and show the answer in the next section.

1.4 Machine learning: new similarity miners

The computer machine, thanks to the development of computer science could show human-like behaviors in various domains. By Pedro Domingos, Domingos, 2015 his book provides an overview of learning algorithms by categorizing the people who use them. The author called each category as a tribe. Each tribe has its own master algorithm for prediction. From the human cognition viewpoint discussed in section 1.2, each tribe has their own way to mine similarity information among phenomena. The critical point here is that the way to modeling their found analogy information either in machine language (black box to human) or understandable (transparent to human). The five tribes in applying Machine learning algorithms show as follows:

- Symbolists: In this tribe, person used the learning method as the inverse of deduction by taking apparent ideas from philosophy, psychology, and logic.
  - In this tribe, the master algorithm is inverse deduction
- Connectionists: Experts who working in this tribe are inspired by neuroscience and physics. The most nominal work is reverse engineer of the brain operation.
  - In this tribe, the algorithm's master is backpropagation
- Evolutionaries: They simulate evolution on the computer and draw on genetics and evolutionary biology.

  - In this tribe, the master algorithm is genetic programming

- Bayesians: They believe that learning is a form of probabilistic inference and have their roots in statistics.
  - The master algorithm frequently use by this tribe is Bayesian inference

- Analogizers: They learn by extrapolating from similarity judgments and are influenced by psychology and mathematical optimization.
  - The most nominal algorithm's master is Support vector machines.

The master algorithm that connects five tribes show how Domingos analyze the role of each tribe to discovering new knowledge. Firstly, existed knowledge expands and envolved through combination, mutation, etc, by Evolutionaries tribe with their algorithms. "The Evolutionaries believe that the mother of all learning is natural selection", Domingos, 2015. That is the reason the genetic algorithm place at the beginning of creating new knowledge process. Next, adjusting connections among new generations is the central role of the Connectionist with their back-propagation algorithm family. The Connectionist put weighting to spawn objects with reference knowledge for reducing searching space. In the fourth step, all learned knowledge either from human or machine as the above process is uncertain. Learning itself is a form of uncertain inference, which is the greatest advance of Bayesians. Finally, Analogizers take their advanced skill in recognizing similarities between situations and then inferring other similarities. They are the final component to make interpretation results from the machine to human.

In fact, the master algorithm is an artifact created by Domingos. The author believes that any permutation of these tribes could contribute to the development of the knowledge creation process. Each tribe has their own belief in explaining and modeling the learning process and as consequence, each tribe provides numerous contributions to the society.

1.5 Structure of the thesis

The thesis is organized by seven main chapters from Chapter 1 to Chapter 7. Chapter 8 shows my publication related to the thesis and an Appendix for proof in Chapter 7. Overviews of Chapters from 1 to 7 are summarized as follows:

Chapter 1 - **Introduction**: This chapter firstly look back few breakthrough discoveries in science and shows how scientific discoveries effect to the change in human society. The key point in scientific breakthrough is to discover the thing different in mechanism than the existing one or to find similarities among things. Concerning this point, this chapter shows an overview of the great ability of human: analogical thinking through distinct phenomena. Taking this viewpoint as the core of research, I consider Machine learning as a new similarity miner then show the need for unifying three fields: materials science, psychology, and machine learning.

Chapter 2 - **A dialogue between material scientists and machine learning experts**: This chapter describes critical points of people working on materials science and machine learning. The materials scientist propose the problem, with the purpose of finding physical rules and finding new structure at the highest priority level. The machine learning expert shows canonical methods as well as state-of-the-art techniques in modeling the similarity terminology to the machine. Explainable artificial intelligence with the first requirement of designing similarity that matches with human understanding is derived as a potential solution.

Chapter 3 - **Relation function: center of similarity concept**: This chapter shows the center idea of the thesis. The common mechanism driving through objects must be considered as the key in measuring similarity, rather than any measurements on description space. The use of supervised learning methods is discussed with variable combination selection. In the work of identifying regression relation, we develop linear regression-based clustering and discuss the use of non-linear regression ensemble learning method. Post-processing models and results regarding interpret the meaning of the clustering work are discussed in detail.

Chapter 4 - **Modeling similarity–dissimilarity concepts**: This chapter focuses on showing three developed methods to measure similarity, dissimilarity information among data instance. Two heuristic voting machines are developed with foundation techniques shown in the previous Chapter 3. I also discuss the meaning of the extracted results from various problem in this chapter. In the last section, an unification method to determine similarity–dissimilarity relationship among data instances by combining multiple pieces of similarity evidence. The method opens a new viewpoint to overcome heuristic measurements. The results reveal that the similarity information is a new layer of data, which is applicable for various purposes simultaneously.

Chapter 5 - **Contributions and limitations of the thesis**: Beside detail contribution described in each work, this chapter summaries an overview of contributions, advantages, and limitations of my thesis.

Appendix **??** - **Appendix for Dissimilarity voting machine**: This appendix shows supplemental information for the dissimilarity voting machine.

Appendix A - **Appendix for combining similarity evidence work**: This appendix shows proof in modeling mass function in for the model in chapter 4.

Chapter 2

A dialogue between material scientists and machine learning experts

2.1 Introduction

The chapter records a conversation between multiple viewpoints from material scientists and machine learning expertises by taking the similarity as the central topic. The similarity is not only the problem about choosing an appropriate measurement to reflect the analogical meaning between objects in machine learning. It is a long term discussion topic between human (domain experts) and human (machine learning experts) and the machine (the similarity miner). The topic widely spread through various questions: how to representation objects, defining data instance-feature viewpoint, model assumptions, the actual reason of selecting the learning model, interpretations.

2.2 The materials scientist

2.2.1 Purposes: new structure and new physical law

Materials problem includes various forms of the concerning objects as well as concerning properties. However, in general, there are only two most canonical purposes in doing the materials research. In the following, I make a brief description about these two purposes from the viewpoint of using machine learning as the solving method.

The first purpose is to discover a new material structure targeted to a given property. Nowadays, people have two main approaches for this. The first approach is to build an artificial scientist, in the form of an Evolutionaire computer program or even a kind of robot scientists King et al., 2004; King et al., 2009; Kevin et al., 2015. All of scientific reasoning process include hypothesis generation; devising experiments to test these hypotheses; physical running the experiments; interpreting the results and repeating the process are made by robots. Everything except the final results are output. In the thesis, the author would like to do experiment and reasoning any part of the process therefore the canonical approach with human interacting with the obtained results is selected.

The second purpose is to find new physical law existed hidden inside data. In the other word, a new unveiling correlations, rules of physical laws should be exposed

and interpretable in explicit form of the result. The terminology interpretable machine learning models will be discussed in the next section. However, since the object to interpretation is human, there is strictly requirement that the extracted knowledge should be in the form of human knowledge.

2.2.2 Provider: materials science's data

Materials scientist, indeed is the data provider in the context of this work. Since materials science contains a large number of objects with different hierarchy levels e.g atom, molecule, crystal, amorphous etc, there are a number of different properties – type of information associated with these objects. From the viewpoint of canonical research approaches in Materials science, each property could also be separated by theoretical, experimental or simulation data. In the following, I give a brief description about the most general form of materials science's data.

Structure information Materials can generally be further divided into two classes: crystalline and non-crystalline. Crystal structure is a description of the ordered arrangement of atoms, ions or molecules and non-crystalline materials are other materials, armorphous with absence of the ordered arrangement Hook, 2010. The symmetry of each crystalline denoted by 230 different space group indexes. There are a number of developed representations as Coulomb matrix, radical distribution function, orbital field matrix, etc that implicitly embed information about the structure

Atomic information A material structure contains a number of atoms. The difference among materials is shown through either arrangement of atoms (structure information) or properties component atoms. The most common chemical properties that used are atomic number, electron negativity, electron affinity, first ionization, etc. There are a number of implicit integrating the information in previous research.

Physical properties Materials exhibit myriad properties, including mechanical properties (stress terms, strength terms, deformation terms); chemical properties (chemical phase, chemical bonding, formation energy, atomization energy, acidity, equilibrium states, energy reactions); electrical properties (electric charge, electric current, electric field, electrochemistry); thermodynamic properties (work, energy, heat); optical properties (diffraction, dispersion, polarization, emission and absorption spectra,); magnetic properties (magnetization, coercivity, Curie temperature)

Processing information Materials science with experimental or simulation approach are all sensitive with processing parameter setting as measure time stamp, preprocessing information (quenching temperature, applied magnetic strength, particle size, type of applied catalyst), postprocessing information (TEM image, SEM image, X-ray ptychography image)

From the author viewpoint, it is ambitious to propose a general framework that is capable to deal with all types of data simultanously. In next chapters, I will describe general approaches for some of special form of data set.

2.2.3 Notable works

Physical modeling acceleration: literal meaning As a part of the first purpose in doing materials science – discovering new structure, finding a method to accelerate the physical modeling process attract a large number of scientist in Materials science community. We list in the following the most well known experts in this field.

Jörg Behler is a notable expert with Behler's representations which is useful for learning potential energy surface in chemistry Behler and Parrinello, 2007; Behler,

2011; J., 2015; Behler, 2016. Almost all his representation targeted to implementations to artificial neural network under the purpose of replacing density functional theory in molecular dynamic simulation.

Mathias Rupp is an expert working on fast modeling atomization energies Rupp et al., 2012; Hansen et al., 2013, finding density functionals Snyder et al., 2012 or by machine learning. From my viewpoint, his most notable work are the molecular fingerprint constructed by fourier series of atomic radial distribution functions Lilienfeld O. Anatole et al., 2015. Beside that, similarity between molecular Rupp and Schneider, 2010, visual interpretation of kernel-based prediction models Katja et al., 2011 others concerning to the topic of unveiling physical laws also attract him.

Considering the same target of developing representation for machine learning we have orbital field matrix representation targeted to properties of valence electron shell Tien-Lam et al., 2018, Extended-connectivity fingerprints David and Mathew, 2010

In finding hidden physical rules Chasing to the topic of finding hidden physical meaning in the data, Matthias Scheffler is one of the most notable people. His works related to identifying interpretable rule-based models that describe materials phenomena by sub-group discovery Goldsmith et al., 2017 or Insightful classification of crystal structures Angelo et al., 2018. In addition, his works also focus on analyzing the role of descriptor in using machine learning to materials science Ghiringhelli et al., 2015 and developing SISSO method for evaluating meaningful features Ouyang et al., 2018. From the viewpoint of similarity, one may seem that materials objects follow the same behavior in physics, e.g a target property are similar together. On the other hand, objects showing different behaviors under the given target are dissimilar, regardless of its description properties. In the case of setting these definitions, the problem of finding meaningful sub-group Goldsmith et al., 2017 or finding an explainable set of descriptors that drive interpretable meaning inside the data set are all converted to the problem of finding similarities that center to a given target property.

2.3 The machine learning expert: explainable AI and similarity modeling

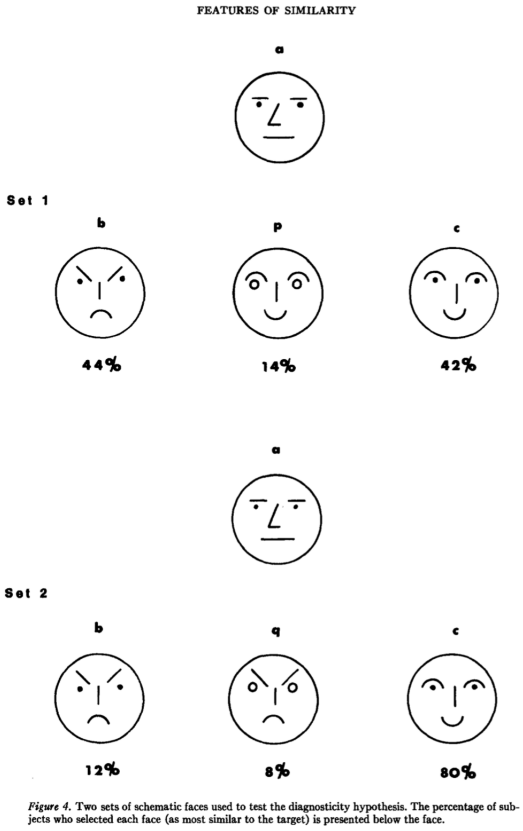
From the viewpoint of the physicist, the machine is cappable to accelerate the process of finding physical relations inside data. However, outcomes of the mining work are either implicit knowledge of the machine or interpretable and transferable to human. In here, I show that the interpretable machine learning or explainable artificial intelligence takes more advantages than the other in a number of sophisticated ways.

Interpret means to explain or to present in understandable terms. In the context of machine learning systems, we define interpretability as the ability to explain or to present in understandable terms **to human**. The need for interpretability stems from an *incompletenes* in the problem formalization, creating a fundamental barrier to optimization and evaluation Doshi-Velez and Kim, 2017. Beside scientific understanding, safety, ethics, mismatched objectives, and multi-objective trade-offs are all serious reasons people need interpretability system. For example, for complex tasks, the end-to-end system is almost never completely testable and it is in computationally and logically infeasible. In these cases, for safety purpose, the system should be represented in well understand and controllable terms by the human. Five general

cognitive chunks are defined in Doshi-Velez and Kim, 2017 as the basic units of explanation. These *chunks* are (1) Form of cognitive chunks, (2) Number of cognitive chunks, (3) Level of compositionality, (4) Monotonicity and interactions between chunks and (5) Uncertainty and stochasticity. The five definitions cover almost all necessary questions for the explanation, e.g what are the basic units of explanation? or how much similar between the features in prototype models and the actual one. There is plenty of necessity works people need to do for making an interpretable machine system. In the following, we show that conventional similarity modeling method in the machine learning mismatch with the human conventional though about similarity.

To model similarity in Machine learning community, one of the biggest challenges is to select an appropriate operator to deal with the variety of data types. Interval-scaled variables, binary variables, categorical variables etc are all need appropriated similarity measurement operator to determine. The most common method is to use mathematical metric. Some of canonical metric forms are Minkowski distance Sung-Hyuk, 2007 (including Manhattan and Euclidean distance), Chord distance Gan G, 2007, Mahalanobis distance BF, 2007, Cosine distance Sung-Hyuk, 2007. These distance measures imply that if distance between two data instances A and B is smaller than A and C then A is similar to B than C. Beside canonical similarity measurement basing on distance forms, a numerous methods have been developed to make the form of their similarity values closer to human interpretation. Joshua B. Tenenbaum with his publications on similarity topic is one of prominent follow the research field of connecting these two phenomena. His papers including Learning the structure of similarity Tenenbaum, 1996, Rules and Similarity in Concept Learning Tenenbaum, 2000; Generalization, similarity, and Bayesian inference Tenenbaum and Griffiths, 2001 make a flow of thought which is consistent and inherited from Tversky's theory Tversky, 1977. Otherwise, Michael I. Jordan and Francis R. Bach Bach and Jordan, 2004 show a widely used model, spectral clustering with the central idea by projecting data instances into new compact dimensions that maximize the similarity inside each partitioning groups. Other well known distance-similarity learning methods should be mentioned are: learning Mahalanobis distance by optimizing Gaussian distribution constrain Davis et al., 2007, similarity estimation by using hash functions Dahlgaard, Knudsen, and Thorup, 2017, semantic central similarity measurement Deudon, 2018 etc.

The similarity meaning, in conventional though by human, especially physicist is totally mismatch with the way of similarity modeling by machine learning experts. According to Amos Nathan Tversky, one of the most distinguish researcher in mathematical psychologist and cognitive science, there is a number of corollary in implying canonical distance metric that contradict to the similarity meaning in human cognition. In his paper, features of similarity in 1977 Tversky, 1977, the minimality axiom of distance metric, $\delta(a,b) \leq \delta(a,a) = 0$ is violate if the identification probability is interpreted as a measure of similarity. Secondly, the symmetry axiom of distance $\delta(a,b) = \delta(b,a)$ is also violated by the similarity judgments of human. We say "the son resembles the father" rather than "the father resembles the son", or "the portrait resembles the person" rather than "the person resembles the portrait." The last axiom about triangle inequality, according to Tversky is also violated or at least, cannot be formulated as ordinal terms as in distance measurement. He stated that: Consider the similarity between countries: Jamaica looks closely similar to Cuba basing on geographical aspect and Cuba looks closely identical to Russia basing on the common viewpoint in political aspect. However, Jamaica and Russia looks like having not much things in common at all.



(A)

FIGURE 2.1: Context dependence of similarity measure, according to Amos Tversky, "Feature of similarity" 1977 Tversky, 1977.

From my viewpoint, the most critical point of similarity from human cognition and also, material scientist is shown in Tversky's work in section Similarity in Context. According to him, changes in context or frame of reference correspond to changes in the measure of the feature space. To better illustrate the idea, there was a test done by Tversky about which face is considered as similar as face "a", results summary in Figure 2.1,. In giving the set 1 with three faces (b) – angry, (p) – smile and (c) smile, the most similar face according to human taking the test is angry one, or the face (a) with 44% voters. By replacing only the face (p) – smile by a face (q) – angry, the result is reversed totally with 80% people believe that the smile one, the face (c) is the most similar to (a). That is an essence example for the concept of diagnosticity hypothesis or later, we called context based similarity measurement. His similarity measurement, Tversky's index was first announced in this paper.

Last but not least, one of the most essential points for all similarity measures is the objective aspect. Any similarity measures are designed under the subjective viewpoint of designers. Since the designers choose the subjective viewpoint by selecting the logical aspect of any given mathematical formula (e.g Euclidean distance, Hamming distance metric). Unfortunately, the logical aspects are not frequently matched with the natural behavior of such objects (now it is the objective aspect). On the other side, the similarity measures designed by domain expert researchers frequently match with the natural behavior of objects. However, there are subjective viewpoints of these designed cases. The first subjective point due to the limitation of natural behavior observations for the researchers. The second subjective point due

to the concept gap in choosing modeling formula that mimics the observations.

To conclude, the gap between mathematical form of distance to the similarity meaning should be filled firstly if human target to knowledge co-creation with the machine.The bridging concept of similarity and corresponding models that used in materials science are shown in detail in 3, 4 and Chapter B.

2.3.1 Similarity in other domains

In the previous section, the mentioned similarity methods e.g canonical Minkowski metrics are objective to any data science problems. These methods are independent to the designed predicting variable, target variable or any relation among these variables. However, in applying to domain problems as drug discovery, natural language processing, a number of similarity measurement was developed for better serving to concerned target.

In drug discovery, the most common similarity measurement was conducted over SMILES (Simplified Molecular Input Line Entry Specification) representation with Tanimoto score Tanimoto, 1957. The Tanimoto score was initially designed focuses on chemical structures of the drug rather than other drug's properties. However, since the main target of designing drugs focuses on the treatment of diseases, other similarity measurement models were developed for building more sophisticated drug-disease network Luo et al., 2016. Other highlight researches in similarity measure in drug discovery are all designed under specific purpose rather than an objective measurement only Lo et al., 2018.

In bioinformatics domain, researchers always need to deal with various segments of genes. People believe that segments has its common expressions under a given reaction context will associate with relations in functional viewpoint **Golan2006**. From this common sense in the bioinformatic community, a number of similarity measure algorithms were developed for grouping that genes with similar expression profiles **Bammer2000**; **Yoo2003**; **Lopez2003**. Others have studied the association between different expression profiles and different cellular conditions. Such associations can help in developing assays that are designed to detect different types of cancers based on the expression patterns of genes **Yeatman2003**.

2.4 The object of research

In the previous section, I pointed about the main purpose of materials science and how possible works in machine learning could serve for the purpose. In addition, the need for explainable artificial intelligence and especially in the materials science field. An interpretable machine learning model not only reduces underspecification cases but also giving more chances to human making analogy reasoning on what knowledge the machine pointing out. The knowledge, from the psychologist, are all produced from analogical thinking. For these reasons, the main goal of my thesis is **to develop machine learning methods to find similarity patterns, defined by referring to a given physical property of materials.**

There are two main contributions from my thesis. From viewpoint of physics researcher community, the first contribution is to propose a framework for modeling an abstract domain knowledge, "the physical similarity" to the machine. The framework contains seven main steps: (1) define the contexts, (2) generate the contexts, (3) collect meaningful contexts, (4) voting system, (5) evaluation, (6) utilizing the similarity information and (7) accelerate the serendipity. The details are shown

in the previous section. From viewpoint of machine learning expert community, the second contribution is shown through partitioning methods in step (2) – linear regression based clustering; three voting system in step (4); pointing out four approaches to utilize the similarity information in step (6). The methods are evaluated by applying to a number of test beds and all published to the community through research articles shown in Chapter 5.

Chapter 3

Relation function – the center of similarity concept

3.1 Introduction

Human conclusions about the similarity between objects are supported by taking its common behaviors or underlying driving mechanisms as supporting evidence R. Hofstadter, 2006; Döbereiner, 1829; Gentner, 2002; Gavetti G, 2005. Taking this viewpoint as the central meaning of similarity, this terminology could be used to interpret the essence of various methods in machine learning. For example, anomaly detection methods focus on identifying the most dissimilar data points; a mixture of regression methods target the grouping of the most similar data poin ts from the viewpoint of regression functions. The work of unifying these methods through the concept of human similarity could be a part of the realization the explainable artificial intelligence Rudin, 2019; Letham et al., 2015.

To break down above the similarity concept, we first define the similarity between two data points considering reference functions passing through these points. Figure 3.1 illustrates and compares similarity measurements using canonical methods (left) and our similarity measurement considering reference functions (right). In the figure, there are three data points A, B, and C with a conventional distance between A and B measured in the descriptive space $\boldsymbol{x}$, where $d(A, B)$ is much smaller than $d(A, C)$. If only the distance of these three points is considered, the similarity between A and B, s_{AB} is larger than s_{AC}. However, in Figure 3.1 (right), reference functions f_1 and f_2 are established considering the appearance of other data points. From the functional viewpoints, A and C belong to the same f_1 or they are similar, but dissimilar to B, which is located in f_2. Comparing the similarity order, s_{AB} is smaller than s_{AC}.

In considering to the subjective and objective viewpoint of similarity measure, the main criteria in measuring similarity in materials science basing on the natural property of materials. In other words, the similarity state of any pair of materials even determined and make judgments by human but observed only from its behaviors itself. To reduce the subjective aspects in selecting observations by designers, designed similarity measurement methods we show in the following are all included a multiple random selection step. Lately, the last attempt in this research to reducing the subjectivity point is using theory of evidence rather than using heuristic designed methods.

In the next following section, different forms of the abstract notation function f e.g linear function, non-linear function are used to investigate and realize the similarity concept. This chapter first introduces the three most commonly used data set in the whole thesis, which takes three physical properties as target variables: formation energy, the lattice constant, and Curie temperature. The next section shows

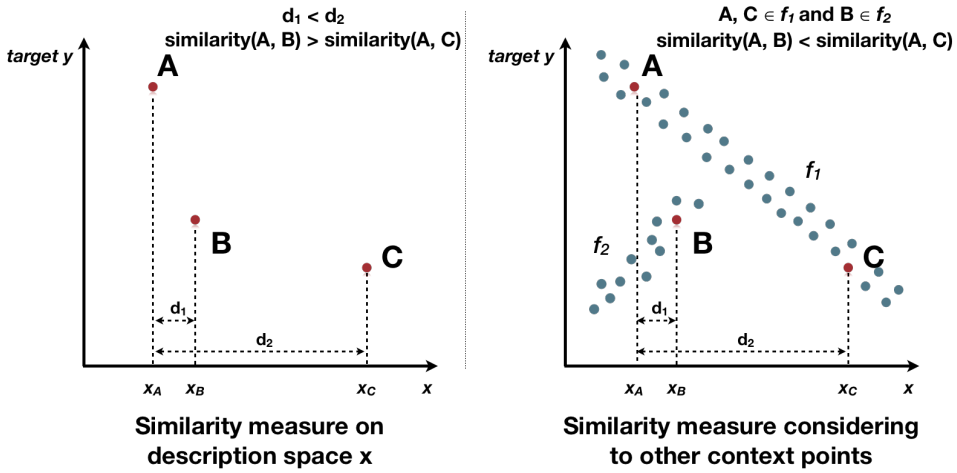


FIGURE 3.1: Left: two data points A and B are much closer, or equivalent, more similar than A and C in the descriptive space. Right: similarities among the three data points change considering the appearance of other data points

a brief introduction about linear and non-linear regression function. These functions later are used as the center of similarity–dissimilarity concepts. The variable selection and evaluation play in an important role in understanding the behavior of non-linear regression shown in the fourth section. The fifth section is a developed model called linear regression-based clustering and related problem. Lastly, the final section shows related work and the bagging method in using non-linear regression as partitioning data space tool.

3.2 Data set

In Chapter 2.2.2, we discuss a various types of information in Materials science including atomic information, structure information, experimental information etc,. In Chapter B, we discuss in detail about problem setting with imaging data set. In this section, I describe three data sets which is to show the common data format of Materials science. These data sets are all well described with materials associated with its representation and calculated or experimental target property.

3.2.1 Notation

In the thesis, we denote a dataset by $\mathcal{D}$ of p data instances. Assume that an instance with index i is described by an n-dimensional descriptive variable vector, $\boldsymbol{x}_i = \left(x_i^1, x_i^2, \ldots, x_i^n\right) \in \mathbb{R}^n$. The dataset $\mathcal{D} = \left\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \ldots (\boldsymbol{x}_p, y_p)\right\}$ is then represented using a $(p \times (n+1))$ matrix. The target variable values of all data instances in the dataset are stored as a p-dimensional target vector $\boldsymbol{y} = \left(y_1, y_2 \ldots y_p\right) \in \mathbb{R}^p$.

3.2.2 Predicting formation energy of $Fm\bar{3}m$ AB materials data

The first data set contains 239 binary AB materials, which are collected from the Materials Project database Jain et al., 2013. All of atoms labeled by A are all metallic forms: alkali, alkaline earth, transition, and post-transition metals, and including lanthanide families. All of atoms labeled by B, by contrast, are mostly metalloids

TABLE 3.1: Designed predicting variables show fundamental physical characteristics of component elements and structure-properties of compounds in E_{form} predicting problem for . The A and B elements compose the AB materials with binary cubic structure identical to that of the $Fm\bar{3}m$ symmetry group.

Category	Predicting variables
Atomic properties of A element	Z_A, r_{ionA}, r_A, IP_A, χ_A, n_{eA}, T_{bA}, T_{mA}
Atomic properties of B element	Z_B, r_{ionB}, r_B, IP_B, χ_B, n_{eB}, T_{bB}, T_{mB}
Structural information	V_{cell}

and non-metallic form. The computed formation energy E_{form} of AB material as the physical target property. In this problem setting, we limited the collected materials to the same cubic structure, $Fm\bar{3}m$ symmetry group (one of the most well known compound exhibits this structure is Sodium Chloride).

To represent predicting variables for all compound, we used seventeen variables which categorize into three types, as summarized in Table 3.1. The first and second categories pertained to the predicting variables of the atomic properties of the element A and element B components; these included eight numerical predicting variables: (1) atomic number (Z_A, Z_B); (2) atomic radius (r_A, r_B); (3) average ionic radius (r_{ionA}, r_{ionB}), (4) ionization potential (IP_A, IP_B); (5) electronegativity (χ_A, χ_B); (6) number of electrons in outer shell (n_{eA}, n_{eB}); (7) boiling temperature (T_{bA}, T_{bB}); and (8) melting temperature (T_{mA}, T_{mB}) of the corresponding single substances. The boiling and melting temperatures are all collected under standard conditions (0°C, 10^5 Pa).

Information of crystal structure shows valuable meaning to any the physical characteristic of materials. For this reason, we build the last category with structure of predicting variables. These variables are estimated from the crystal structures of the materials. In this dataset, basing on the similar of all materials in the dataset, we use only the unit cell volume (V_{cell}). This variable only represent for the structural predicting variable. The target variable in this experiment is E_{form}.

3.2.3 Predicting lattice parameter of body-centered cubic material data

The second data set contains 1541 binary AB body-centered cubic (BCC) crystals with a 1:1 element ratio from Ref.Takahashi et al., 2017. The data set associates with computed lattice constant value L_{const} of the crystals. The A elements corresponded to almost all transition metals Al, As, Au, Co, Cr, Cu, Fe, Ga, Li, Mg, Na, Ni, Os, Pd, Pt, Rh, Ru, Si, Ti, V, W, and Zn and the B elements corresponded to those with atomic numbers in the ranges of 1–42, 44–57, and 72–83. The collected data set includes materials which is non-existed in reality, e.g the binary compound of AgHe, which is one of the constituent He, is a noble gase element and well known for unlikely to form a solid.

TABLE 3.2: Designed predicting variables show fundamental physical characteristics of component and structural characteristic of materials in the lattice parameter prediction problem. A and B are elements of the binary AB *BCC* materials.

Category	Predicting variables
Atomic characteristics of metals A	r_{covA}, m_A, Z_A, n_{eA}, ℓ_A, χ_A, ρ_A
Atomic characteristics of metals B	r_{covB}, m_B, Z_B, n_{eB}, ℓ_B, χ_B, ρ_B
Structural & additional information	ρ, d_χ, sum_{AD}

To represent predicting variables of each compound, we used seventeen variables which categorize into three types, related to fundamental physical characteristics of the A and B construction elements. Details of the designed predicting variable are summarized in Table 3.2. The concerning physical characteristics are: the (1) atomic radius (r_A, r_B); (2) mass (m_A, m_B); (3) atomic number (Z_A, Z_B); (4) number of electrons in outermost shell (n_{eA}, n_{eB}); (5) atomic orbital (ℓ_A, ℓ_B); and (6) electronegativity (χ_A, χ_B). The atomic orbital values are used in this work under categorical variables with elements: s, p, d, f respect to numerical values of orbitals, *i.e.* 0, 1, 2, 3, respectively. For integrating the structure information, we add four more properties : (7) the density of atoms per unit volume (ρ_A, ρ_B); (8) the unit cell density ρ; (9) the difference in electronegativity d_χ; and (10) the sum of the atomic orbital B and difference of electronegativity sum_{AD} (see Ref.Takahashi et al., 2017).

3.2.4 Predicting experimental observed Curie temperature of rare-earth–transition metal alloys

The third data set contains 101 binary alloys which is combination of one transition metal element and one rare-earth metal element. The data set was collected from the NIMS AtomWork database Villars et al., 2004; Xu, Yamazaki, and Villarsß, 2011, which included the detail information of all structures alloys and its experimental Curie temperatures T_c.

For representing the structure properties and physical properties of all these binary alloys, we designed twenty one predicting variables in three categories. All variables are described in the Table 3.3. All variables in the first category and the second category all represents the atomic properties of the transition metal elements and rare-earth elements, respectively. The designed physical pproperties are listed as follows: (1) atomic number (Z_R, Z_T); (2) covalent radius (r_{covR}, r_{covT}); (3) first ionization (IP_R, IP_T); and (4) electronegativity (χ_R, χ_T). Moreover, predicting variables related to the magnetic properties include: the (5) total spin quantum number (S_{3d}, S_{4f}); (6) total orbital angular momentum quantum number (L_{3d}, L_{4f}); and (7) total angular momentum (J_{3d}, J_{4f}). For R metallic elements, additional variables $J_{4f}g_j$ and $J_{4f}\left(1-g_j\right)$ are added, due to the strong spin-orbit coupling effect.

The third category variable was chosen which contained values calculated from the crystal structures of the alloys reported in the AtomWork database Villars et al., 2004; Xu, Yamazaki, and Villarsß, 2011. The designed predicting variables included

TABLE 3.3: Designed predicting variables describing fundamental characteristics of component elements and structural characteristics. The Curie temperature, T_c is set as target variable in predicting of the rare-earth–transition metal alloys.

Category	Predicting variables
Atomic characteristics of transition metals	Z_T, r_{covA}, IP_T, χ_T, S_{3d}, L_{3d}, J_{3d}
Atomic characteristics of rare-earth metals	Z_R, r_{covR}, IP_R, χ_R, S_{4f}, L_{4f}, J_{4f}, $J_{4f}g_j$, $J_{4f}\left(1-g_j\right)$
Structural information	C_T, C_R, r_{TT}, r_{TR}, r_{RR}

the transition (C_T) and rare-earth (C_R) metal concentrations. It should be noted that, by using the atomic percentage for the concentration, there is a correlation between these two properties. For this reason, we use the concentrations in units of atoms/Å^3. This unit is more informative than the atomic percentage since it contains information on the constituent atomic size. As a consequence, (C_T) and (C_R) are not completely dependent. Other additional structure variables are also added: the mean radius of the unit cell between two rare-earth-elements r_{RR}, between two transition-metal-elements r_{TT}, and between transition and rare-earth-elements r_{TR}. The experimentally observed T_c is used as the target variable.

3.2.5 Appropriated physical phenomena

In this section, we discuss the main underlying reason why we apply our similarity measurement methods developed. As the main problems which interest us in this thesis are the similarity / dissimilarity between the material objects, for example, the binary, quaternary, ternary compounds, the alloys, etc. pairs or groups of objects that are similar and different from others. There are two underlying assumptions: (1) all hidden phenomena are observable within the confines of data collection / representation, data selection and verification (2), the mixture physical phenomena could be identifiable under the limit of functional representation and data representation.

Regarding the first hypothesis, our developed method could not work for missing information cases or the data representation case could not capture information about the phenomena. These cases are called under-specification cases. In this situation, we need experts or domain knowledge to better guide the representation of data from the first step.

Regarding the second hypothesis, we must assume that the physical phenomena involved could be detected separately with our design problems. Without this assumption, the similarity / dissimilarity results we obtained are all identical. Therefore, we could not compare the results together to reject or reject our original hypothesis; or revise the hypothesis by selecting other methods for measuring similarity.

3.3 Regression function

The terminology "relation function" between materials is complicated. In Physics, a physical law function represents the relation between description properties of objects to a given target property. For example, the Coulomb force, $F = k_e \frac{q_1 q_2}{r^2}$ represents repulse – attract force between two charge particles $q_1, q2$ with r distance between them. By increasing the distance r twice times, the magnitude of the force

decrease four times. The relation described by the deterministic equation, or deductive reasoning as the Coulomb force is considered as relation function by the physicists.

The relation function in machine learning could be represented in various forms. The most common way to categorize is by parametric and non-parametric models. The parametric model, e.g linear regression, linear support vector machine, neural network, etc, assume the data instances follows a predefined formula of the relation among descriptive variable and the target variable. On the other hand, the non-parametric model, e.g k-nearest neighbor, kernel ridge regression function assumes the function's value at any point in representation space by a parametric function to other points. In the thesis, I investigate the type functions: linear regression (parametric model) and kernel ridge regression function (non-parametric model).

3.3.1 Linear regression

A linear regression function $\hat{f}$ represents the relation between descriptive variable $\boldsymbol{x}$ to a given target variable $\boldsymbol{y}$ by a linear relation form as follows:

$$\boldsymbol{y} = f(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{\beta} + c + \boldsymbol{\epsilon} \tag{3.1}$$

with $\boldsymbol{\beta}$ is a p dimensional parameter vector and c is a intercept, $\boldsymbol{\epsilon}$ represents as error term. The parameter vector $\boldsymbol{\beta}$ are determined by minimizing

$$\sum_{i=1}^{p} [f(\boldsymbol{x_i}) - y_i]^2 + RegTerm \tag{3.2}$$

with *RegTerm* represents for regularization term which differ from various regreession model: Lasso $\lambda \sum_{i=1}^{p} |c_i|$, ridge regression $\lambda \sum_{i=1}^{p} |c_i|_2^2$, elastic net with balancing the Lasso and ridge regression, etc. The *RegTerm* is used not only to reduce over fitting but also select variables which is non-redundant. Detail formulation and interpretation are shown in Murphy, 2012a.

3.3.2 Kernel-ridge regression

To learn a regression function $\hat{f}$ for predicting the target variable, we utilize the kernel ridge regression (GKR) technic Murphy, 2012a, which has been recently used with a lot of successful in materials science studies Matthias, 2015; Botu and Ramprasad, 2014; Pilania et al., 2013.

Kernel regression or specifically, kernel-ridge regression is one of non-parametric technic in statistics for estimating the conditional expectation of a random variable. The general objective of the kernel-ridge technique is for finding a non-linear relation between a pair of random variables which generally called descriptive/predicting variable $\boldsymbol{x}$ and a target variable $\boldsymbol{y}$. In non-parametric models, the relation between these variables is modeled by the conditional expectation of $\boldsymbol{y}$ relative to $\boldsymbol{x}$ as follows:

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}) = f(\boldsymbol{x}) \tag{3.3}$$

with an unknown function f. The very first and most general estimation of f is proposed by Nadaraya, 1964 and Watson, 1964, by locally weighted average by a *kernel* function. There are number of canonical kernel functions as Gaussian, Laplacian, Cosine, Linear etc as follows:

Cosine kernel:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|} \tag{3.4}$$

Linear kernel:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i \cdot \boldsymbol{x}_j + c \tag{3.5}$$

Polynomial kernel:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\alpha \boldsymbol{x}_i \cdot \boldsymbol{x}_j + c\right)^d \tag{3.6}$$

Gaussian kernel:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \tag{3.7}$$

Laplacian kernel:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sigma}\right) \tag{3.8}$$

where $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| = \sum_{a=1}^{m} |x_i^a - x_j^a|$; $\boldsymbol{x}_i \cdot \boldsymbol{x}_j = \sum_{a=1}^{m} x_i^a x_j^a$, and m is the number of dimensions; σ, α are the variance of the Gaussian/Laplacian and the hyper parameter of polynomial kernel functions, respectively; d is the polynomial order of the polynomial kernel—we set $(d = 3)$ in all our experiments; c is a constant.

For a given new data instance $\boldsymbol{x}_*$, the predicted value $\hat{f}(\boldsymbol{x}_*)$ is expressed by taking summation of all weighted kernel functions as follows:

$$\hat{f}(\boldsymbol{x}_*) = \sum_{i=1}^{N} c_i k(\boldsymbol{x}_*, \boldsymbol{x}_i) \tag{3.9}$$

where N is the number of training materials. The coefficients c_i which represent for the weight of corresponding materials $\boldsymbol{x}_i$ are determined by minimizing

$$\sum_{i=1}^{N} [\hat{f}(\boldsymbol{x}_i) - \boldsymbol{y}_i]^2 + \lambda \sum_{i=1}^{N} ||c_i||_2^2. \tag{3.10}$$

Both regularization parameter λ and hyper parameter σ are generally selected by using cross-validation Stone, 1974; Picard and Cook, 1984 evaluation method. The cross-validation method excludes a number (e.g 10%) of the data instances during the training process, and maximizing the prediction accuracy for these excluded data instances. The prediction accuracy is evaluated by different scores: coefficient of determination R^2 score and mean absolute error MAE, shown in Section 3.3.3. We consider the component $c_i k(\boldsymbol{x}_*, \boldsymbol{x}_i)$ in Eq. 3.9 as the contribution of the training material $\boldsymbol{x}_i$ to the prediction model $\hat{f}$. In section 4.3 we will see back the utilization of the contribution term, $c_i k(\boldsymbol{x}_*, \boldsymbol{x}_i)$, in investigating dissimilarity effect among data instances.

3.3.3 Evaluation

Coefficient of determination R^2 is the very common and widely accepted to evaluate the prediction ability of prediction models in Machine learning. In regression, the R^2 coefficient of determination, or adjusted version of R^2 score Kvalseth, 1985, measures of how well the regression predictions approximate the distribution of real data instances. In the canonical form by, the score evaluates the relation between the sum of squares of residuals $\sum_{j=1}^{p_{vld}} \left[f(\boldsymbol{x}_j) - y_j\right]^2$ and the total sum of squares $\sum_{j=1}^{p_{vld}} \left[\bar{y} - y_j\right]^2$ as follows:

$$R^2 = 1 - \frac{\sum_{j=1}^{p_{vld}} \left[f(\boldsymbol{x}_j) - y_j\right]^2}{\sum_{j=1}^{p_{vld}} \left[\bar{y} - y_j\right]^2} \tag{3.11}$$

Here, p_{vld} is the number of validation points, $f(\boldsymbol{x}_j)$ is the predicted value by a learning function f at point $\boldsymbol{x}_j$. y_j is the observed value of data instance with index j in the validation set and $\bar{y}$ is the average of the validation set used to compare the values predicted for the excluded data instances with the known observed values.

An value of R^2 score, e.g $R^2 = 0.8$ frequently is interpreted as: "Eighty percent of the variance in the response variable could be interpreted by the predicting/descriptive variables. The remaining twenty percent could represent for unknown variability." An R^2 that reach the maximum value at 1.0 will indicates that the regression predictions perfectly fit the data. An R^2 is smaller than 0.8 could be seem as the regression predictions could not capture the nature between predicting variable combination to the target variable. Values of R^2 that reach out of the range from 0 to 1 might occur if the model fits the data worse than a horizontal hyperplane.

Mean absolute error (MAE) is a measure of difference between two continuous variables, in the most common case, are predicted and observed values of the target variable. In all thesis, MAE is defined as follows:

$$MAE = \frac{\sum_{j=1}^{p_{vld}} |f(\boldsymbol{x}_j) - y_j|}{p_{vld}} \tag{3.12}$$

with p_{vld} is the number of validation points, $f(\boldsymbol{x}_j)$ is the predicted value by a learning function f at point $\boldsymbol{x}_j$, y_j is the observed value of data instance with index j in the validation set.

3.4 Kernel regression-based variable evaluation

There are two most important objects in any data science problem, instances and descriptive variable/feature. In almost all problems, people need to design data instances and design the descriptive variable. However, there is always a subset of the designed variables that serve well for a certain modeling method. In this chapter, we focus to regression modeling methods targeted to a given property. Therefore, the need of understanding the role of descriptive variable or each variable combination is a critical questions. To develop a better understanding of the processes that generated the data, we choose the most completeness method as exhaustive searching and evaluating all possible variable combinations Kohavi and John, 1997; Liu and Yu, 2005; Blum and Langley, 1997 to identify and remove irrelevant and redundant variables Duangsoithong and Windeatt, 2009; Almuallim and Dietterich, 1991; Biesiada and Duch, 2007.

3.4.1 Subset prediction ability: PA

One can easily presume that incorporating irrelevant variables into the GKR model may impair its prediction accuracy. For that reason, we denote the prediction ability $PA(S)$ of a set S by the maximum prediction accuracy that the GKR model gain by using the subset of variables s of S as follows:

$$PA(S) = \max_{\forall s \subset S} R_s^2; \;\; s_{PA} = \arg\max_{\forall s \subset S} R_s^2, \tag{3.13}$$

where R_s^2 is the coefficient of determination R^2 value Kvalseth, 1985 gained by the GKR using a variable set s as the independent predicting variable combination. s_{PA} is the subset of variable of S that yields the prediction model having the maximum prediction accuracy.

For gathering a set of concerned variable combinations which predict the target variable at relatively high level of accurate, we train the GKR models for all possible combinations of designed predicting variables. Since we do not know yet the effect of each predicting variable on the target quantity, all the numerical descriptive variables are normalized in the same manner through all analysis. In this study, R^2 coefficient determination score is used to measure of PA. For accurate PA estimation, cross-validation evaluation to the GKR Stone, 1974; Picard and Cook, 1984; Kohavi, 1995 using the data repeatedly is used. For all possible combinations, the regularization parameters are performed grid search targeted to maximize PA of the corresponding GKR models. Each predicting variable combinations contributes a perspective on the correlation between the target and the predicting variables. For that reason, an ensemble averaging Tresp, 2001; Dietterich, 2000; Zhang and Ma, 2012 technic can be applied to combine all the pre-screened regression models to improve the PA. Furthermore, the material's similarity regarding the mechanism of the chemical and physical phenomena equipped to the target quantity is investigated more predominantly if we integrate information from all possible perspectives.

3.4.2 Strongly relevant and weakly relevant features

In this section, we focus on analyzing relation among descriptive variables in the context of contributing to the prediction ability PA of predicting model under a given target variable. From the foundation established by Eq.(3.13), we evaluate the relevance Yu and Liu, 2004; Visalakshi and Radha, 2014 of a variable in prediction work of T_C using the expected reduction in the prediction ability resulting from the removal of this variable from the full set of the variables. Let D be a full set of variables, d_i a concerned variable, and $D_i = D - \{d_i\}$ the full set of variables created by removing the variable d_i. The degree of the relevance of variables can be formalized as follows:

1. Strong relevance: a variable called strongly relevant if and only if

$$PA(D) - PA(D_i) = \max_{\forall s \subset D} R_s^2 - \max_{\forall s \subset D_i} R_s^2 > 0. \tag{3.14}$$

If removing a given variable which causes a larger reduction of prediction ability, the variable should be considered as a strong variable. The degree of relevance of a strongly relevant variable can be computationally estimated by using the leave-one-out approach, i.e., by leaving out a variable in the currently considered variable set for the GKR analysis and evaluating the extent to which the prediction accuracy is impaired.

2. Weak relevance: a variable called weakly relevant if and only if

$$PA(D) - PA(D_i) = \max_{\forall s \subset D} R_s^2 - \max_{\forall s \subset D_i} R_s^2 = 0$$

and

$$\exists D_i' \subset D_i \,\text{s.t}\, PA(\{d_i, D_i'\}) - PA(D_i') > 0. \tag{3.15}$$

Eq. (3.15) shows the estimation of the degree of relevance for the weakly relevant

TABLE 3.4: Prediction accuracy of ensemble learning model with different kernel matrices

	Cosine kernel	Linear kernel	Polynomial kernel	Gaussian kernel	Laplacian kernel
PA	0.572	0.569	0.982	0.982	0.973
s_{PA}	Z_R, J_{4f}, Z_T, r_{covT}, S_{3d}, r_{TT}, C_R	r_{covR}, J_{4f}, r_{covT}, L_{3d}, r_{RR}, r_{TT}	Z_R S_{4f}, L_{3d}, J_{3d}, r_{RR}, C_T, C_R	Z_R, χ_T, J_{3d}, r_{TR}, C_T, C_R	χ_R, χ_T, $J_{4f}\left(1-g_j\right)$, Z_T, r_{covT}, IP_T, S_{3d}, L_{3d}, J_{3d}, C_R

variables cannot be carried out in a straightforward manner, as with the strongly relevant variables. In actual meaning derivation, weakly relevant variables are relevant for prediction, but they can be replaced by other variables.

3.4.3 Result

To investigate the scientific connection Dam et al., 2018 between the variables and the actuation mechanisms of the physical phenomenon of T_C in these bimetal systems, we examine whether T_C can be predicted by using the designed variables of the compounds. A screening was conducted for all possible variable combinations, $2^{21}-1 = 2,097,151$ that respectively derive the same number of prediction models. A given kernel metric formula associates with a method to measure the similarity between compounds. Therefore, five kernel metrics- cosine, linear, Gaussian, polynomial, and Laplacian are all implemented and analyzed in this section. Leave-one-out cross-validations Stone, 1974; Picard and Cook, 1984 is performed to evaluate the prediction accuracy of these models.

Prediction ability *PA*

The prediction abilities PA (equation 3.13) of the designed descriptive variables set for different kernel-type predictors are summarized in the table 4.1. The highest PA - R^2 score 0.982 is achieved by two models deriving from variable combinations $s_{PA} = \{Z_R, \chi_T, J_{3d}, r_{TR}, C_T, C_R\}$ with the Gaussian kernel and $s_{PA} = \{Z_R, S_{4f}, L_{3d}, J_{3d}, r_{RR}, C_T, C_R\}$ with the polynomial kernel. The PA value with the Laplacian kernel experiment achieves an R^2 score of 0.973, with its $s_{PA} = \{\chi_R, \chi_T, J_{4f}\left(1-g_j\right), Z_T, r_{covT}, IP_T, S_{3d}, L_{3d}, J_{3d}, C_R\}$.

It should be noted that the prediction model created by ensembling top models that yield highest R^2 score, archieves a higher prediction accuracy than PA. Figure 3.2 shows result of ensembling top 5 models that yield highest R^2 score to with R^2 score of 0.984 and MAE: 31.21 (K), higher than PA of Gaussian kernel experiment.

In this screening result, there are **892,612** variable combinations associated with the Gaussian kernel; **284,649** variable combinations associated with the polynomial kernel and **1,317,193** variable combinations associated with the Laplacian kernel that yield regression models with R^2 scores exceeding 0.90. It is noted that, even with the same kernel metric, several regression models achieved a similar excellent prediction accuracy because the designed variables are not independent variables.

On the other hand, the PA values with linear and cosine kernels were lower than 0.8, i.e., the T_C variable cannot be predicted by our designed descriptive variable set with these kernel regression models.

Finally, the PA analysis indicates that with all the designed variables, it is possible to accurately predict the values of T_C of the rare-earth transition bimetal alloy

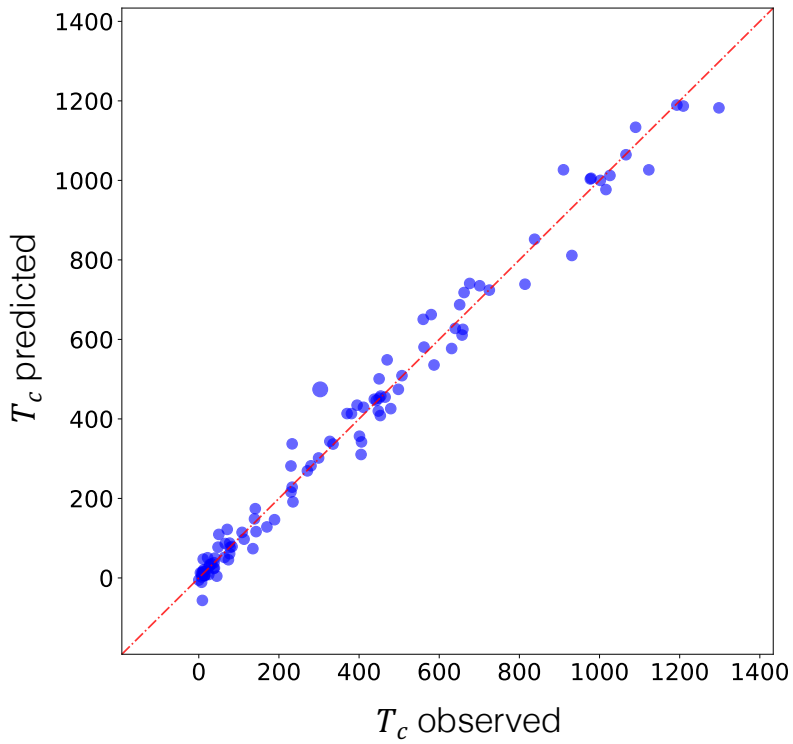
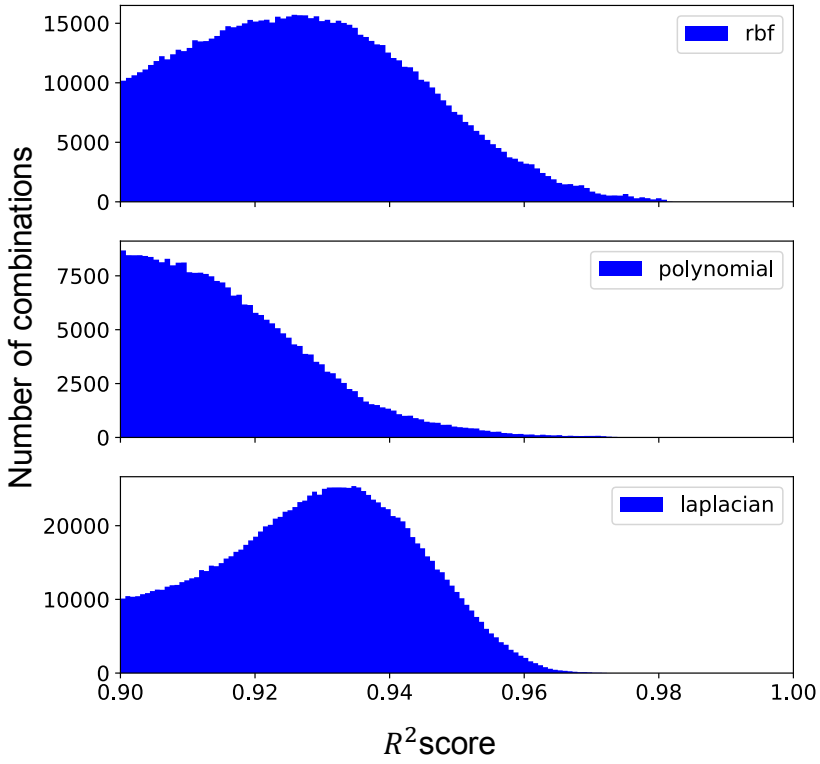


FIGURE 3.2: Observed and predicted T_C for 101 transition–rare-earth bimetal alloy compounds by an ensemble Gaussian kernel regression. The prediction model is constructed by taking ensemble averaging of top 5 models that yield highest R^2 score after kernel regression-based variable evaluation. The prediction accuracy of this model, R^2 score of 0.984, MAE: 31.21 (K), achieves a higher prediction accuracy than *PA* of all our designed variable sets.

compounds by using Gaussian, polynomial, and Laplacian GKR models with the designed variables. In the subsequent sections, we will discuss the method to improve the maximum prediction accuracy of the models as well as the physical meaning of strongly relevant variables.

Strong-weak relevant features

Figure 3.4 shows the dependence of the best prediction accuracy *PA*—red lines on the number of variables recruited in the Gaussian - polynomial - Laplacian - Sigmoid kernel regression models. In general, the prediction accuracy in all the experiments reaches the highest value with the number of descriptive variables from 6 to 8, and then gradually decreases when the number of recruited variables increases. The small subset of the designed descriptive variables and the large number of high-accuracy models described above originate from the fact that the overuse of many weakly relevant variables Duangsoithong and Windeatt, 2009; Almuallim and Dietterich, 1991; Biesiada and Duch, 2007 weakens the correlation between the similarity of the compounds, which is measured using the kernel of the variables, and the differences in their T_C values.



(A)

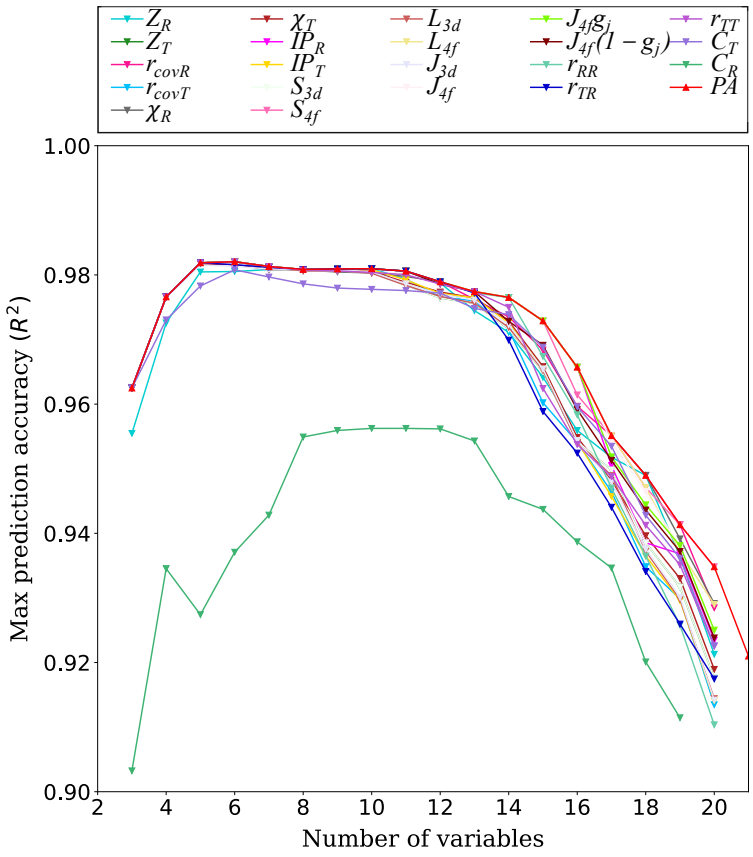
FIGURE 3.3: The distribution of R^2 score larger than 0.9 in exhaustive search of all variable combinations models with four kernels: Gaussian with 892,612 models - polynomial; 284,649 models - Laplacian; 1,317,193 models

Next, we evaluate the relevance of each variable for the prediction of T_C. We compare $PA(D)$ of the full set of variables D and $PA(D - \{d_i\})$ for all the variables d_i. We found that most of the variables are weakly relevant, and the prediction accuracy does not significantly change, except in one case—when removing the variables of the concentration of the rare-earth metal C_R. It is clearly seen in Figure 3.4 that the absence of C_R in the Gaussian and polynomial kernel model results in a dramatic decrease in the accuracy: $PA(D) < PA(D - \{C_R\})$; therefore, C_R is surely assigned as a strongly relevant variable in terms of the prediction of T_C).

This result is consistent with the understanding so far that the values of T_C of binary alloys consisting of $3d$ transition-metal and $4f$ rare-earth metals are mainly determined by the magnetic interaction in the transition-metal sublattice. In terms of molecular field theory P.Myers, 1997, we have

$$T_C \sim n_{TT} M_T^2 / 3k_B, \tag{3.16}$$

where M_T and n_{TT} are the magnetization and molecular field coefficients of the transition-metal sublattice, respectively. Both M_T and n_{TT} strongly depend on C_R. The dependencies of M_T and n_{TT} on C_R are different in compounds with different combinations of rare-earth metal (R) and transition metal (T). In Co-based and Ni-based compounds, M_T and n_{TT} tend to decrease when C_R increases. This leads to a rapid decrease in T_C with an increase in C_R. For example, in Gd–Co compounds,



(A)

FIGURE 3.4: Dependence of the best prediction accuracy on the number of variables in the Gaussian kernel regression model. The red line represents the maximum prediction ability $PA(\mathrm{D})$ of the full descriptive variable set D with respect to the different numbers of variables. The other lines represent the trend of $PA(\mathrm{D}-\{d_i\})$ by removing variable $\{d_i\}$ from D in the same manner. The significant decrease of $PA(\mathrm{D}-\{C_R\})$ shows that the concentration of the rare-earth element C_R is strongly relevant to T_C.

T_C decreases from 1,404 K to 143 K with the increase in the concentration of Gd from 0% to 75%. In contrast, in R–Fe compounds, M_T and n_{TT} tend to increase with increasing C_R. This leads to an increase in T_C with increasing C_R. Indeed, in Gd–Fe compounds, T_C rapidly increases from 0 K to 827 K with the increase in the concentration of Gd from 0% to 33%. Detailed correlation between C_R and T_C in different transition metal-based compound is shown in Figure 4.13.

Prediction of T_C for new compounds

In this section, we discuss the ability of ensemble learning using the obtained GKR models in the prediction of T_C for new compounds. The test set of new compounds includes five Fe–based compounds. The T_C of two of these compounds has been determined experimentally: $SmFe_{12} - 555(\mathrm{K})$Hirayama et al., 2017, $YFe_{12} - 483(\mathrm{K})$ Suzuki, 2017. No T_C information is available for $DyFe_{12}$, $GdFe_{12}$, $NdFe_{12}$.

From the discussion in Section 3.4.3, we know that C_R is the most strongly relevant to T_C. The C_R values of the five compounds in the test set are approximately

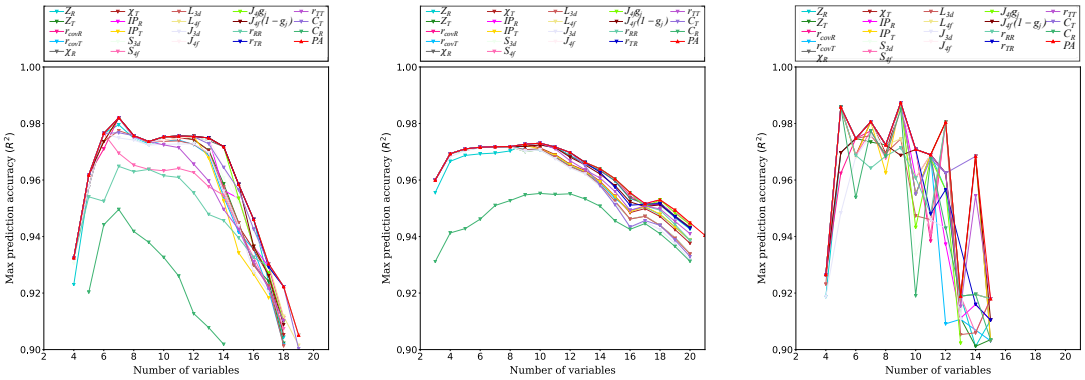


FIGURE 3.5: Dependence of the best prediction accuracy on the number of variables in the polynomial(a) - Laplacian(b) - Sigmoid (c) kernel regression model. The red line represents the maximum prediction ability $PA(\mathrm{D})$ of the full descriptive variable set D with respect to the different numbers of variables. The other lines represent the trend of $PA(\mathrm{D} - \{d_i\})$ by removing variable $\{d_i\}$ from D in the same manner. Supporting to the result from Gaussian kernel in Figure 3.4, the significant decrease of $PA(\mathrm{D} - \{C_R\})$ shows that the concentration of the rare-earth element C_R is strongly relevant to T_C.

0.006 atoms/Å^3, which is much lower than the smallest C_R value for all the other Fe–based compounds (which is 0.0075 atoms/Å^3 of Fe_{17} compounds–based) (see Figure 4.13). The only compound having a similar C_R value is $LaCo_{13}$ with C_R of 0.0055 atoms/Å^3. Further, the compounds in the test set are recently synthesized and have a crystal structure significantly different from that of all the other compounds in the data set. Therefore, the prediction for the T_C of these compounds could be seem closer to an *extrapolative* than an interpolative prediction problem.

Figure 3.7 shows the predicted value distributions for all test compounds (black line histogram) and the observed values (red dash line). It is easy to recognize that the distributions of the predicted T_C for these test compounds can be approximated by a mixture of three separate Gaussian distributions. We represent these distributions using red, blue and green, in increasing order of their mean values. From this Gaussian distribution decomposition, we can suggest that at least three distinguish functions $\hat{f}(\boldsymbol{x})$ can be regressed from the observed data to model the T_C phenomenon. Further, since the functions are learned from sub-groups of all the data set by the bagging method, we can suggest that there are at least three sub-groups of compounds corresponding to these three models.

Next, we analyze the relation of the experimentally observed T_C of the two test compounds $SmFe_{12}$ and YFe_{12} to their predicted T_C distribution. It is obvious that both the experimentally observed values correspond to the green Gaussian distribution, i.e., the largest T_C group (see Figure 3.7a, 3.7b). The mean value of these green Gaussian distributions is close to the observed value of 535.1 K, as compared with 555.0 K of $SmFe_{12}$, and 443.7 K as compared with 483.0 K of YFe_{12}. From the fact that all the five test compounds are Fe_{12}–based compounds with the same crystal structure, we can infer that the experimental values of the remaining three compounds also correspond to the green distribution. Therefore, we predict that the T_C of $DyFe_{12}$, $GdFe_{12}$, and $NdFe_{12}$ are 444.7 K, 482.7 K, and 488.5 K, respectively. There is also the potential for further experimental study on the T_C of these compounds.

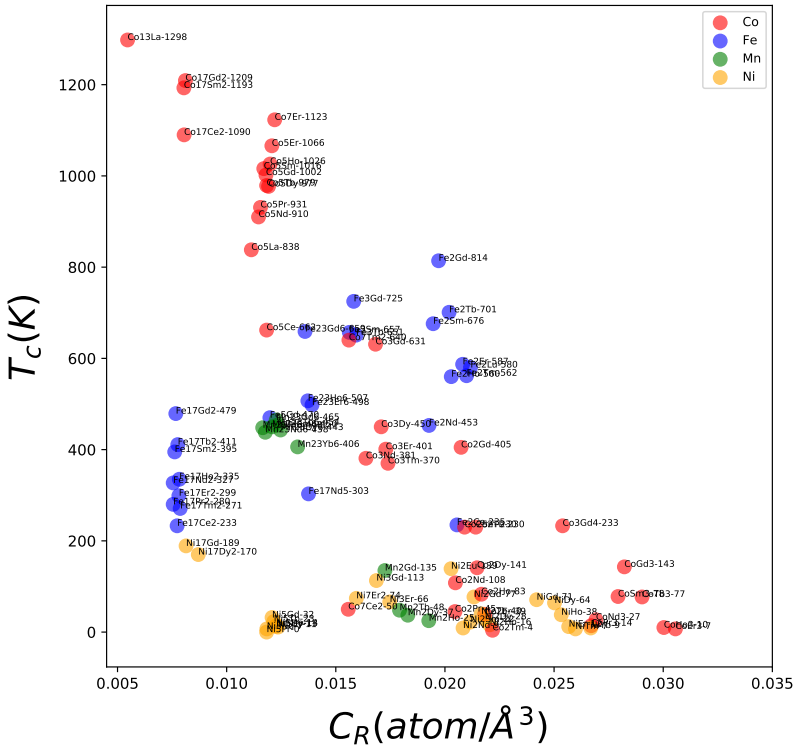


FIGURE 3.6: The dependence of T_C on the concentration of the rare-earth metal (C_R) in binary alloy compounds.

Lastly, all the above investigations show that further research is required to extract hidden functions in data as well as an alternative method of integrating predicted results from ensemble learning algorithms.

3.5 Linear regression-based clustering

In reality, a simple linear model is often contain a number of limitations to model the relationship existed in the data set. In almost cases, the data set contains non-linear relationship or the data itself can be heterogeneous and contain multiple subsets. Different subsets of data could fit best with different form linear model. However, in traditional data analysis, linear models are often preferred because of their interpretability. In the meaning covered by linear model , one might qualitatively estimate and intuitively understand how the predicting variables contribute to the target variable. Accordingly, several efforts have been devoted to develop subspace partitioning technics to decompose a high-dimensional data set into a set of disjoint small data sets, each of which might be approximated by a number of linearity subspaces by employing principal component analysis Fukunaga and Olsen, Feb. 1971; Vidal, Ma, and Sastry, Dec. 2015; J., L., and C., 2008.

In this experiment, our initial interest is the local linearity between the predicting variables and the target variable. The linearity relation reflect the nature of the underlying physical mechanism at the subspace of observations. To reach this purpose, a simple strategy by using subspace segmentation is used. The method integrate the key ideo of conventional clustering methods as well as linear regression analysis.

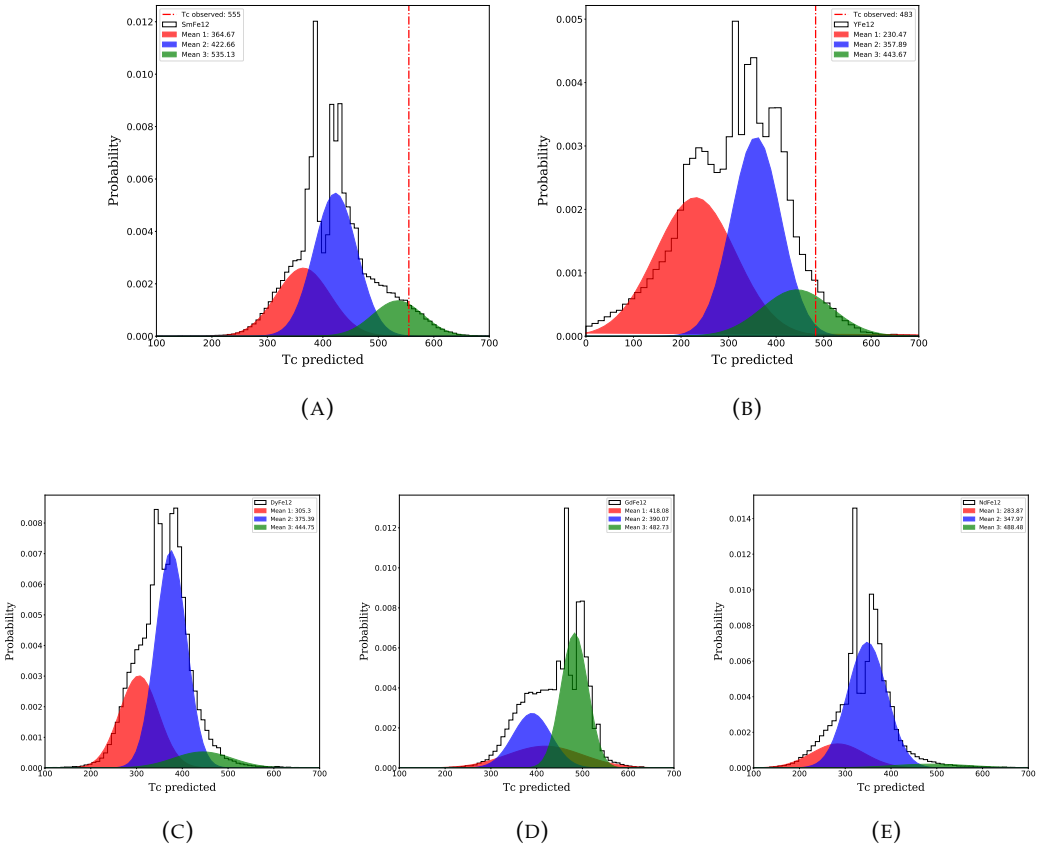


FIGURE 3.7: Predicted value distribution of test compounds by using a bagging model constructed from top–5 highest prediction accuracy kernel regression model with Gaussian and Laplacian kernels. The distributions appear as a mixture of several Gaussian distribution components. This serves as evidence to show that there are a number of functions that generate such components. The black dashed lines show the predicted value obtained by taking the average of all possible values. The red dashed lines show the observed values.

Infact, the possible sub spaces could have fewer dimensions than the whole space. Hence, we apply the sparse linear regression model that use $L1$ regularization Tibshirani, 1996 instead of the ordinary least square method.

In here, we introduce regression-based clustering method. The method base on the well-known K-means clustering associated with two major modifications. The first modification is the sparse linear regression model derived from data associated with materials in a particular cluster will be considered as the common characteristic (center of clustering model). The dissimilarities of the actual target property of each material in a group relative to the common nature relation of that group (the distance to the center). The dissimilarity values are estimated by the deviation from the observed value the corresponding linear regression model. (2) The sum of the differences of all materials in a group from the corresponding linear regression model of another group is used to measure the dissimilarity in the characteristics of that group with regard to the other group. The summation of all dissimilarities between one group to another determine in the reverse direction are used to assess the divergence between the two groups.

3.5.1 Methodology

As follows of the variable evaluation 3.4 step, we assume that a number combinations of predicting variables that yield non-linear regression models of high *PA* are collected. Under of a given selected combinations, m' numerical variables are selected from the original m numerical variables. Therefore, a material in the data set is described by an m'-dimensional predicting variable vector $\boldsymbol{x}_i' = (x_i^1, x_i^2, \ldots, x_i^{m'}) \in \mathbb{R}^{m'}$, and the data are represented using a $(p \times m')$ matrix.

Under a given data set $\mathcal{D}$ of p materials represented by m'-dimensional numerical vectors, a natural number $k \leq p$ is denoted to represent the number of clusters. Our first assumption is the existence of k linear regression models hidden in the data set. Any materials in $\mathcal{D}$ follows one of them. The purpose is to determine those k linear regression models, accordingly, to divide $\mathcal{D}$ into k non-empty disjoint clusters. Our algorithm searches for a partition of $\mathcal{D}$ into k non-empty disjoint groups $(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$ which minimize the overall sum of the residuals between the observed and predicted values (using the corresponding models) of the target variable. The problem is formulated as an optimization problem as follows.

For a given experiment with cluster number k, minimize

$$P(W, M) = \sum_{i=1}^{k} \sum_{j=1}^{p} w_{ij} \parallel y_j - y_j^{M_i} \parallel \tag{3.17}$$

subject to

$$\forall j : \textstyle\sum_{i=1}^{k} w_{ij} = 1, w_{ij} \in \{0, 1\} \tag{3.18}$$

$$1 \leq k \leq p, 1 \leq i \leq k, 1 \leq j \leq p \tag{3.19}$$

where y_j and $y_j^{M_i}$ are the observed value and the value predicted by model M_i (of k models) for the target property of the material with index j; $W = \left[w_{ij}\right]_{p \times k}$ is a partition matrix (w_{ij} takes a value of 1 if object x_j belongs to cluster $\mathcal{D}_i$ and 0 otherwise), and $M = (M_1, M_2, \ldots, M_k)$ is the set of regression models corresponding to clusters $(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$.

P can be optimized by iteratively solving two smaller problems:

- fix $M = \hat{M}$ and solve the reduced problem $P(W, M)$ to find $\hat{W}$ (re-assign data points to the cluster of the closest center);
- fix $W = \hat{W}$ and solve the reduced problem $P(W, M)$ to find $\hat{M}$ (reconstruct the linear model for each cluster).

Our regression-based clustering algorithm comprises three steps and iterates until $P(W, M)$ converges to some local minimum values:

1. The dataset is appropriately partitioned into k subsets, $1 \leq k \leq p$. Multiple linear regression analyses are independently performed with the $L1$ regularization method Tibshirani, 1996 on each subset to learn the set of potential candidates for the sparse linear regression models $M^{(0)} = \left\{M_1^{(0)}, M_2^{(0)}, \ldots, M_k^{(0)}\right\}$. This represents the initial step $t = 0$;
2. $M^{(t)}$ is retained and problem $P(W, M^{(t)})$ is solved to obtain $W^{(t)}$, by assigning data points in $\mathcal{D}$ to clusters based upon models $M_1^{(t)}, M_2^{(t)}, \ldots, M_k^{(t)}$;

3. $W^{(t)}$ is fixed and $M^{(t)}$ is generated such that $P(W, M^{(t+1)})$ is minimized. That is, new regression models are learned according to the current partition in step 2. If the convergence condition or a given termination condition is fulfilled, the result is output, and the iterations are stopped. Otherwise, t is set to $t+1$ and the algorithm returns to step 2.

3.5.2 Determine number of clusters

The number of clusters k is determined by balancing two criteria: high linearity between the predicting and target variables for all members of the group, and no model representing two different groups. The first criterion is associated with higher priority and can be quantitatively evaluated by using the Pearson correlation scores between the predicted and observed values for the target variable of the data instances in each group, by applying the corresponding linear model. The second criterion is implemented to avoid any high linearity group is further divided into two or more subgroups which might be represented by the same linear model. The determination number of k, from these intuitive concepts, can be formulated in terms of an optimization problem, as follows:

$$k = \arg\min_{k \leq p} \left[\log \frac{1 - \min_{1 \leq i \leq k} R^2_{i,i}}{\min_{1 \leq i \leq k} R^2_{i,i}} + \max_{1 \leq i \neq j \leq k} R^2_{i,j} \right] \tag{3.20}$$

where $R^2_{i,i}$ and $R^2_{i,j}$ are the Pearson correlation scores between the predicted and observed values for the target variable when we apply the linear model M_i to data instances in clusters i and j, respectively.

The first term in the function decreases regarding to the range of $\min_{1 \leq i \leq k} R^2_{i,i}$ varying from 0 to 1. Since the value $\min_{1 \leq i \leq k} R^2_{i,i}$ approaches 1 (the entire cluster exhibits almost perfect linearity between the target and predicting variables), the optimization function significantly decrease in a log scale targeted to emphasize the expected region. In contrast, the optimization function exponentially increases when $\min_{1 \leq i \leq k} R^2_{i,i}$ approaches 0, the case of any given cluster shows no linearity between the target and predicting variables. The last component of the optimized function is for avoiding overestimation of k. A group associated with high linearity further is prevented to divide into two or more sub-groups which are all represented by the same linear regression formula. In this work, the criterion for determining k is identical with the criterion to evaluate and compare a given linear regression-based clustering model to the other. Furthermore, any material associated with a number of cluster labels without explicit information of the target physical property's value. In reality, the value could be estimated by obtain from a prediction models, *e.g.* a non-linear regression model.

3.5.3 Interpreting cluster structure by decision rule

In order to establish the rule of determination of the material groups, we carry out a multi class classification analysis using a decision tree **Quinlan86**; Rokach and Maimon, 2008 regression model. The aim is to represent the dependence of the group (cluster) on the features of the materials. By utilizing the regression-based clustering method, the cluster label of the p materials are stored as a p-dimensional categorical vector $\boldsymbol{c} = (c_1, c_2, \ldots, c_p)$ where $c_i \in \{1, 2, \ldots, k\}, 1 \leq i \leq p$ and is considered as a new target vector. We learn the model to predict the new target c by using all the

original m numerical variables $(\boldsymbol{u})$ and n categorical variables $(\boldsymbol{v})$ for data representation. The description of materials in the data set $\mathcal{D}$ of p materials is subsequently represented using a $(p \times (m+n))$ matrix of both numerical and categorical values.

Categorical variables which are used in this step has several layers of meaning. Firstly, a group of numerical variables could explicitly describe a group of objects, however, it has a drawback of too explicit to interpret, especially by decision tree method. For example, to describe a group of element: B, Si, Ge, As, Sb, Te and At by two numerical variables: number of electrons in outer shell n_e and number of electron shell n_{shell}, the final decision tree is complex to acquire meaningful. However, a categorical variable *Type* could solve easier with *Type* = 'Metalloids'. A reflect of physical history reminds us the same story of categorizing complexity numerical elements in the discovery of alkali, alkaline earth, noble gas groups in periodic table; the grouping fundamental particles into fermion and bosons, lepton and quarks in particle physics, etc.

The concept of the decision tree means that the prediction model is broken down into a set of choices for each descriptor element - *i.e.* starting at the root of the tree and progressing to the leaves, where the prediction result is received. The goal is to create a model that predicts the value of a target variable by determining simple and interpretable decision rules inferred from the data features. In this study, the aim is to learn interpretable decision rules for determining the group that a material belongs to. An integration of the regression-based clustering analysis and the decision tree for multiclass classification analysis can be utilized for both the interpretation and prediction purposes.

It should be note that a material can be assigned to clusters based upon its deviation from the corresponding linear models. If there is no information on the value of a material's target quantity, the cluster label can be predicted by applying the learned decision rules or by utilizing the predicted value of the target quantity regressed by the best constructed GKR model.

3.5.4 Group index prediction for new instance

In this section, we discuss about the group index estimation for a new instance and prediction ability of the regression-based clustering (RBC) model, which inherits the variable combination set from the initial model. A new instance will associate with k predicted values which respect to outcome of k linear mixture models in our method. The problem of identifying which group the true value belongs to is raised. To deal with this, the outcome of the best performance non-linear predictive model is used as a temporal true value y_{tmp} and then the group label is estimated by using the following formula:

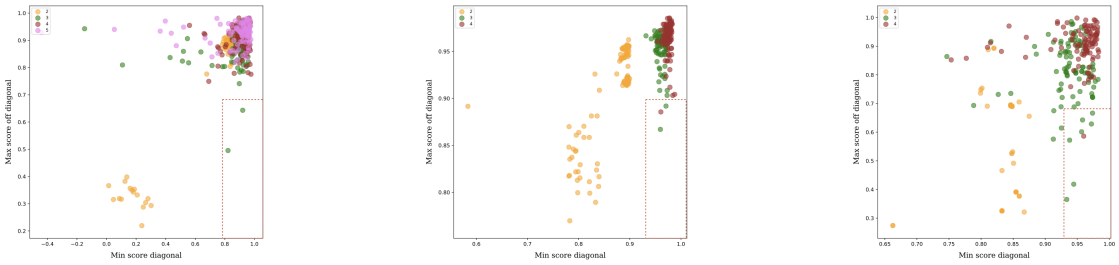
$$g_{idx} = \arg\min_{i \in \{1:k\}} |y_{tmp} - y_i| \tag{3.21}$$

This method is constructed by the clustering evaluation criteria in formula 3.20. Since any pair of linear models are minimized the correlation, the most closest value generated by those models to the value predicted by non-linear model will indicate the group of the instance belongs to.

3.5.5 Result

Determine number of clusters

Regression-based clustering method as other widely used mixture model: Gaussian mixture model, K-means ... face with a problem of traditional parameter estimation



(A) (B) (C)

FIGURE 3.8: Determination of the number of clusters and the results of regression-based clustering technic using a map of two evaluation objectives: (1) maximize prediction R^2 score for all models and (2) maximize the dissimilarity among models in evaluation of different problems: 3.8a AB compound, 3.8b lattice parameter, and 3.8c Tc magnetic phase transition temperature. The red rectangle denotes the region in which the mixed linear model shows the best prediction ability and the highest degree of dissimilarity from other models.

basing on maximum likelihood estimation (minimizing absolute error in equivalent). One possible solution is to construct and optimize the tight lower bound of the data marginal likelihood by variational methods Blei, Kucukelbir, and McAuliffe, 2017; Corduneanu and Bishop, 2001; Wang and Titterington, 2006. However, the likelihood function of a mixture model is usually multimodal let the final maximum likelihood result easy to trap into local maxima. Several methods are proposed to overcome this problem by repeatedly optimize the likelihood under different initialization strategies Christophe Biernacki, 2003; R, 2009; Fraley, 2006.

In our method, we pick the random initialization procedure and define the evaluation criteria for grouping work by two conditions: (1) maximize the prediction accuracy inside a group (high R^2 score and low MAE)and (2) minimize the prediction ability between any two separated groups.

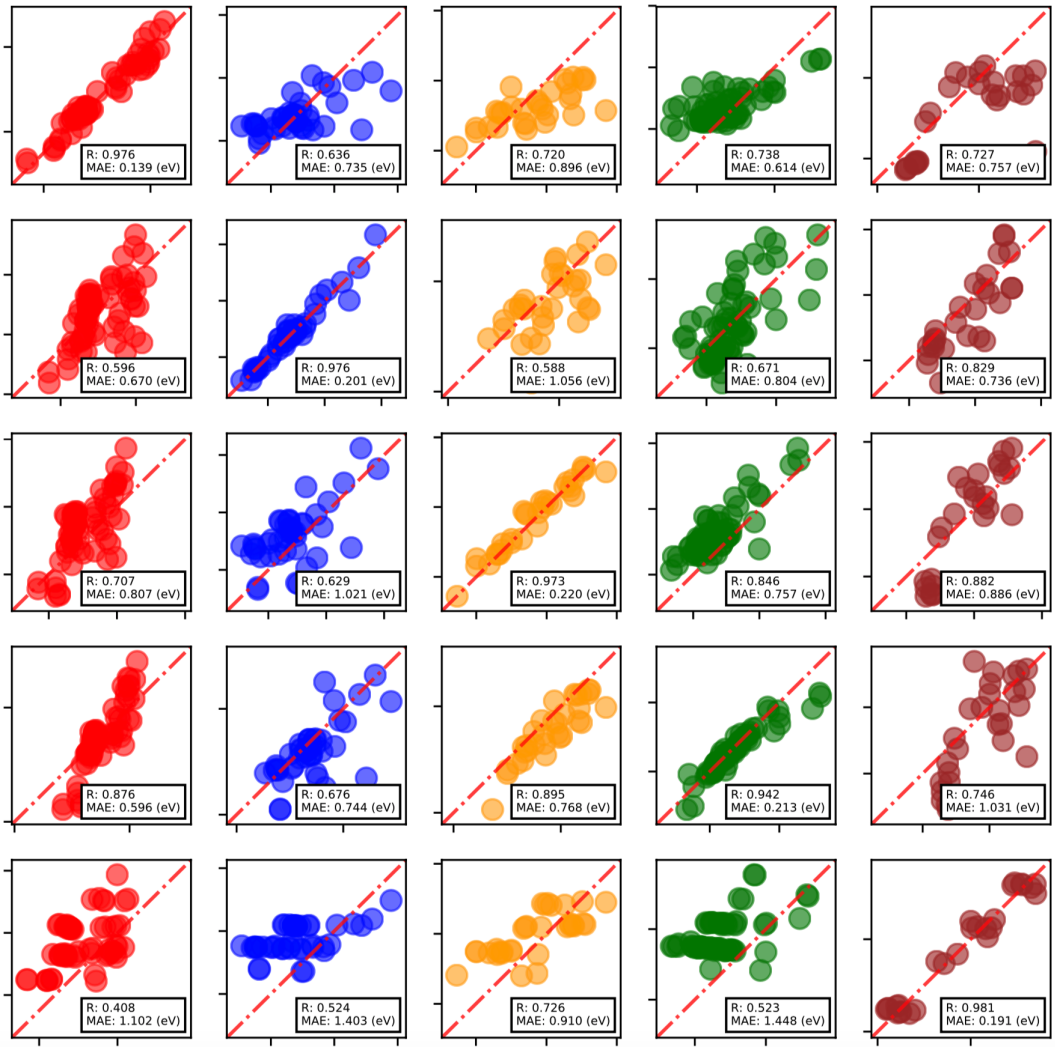
Figure 3.8 shows us the result of this clustering evaluation strategy in all three experiments shown above. The right corner in each figure shows us the high expected clustering result location.

Learning decision rule from cluster structure

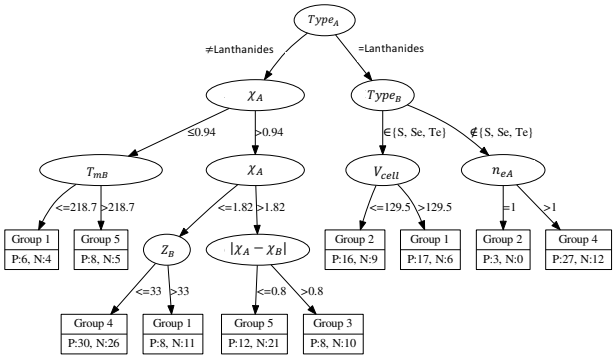
Predicting formation energy of AB compound

We carried out the regression-based clustering analysis on this data. For determining the number of clusters and the weighted average of the linearity (measured by using Pearson correlation as described above) for evaluating the performance of the clustering analysis, we found that the best set of variables are $\{V, Z_A, \chi_A, n_{eA}, IP_A, T_{mA}, Z_B, n_{eB}\}$ and the number of clusters is five. The prediction ability afforded by the constructed linear mixture models reaches 0.941 (MAE: 0.188 eV), Figure 3.9c. The confusion matrix, Figure 3.9a shows high prediction ability for individual models and high dissimilarity between the models.

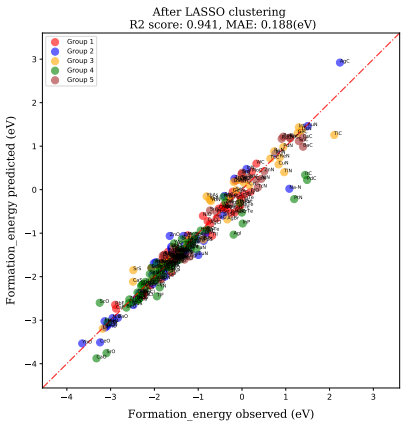
The group label obtained for each compound via the regression-based clustering analysis is subsequently set as the target variable for the decision tree analysis. Figure 3.9b shows the obtained tree, which yields precision of 72.0%. Following the tree, it is apparent that the dominant elements in group 1, 2 and 4 are compounds



(A)

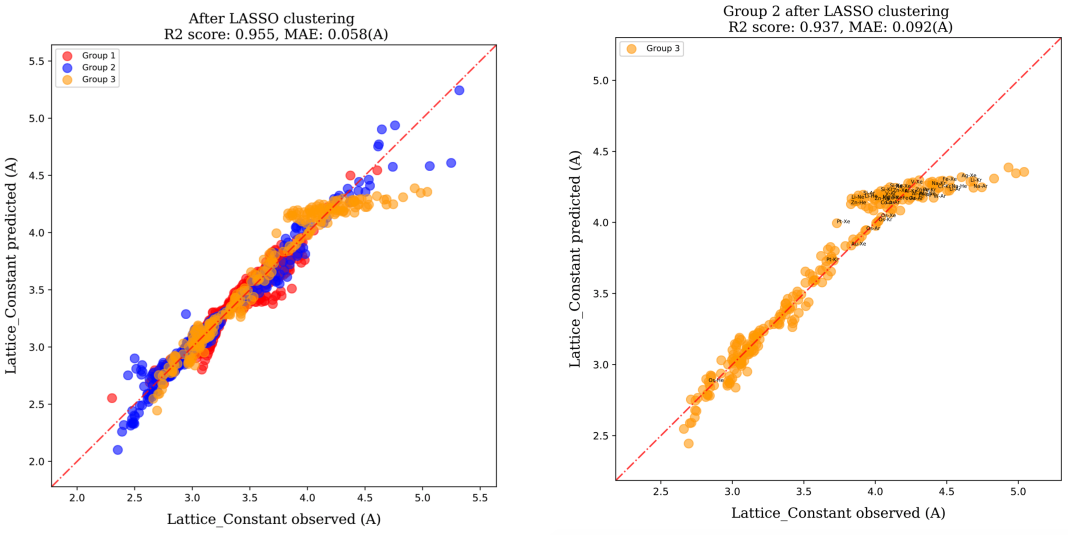


(B)



(C)

FIGURE 3.9: Binary AB compound – (A): confusion matrix describes the dissimilarity among the models employed (B): the overall prediction accuracy achieved by combining 5 clusters. (C): The decision tree used to classify group indices determined using regression-based clustering.



(A) (B)

FIGURE 3.10: Lattice parameter data set – Results of regression-based clustering for binary compounds with L_{const} is set as the target variable. The compounds in the data set are divided into 3 separated groups. (a): the overall prediction accuracy achieved by combining 3 clusters is R^2 score 0.955 (MAE: 0.058 Å). (b): Unrealistic alloys - noble gas compounds are almost allocated on small linearity "edge" that bend an angle with the dominated component in group 3. It shows that all of this unrealistic compounds are belong to minority group differ from normal ones in this group

with A elements are lanthanide. More detail, in the same condition of B elements is either S, Se or Te, group 1 compounds have smaller unit cell volume than group 2 compounds. If B elements rather S, Se or Te, almost compound are assigned into group 4. In contrast, with A elements rather than lanthanide, the criteria for assigning elements into group 3 and 5 is the electron negativity difference, which is canonical rule for identifying type of bonding and formation energy characteristic.

Lattice parameter data for body centered cubic structure

For determining the number of cluster and the weighted average of the linearity for evaluating the performance of the clustering analysis, we found that the best set of variables is $\{\rho, m_A, m_B\}$ and the number of cluster is three. The prediction ability achieved by combining all 3 linear models reaches R^2 score 0.955 (MAE: 0.058 Å, Figure 3.10a). The confusion matrix in Figure 3.11b shows the high linearity of each individual model and the dissimilarity nature among groups in the data set.

One thing to notice in this experiment is a linearity separation of noble gas compounds shown in group 3. Following the figures 3.10b, by annotating name of noble gas instances, one can recognize that they are all lying a small “edge”. This part shows the linearity relationship of the group of noble gas that “bend” an angle comparing with the main trend in group 3. This result of linearity group separation shows an useful feature of our method in the attempt of understanding structure component of a data set.

The group label obtained for each compound using the regression-based clustering analysis is subsequently set as the target variable for the decision tree analysis. Figure 3.11a shows the obtained tree, which yields a correct classification

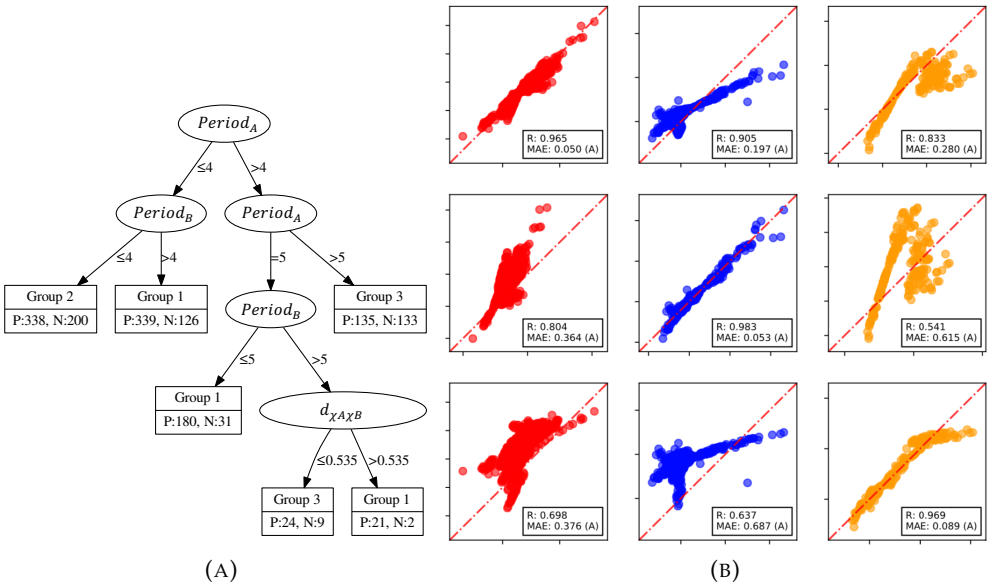


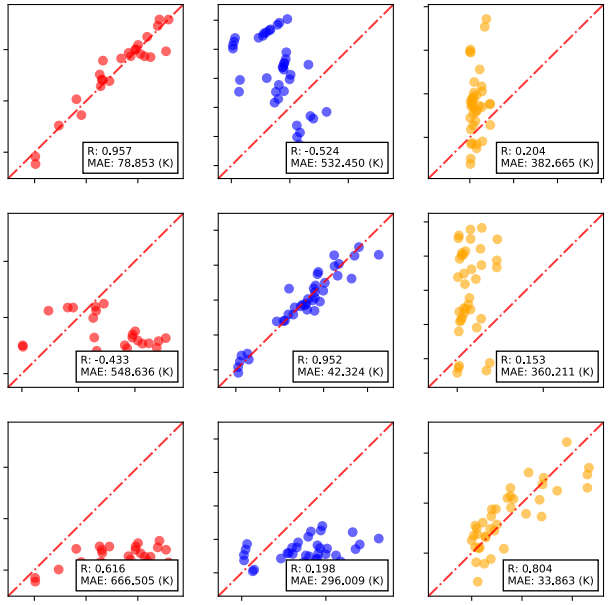
FIGURE 3.11: Lattice parameter data set – (a): Confusion matrix describes high dissimilarities among models. (b): The decision tree takes group index learned from regression-based clustering as target variable.

result 79.0% of the time, as evidenced by 10-fold cross validation. Following the tree, it is apparent that group 1 contains AB compounds with a light element (either $Period_A \leq 4$ or $Period_B \leq 5$) associate with other massive element ($Period_B \geq 4$ or $Period_A = 5$ respectively). In contrast, all compounds in group 2 constructed by pair of light elements ($Period_A \leq 4$ and $Period_B \leq 4$) and group 3 with massive elements ($Period_A > 5$).

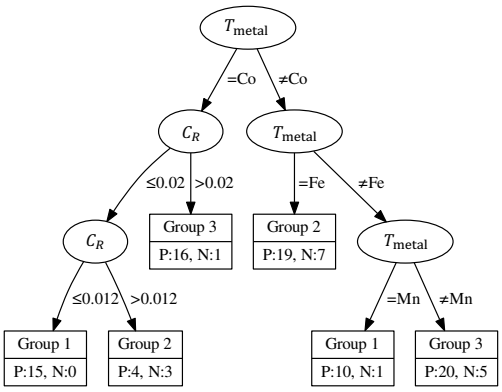
Curie temperature data set

In the next step, we carry out regression-based clustering analysis. For determining the number of cluster and the weighted average of the linearity for evaluating the performance of the clustering analysis, we found that the best set of variables is $\{C_R, S_{3d}, L_{3d}, S_{4f}\}$ and the number of clusters is three. The R^2 score of the mixed model constructed from the three linear models reaches 0.963 (MAE: 48.183 K, Figure 3.12c), which is comparable with the score of the best performing non-linear prediction model. Figure 3.12 shows a matching matrix of linearity of data in each group and the deviation of data in a group from the linear models of the other groups, allowing us to confirm the dissimilarity between the linear models and the linearity of data within in each model.

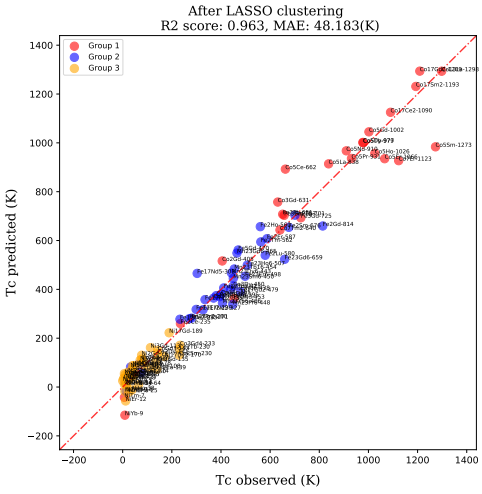
The group label obtained for each compound using the regression-based clustering analysis is subsequently set as the target variable for the decision tree analysis. Figure 3.12b show the obtained tree, which yields a correct classification result 80.0% as evidenced by 10-fold cross validation. It is apparent that the decision rules, employed for the determination of the group an alloy should belong to, are based on the species of the constituent transition metal and the concentration of the rare-earth metal. The critical role of C_R can be confirmed easily from results shown in Fig. 4.13, where the upper limit of T_C depends linearly on C_R. Furthermore, the dependence of T_C on C_R is qualitatively different for different transition metals. For Mn and Co as transition metals, T_C tends to decrease with C_R. By contrast, it tends to increase for Fe. In addition, for Ni, the T_C is rather insensitive to C_R. It is important to note



(A)



(B)



(C)

FIGURE 3.12: Curie temperature data set – (A) confusion matrix describing the dissimilarities among models, (B) the overall prediction accuracy achieved by combining 3 clusters. (C) The decision tree for the classification of group indices determined using regression-based clustering.

TABLE 3.5: Prediction accuracy of regression-based clustering

	T_c	E_{form}	L_{const}
Lasso	R^2 : 0.467, MAE : 213.965	R^2 : −0.108, MAE : 0.938	R^2 : 0.759, MAE : 0.136
GKR	R^2 : 0.911, MAE : 71.331	R^2 : 0.958, MAE : 0.162	R^2 : 0.981, MAE : 0.015
RBC	R^2 : 0.952, MAE : 51.199	R^2 : 0.960, MAE : 0.156	R^2 : 0.982, MAE : 0.017

that the decision tree model can reflect this situation quite well. From the decision tree, we can see that the group 1 includes bi-metal alloys of Co with low concentration of rare-earth metals and bi-metal alloys of Mn. The group 2 is dominated by bi-metal alloys of Fe, whereas group 3 consists mostly of alloys of Co and rare-earth metal at high concentrations, and bi-metal alloys of Ni. The obtained results confirm that our analysis flow can learn simultaneously and correctly the group of the material and the relationships between the descriptors and the corresponding physical phenomenon, *i.e.* T_C in the present case, for each group.

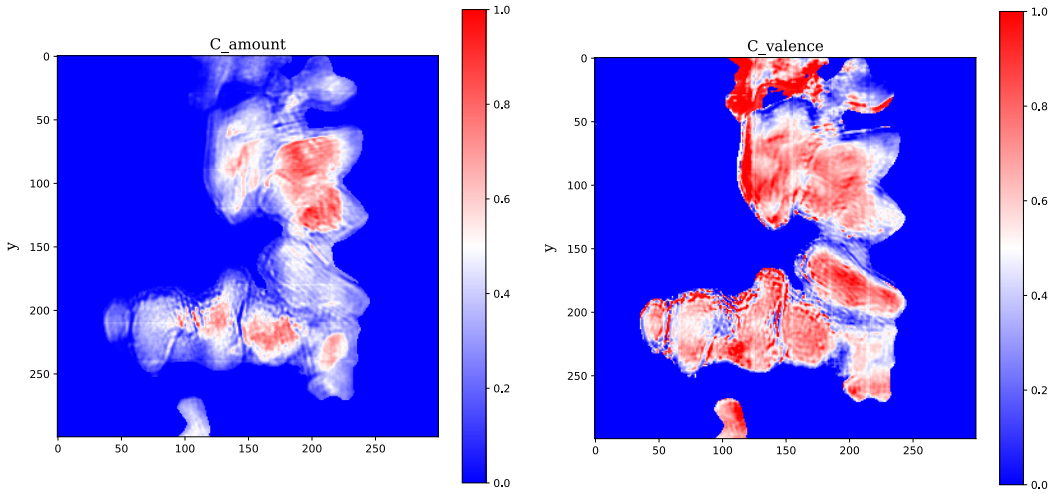
Prediction ability of regression-based clustering

The result in this part is measured by leave-one-out test set separation in T_c; 10-folds test set separation in AB compound E_{form} and L_{const} problem. This process is iteratively conducted over all data instances of three data sets.

In T_c experiment, the considered variable combination $\{C_R, S_{3d}, L_{3d}, S_{4f}\}$ has a kernel ridge regression prediction ability R^2 score of 0.911 (MAE: 71.331), and the temporal predicted value is taken from the highest predictive model. Finally, a remarkable improvement in the regression-based clustering model is achieved at a R^2 score of 0.952 MAE : 51.2. For other two problems, the prediction abilities are improved with the R_2 score from 0.958 to 0.969 in the E_{form} problem and remaining the same in range of 0.981 - 0.982 with the L_{const} problem. The detailed conclusion is shown in Table 4.1, including a comparison of the improvement to a simple linear model.

In the E_{form} and L_{const} problem, the regression-based method under the "temporal" prediction value from the highest prediction ability model could not reach a higher prediction ability than the top 1. However, compared with the initial nonlinear model, they significantly improve the accuracy: from R^2 : 0.958, MAE : 0.162 to R^2 : 0.959, MAE : 0.156 with the E_{form} AB compound problem and from R^2 : 0.981, MAE : 0.015 to R^2 : 0.988, MAE : 0.014 with the L_{const} binary cubic crystal problem.

Due to the gap between initial nonlinear model (the variable combination) and the top 1 model in these two cases being relatively small, detail in 4.1 and also lacking of domain knowledge (shown in initial state setting, the meaningful of variable combination), the regression based clustering method could not show too much effect like the case of T_c problem. One more thing we should emphasize here is that the suspected variable set does not correspond to the highest prediction ability among the non-linear models. These criteria along with the prediction ability improvement above show a potential strategy to investigate further the actual structure of data, which is out of scope for this work.



(A) (B)

FIGURE 3.13: Two original images of the cerium density ρ_{Ce}(left) and valence val_{Ce}(right) of Pt/Ce2Zr2Ox(x=7–8)

Identify different behavior groups in catalyst data set

In this section, we show a special data set that applicable by linear regression-based clustering method. Two images 555×550 that measure the cerium density ρ_{Ce} and valence val_{Ce} in micrometer-size platinum-supported cerium–zirconium oxide Pt/Ce2Zr2Ox(x=7–8) three-way catalyst particles are collected from M et al., 2018. The original images shown in refFigCatalyst2D. These images were successfully mapped by hard X-ray spectro-ptychography (ptychographic-X-rayabsorption fine structure, XAFS). By assuming the dependece of val_{Ce} to ρ_{Ce} originates from various linearity relation, we can conduct the linear regression-based clustering method as above.

By estimating the number of clusters as four, we have a map reflect back the distribution of each group as in Figure 3.13. Four groups denoted as G_i, with $i = \{1, 2, 3, 4\}$. The mixture of linear functions is simple in form:

$$val^i_{Ce} = a_i * \rho^i_{Ce} + b_i \tag{3.22}$$

with i as index of each group. The result from linear clustering method show coefficients a_i, b_i for each group as follows $a_1 = -0.33, b_1 = 3.90$, $a_2 = -0.15, b_2 = 3.60$, $a_3 = 0.34, b_3 = 3.03$ and $a_4 = 0.09, b_4 = 3.32$. It should be noticed that, there is a large difference among groups, especially in the linear coefficient a_i. Since the group G_1, (in green in 3.14) associates with largest negative slope of coefficient $a_1 = -0.33$, the group G_3 could be described under the largest positive coefficient $a_3 = 0.34$ (in purple in 3.14). The G_2 contains the slightly negative correlation between val_{Ce} and ρ_{ce} and the group G_4 could be considered as no linear correlation between these two variables. All these analysis is consistent with published result in M et al., 2018. Further than that, our groups detection shows extraodinary resolution comparing with the methods shown in M et al., 2018 and interestingly, the shape of identified group matches with 3D SEM images shown in the paper.

To investigate the behavior of each voxel by considering to linearity of neighbor, cubic 3×3 of neighbors for all voxels are collected. We perform fitting a simple

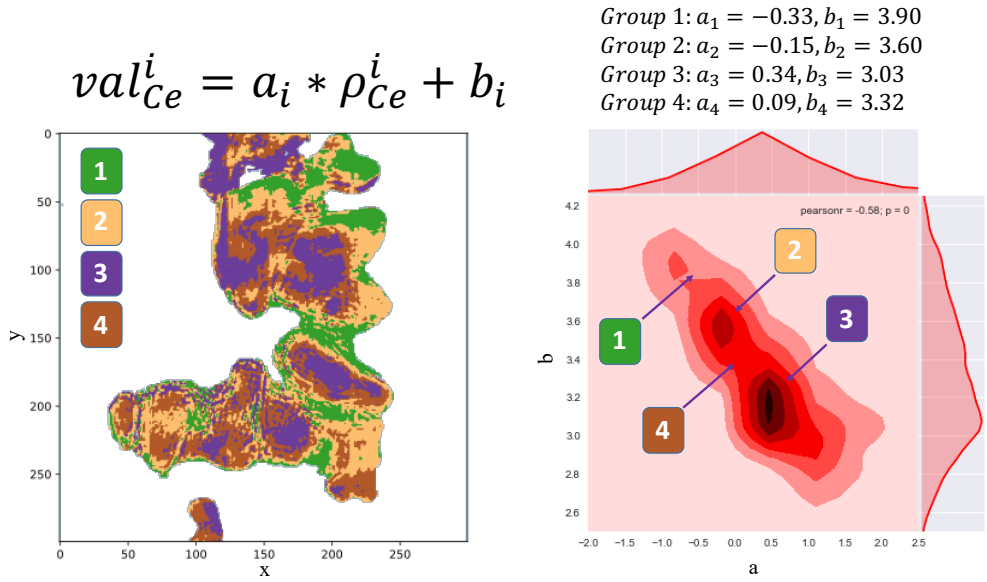


FIGURE 3.14: Linear regression based clustering results with four groups. Left: distribution of four groups over the original map. Right: joint distribution of all linear correlation coefficients a and intercepts b

linear relation on the neighbors then create joint distribution of all linear correlation coefficient a and intercept b, Figure 3.14.

3.6 Non linear regression ensembling

3.6.1 Overview

In a lot of scientific problems, we frequently observe data which are believed to be generated under distinct mechanisms with different setting contexts. From this aware, even non-linear supervised machine learning models for predicting physical properties of materials are used more and more frequent by empirical models, finding a model that qualitatively determines the mixture effect is always a demanded task in both theoretical establish and building experimental model. The application of unsupervised learning technics, with the ability to screen predefined correlations at different data scales, can be a promising approach LeCun, Bengio, and Hinton, 2015, 05. Conventional unsupervised learning technics for unveiling of mixture models in descriptive space are implemented using clustering methodsXu and Tian, 2015 such as the hierarchical clustering model, K-means method, etc.

Besides, the revealing of mixture models under the use of supervised models as the centers, has not gained much attention by experts working on machine learning. One of the well-known methods in this research direction is the mixture of experts model Jacobs et al., 1991; Seniha, Joseph, and Paul., 2012, which learns the gating functions to appropriately partition the descriptive space for identifying the components of mixture models. Furthermore, a number of linear regression-based clustering was recently developed in previous section or Eto et al., 2014; Hayashi and Fujimaki, 2013; Nguyen et al., 2018 without partitioning the descriptive space. However, the models by including a number of parameters as number of disjoint clusters, the complexity of the learners often show diverge performance.

In this section, we propose a method to unveil the mixture of information on the mechanism of physical properties of materials by using nonlinear supervised learning technics. The method is based on an ensemble method with Kernel ridge regression as the predicting model. We apply a bagging algorithm to carry out random subset samplings of the materials for generating multiple prediction models. The distribution of the predicted values for each material is then approximated by a Gaussian mixture model. Further, the contributions of the reference training materials to each of the corresponding models are investigated in detail. Reference training materials that are avoided and do not contribute to a predictive model, which accurately predicts the physical properties of a particular material, are considered dissimilar to that material.

3.6.2 Related methods

Canonical methods in unveiling mixtures of non linear regression models show through Mixtures of Local Experts with over twenty years of development Seniha, Joseph, and Paul., 2012; Jacobs et al., 1991, Mixture of Gaussian Process Rasmussen and Ghahramani, 2002; Meeds and Osindero, 2006; Ross and Dy, 2013; Souza and Heckman, 2014; Lázaro-Gredilla, Vaerenbergh, and Lawrence, 2012. In the thesis, there are two developed models which is possible to unveil mixture of non-linear regression functions does not require prior assumptions about using gating functions (mixture of experts) or not (mixture of Gaussian processes).

3.6.3 Methodology

In this research, we tend to unveil the mixture of information in prediction model space. The prediction space we emphasize here is a linear combination of kernel functions constructed by training materials/data instances. Applying the bagging algorithm Baldi and Sadowski, 2014, we carry out random subset samplings of the materials dataset to generate multiple prediction models. For each sampling, we prepare two separated data sets: bagging data set, $\mathcal{D}_{bagg}$, and testing data set, $\mathcal{D}_{test}$. These two data sets satisfy the condition $\mathcal{D}_{bagg} \cap \mathcal{D}_{test} = \emptyset$ and $\mathcal{D}_{bagg} \cup \mathcal{D}_{test} = \mathcal{D}$. With each of the two datasets $\mathcal{D}_{bagg}, \mathcal{D}_{test}$, we generate a prediction model by regressing the bagging datasets $\mathcal{D}_{bagg}$ using a cross-validation technic Stone, 1974; Picard and Cook, 1984. For each obtained prediction model, we collect the predicted values of the target property for all the materials in the corresponding testing data set $\mathcal{D}_{test}$. The canonical size of $\mathcal{D}_{bagg}$ is selected as 66% of the total number of data instances. By repeating the bagging process, each material $\boldsymbol{x}_i$ has an equal chance to appear in the test set $\mathcal{D}_{test}$. As the result, we obtain a predicted values of target property distribution $p(\hat{y}(\boldsymbol{x}_i))$ for all considered materials.

The null hypothesis represents for for an assumption of the homogeneity of the dataset in the kernel space or the existence of a single regression function. If the null hypothesis is true, the distribution $p(\hat{y}(\boldsymbol{x}_i))$ should be Gaussian for every material $\boldsymbol{x}_i$; else, we can significantly approximate the distribution $p(\hat{y}(\boldsymbol{x}_i))$ for a particular material $\boldsymbol{x}_i$ in the form of a mixture of Gaussian distributions. By examining the distribution $p(\hat{y}(\boldsymbol{x}_i))$, we can test the hypothesis on the homogeneity of our dataset.

The distribution $p(\hat{y}(\boldsymbol{x}_i))$ could be approximated by a mixture of K number of Gaussian distributions Murphy, 2012b as following:

$$p(\hat{y}(\boldsymbol{x}_i)|\theta) = \sum_{k=1}^{K} \pi_i^k \mathcal{N}(\mu_i^k, \sigma_i^k), \tag{3.23}$$

where π_i^k, μ_i^k, and σ_i^k are the weights, centers, and coefficient matrices of the constituent Gaussians components. Under a given number of mixture components, the parameters are estimated using an expectation-maximization algorithm, which is explained in detail in Murphy, 2012b. Maximizing Bayesian information criterion G., 1978 process is used to determine the number of mixture components. In practice, the evaluation process is performed by applying several different trials to randomize the initial states then selecting the maximize the Baysian information score. By combining the information from the bagging process, e.g investigate predicted value distribution, the correlation between training dataset we could reproduce mixture of regression works. In next chapter, we show that the use of bagging method as a partitioning data space produce very high potential to either building heuristic voting method shows or envidence combining based method shows in detail in chapter 4.

Chapter 4

Modeling similarity–dissimilarity concepts

4.1 Introduction

In this Chapter, three methods to qualitatively measure similarity–dissimilarity respect to a given target property are developed. The essence idea about similarity between two data instance A and B relying on whether or not the appearance of regression functions passing through them. A committee voting machine for similarity is constructed by result from linear regression-based partitioning methods in section 4.2. Section 4.3 shows the dissimilarity voting machine that take non-linear relation as the center. Lastly, an unify method for modeling similarity–dissimilarity simultanously by using evidence combining method from Dempster-Shafer theory is shown in section .

4.2 Committee voting machine for similarity measurement

4.2.1 Similarity voting machine

We developed a method for measuring the similarity between materials/data instances, focusing on specific a target physical property. The collected information can be utilized to understand the underlying mechanisms and to support the prediction of the physical properties of materials. The method consists of three steps: evaluating variable/variable combination based on non-linear regression, linear regression-based clustering in Chapter 3, and this chapter target to the similarity measurement work with a committee machine constructed from the clustering results. The entire data analysis flow is shown in Figure 4.1.

Three data sets of crystalline materials shown in Chapter 3 represented by critical atomic predicting variables are used as test beds. Three target variables are formation energy, lattice parameter, and Curie temperature respectively. Based on the information collected on the similarities between the materials, a hierarchical clustering technique is applied to learn the cluster structures of the materials that facilitate interpretation of the mechanism, and an improvement of regression models is introduced for predicting the physical properties of the materials. Our experiments show that rational and meaningful group structures can be collected and that the prediction accuracy of the materials' physical properties can be significantly increased, confirming the rationality of the proposed similarity measure.

The proposed linear regression-based clustering is applied for a number of predicting variables combination. The model shows a specific partitioning of the data set into groups in which the linear correlations between the predicting and target variables can be observed. The materials belonging to the same group potentially

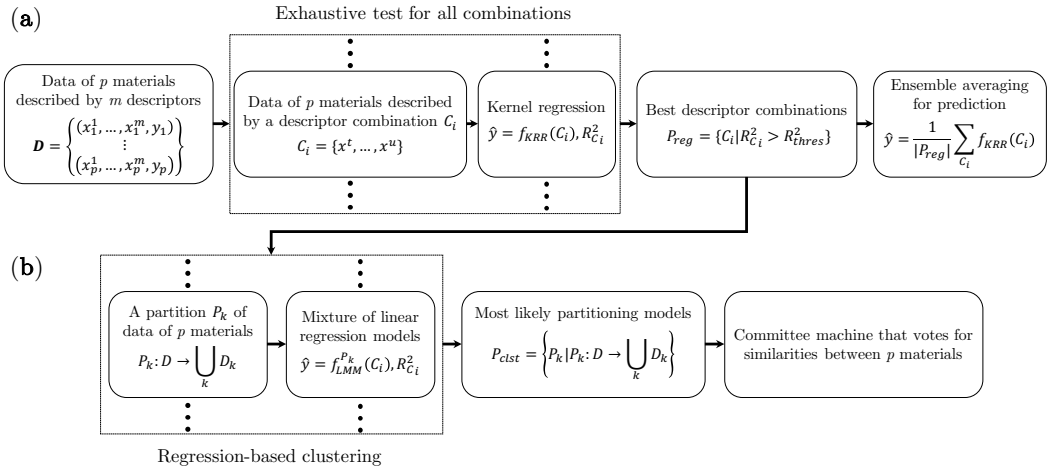
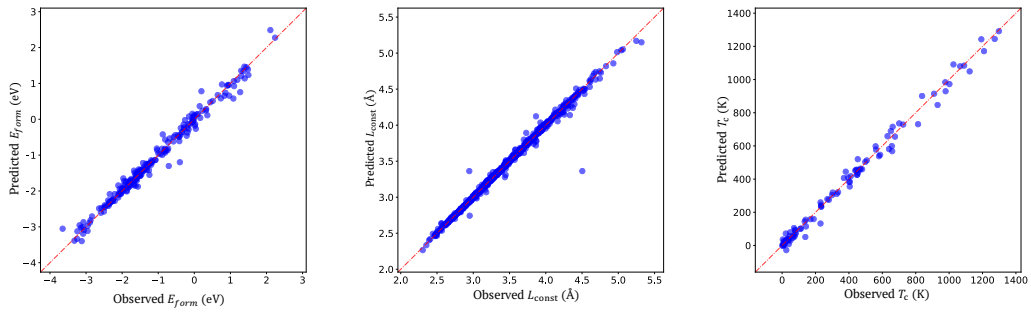


FIGURE 4.1: The data flow of our proposed method to measure similarity between materials, regarding to a given target physical properties. The method is illustrated under Map-Reduce language. The method consists of two sub-processes. The first process is kernel-regression based variable evaluation step: an exhaustive screening for all predicting variable combinations. By applying this step, one select the best variable combinations yielding the most likely regression models. The second process is an utilization of the regression-based clustering technique to search for partition models. break down the data set into a set of separated smaller data sets, so that each target variable can be predicted by a different linear model. We can obtain a prediction model with higher predictive accuracy by taking an ensemble average of the yielding models in (a). We use the collected partitioning models in (b) to construct a committee machine that votes for the similarity between materials.

have the same actuating mechanisms for the target physical property. However, materials that actually have the same actuating mechanisms for a specific physical property should be observed similarly in many circumstances. Therefore, the similarity between materials, focusing on a specific physical property, should be measured in a multilateral manner. For this purpose, for each pre-screening of the sets of predicting variables that yield non-linear regression models of high *PA* (section 3.4), we construct a regression-based clustering model. A committee machine which votes for the materials's similarity is then constructed from all collected clustering models. Two materials can be measured its similarity naively by using the committee algorithm Seung et al., 1992; Settles, 2010, by counting the number of clustering models that partition the two materials into the same cluster. The affinity matrix A of all pairs of materials in the data-set is then constructed as follows:

$$A_{a,b} = \frac{1}{|S_h|} \sum_{\forall S \in S_h} \sum_{i=1}^{k_S} w_{ia}^S w_{ib}^S \tag{4.1}$$

where S_h is the set of all pre-screened combinations of predicting variables that yield non-linear regression models of high *PA* and k_s is the cluster number. Further, $W^S = \left[w_{ij}^S \right]_{p \times k_S}$ is the partition matrix of the linear clustering models with the use of variable predicting variable combination. Each cell S (w_{ia}^S receive 1 if material a belongs to cluster i and 0 otherwise). By using this affinity matrix, one can



(A) (B) (C)

FIGURE 4.2: From left to right, observed and predicted target variable by taking ensemble averaging of 139 (E_{form} problem), 57 (L_{const} problem) and 59 (T_c problem) best prediction models including similarity measure information. By ensembling top 5 largest accuracy models yield a *PA* with R^2 scores of 0.982 (MAE: 0.101 eV) for predicting E_{form} problem, 0.992 (MAE: 0.011 Å) for predicting L_{const} problem and 0.991 (MAE: 24.16 K) for predicting T_c problem.

easily implement a hierarchical clustering technique Everitt et al., 2011 to obtain a hierarchical structure of groups of materials that have similar correlations between the predicting and target variables.

4.2.2 Experiments

Experiment 1: Formation energy of $Fm\bar{3}m$ AB materials data set

In this experiment, we perform three times ten folds cross validation to evaluate all prediction models derived by all possible combinations of our seventeen designed variables. The total number of all variable combination are 2^{17} - 1 = 131,071. Then, after analyze these result, we finally obtain 34,468 variable combinations associated with Gaussian kernel ridge regression models with R^2 scores larger than 0.90 (Fig.4.2). In particular, there are 139 prediciton models accompanied with R^2 score value larger than 0.96. Those designed variable combinations are used for the next analysis work. We also obtain he highest prediction accuracy *PA* at level of R^2 score 0.967 and the MAE: 0.122 eV. The s_{PA} is $\{V_{cell}, \chi_A, n_{eA}, n_{eB}, IP_A, T_{bA}, T_{mA}, r_B\}$. None of the less, we archive even a prediction model with the predicton ability higher than the *PA* with R^2 score is 0.972 (MAE: 0.117 eV) by averaging Tresp, 2001; Dietterich, 2000; Zhang and Ma, 2012 of the 139 prediction models mentioned above.

In the work of applying linear regression-based clustering, we use 139 prediction models that derived from variable combinations mention aboved. Each experiment was performed under one thousand initial random initial states for each variable set. From 3.20), we collected in total over 200 best clustering results after the use of applying the same criteria for determining the number of clusters. These results are used to build a voting machine to measure the similarity among all materials. The collected similarity matrix for all the $Fm\bar{3}m$ AB materials shown in Fig.4.3. The similarity value between each pair of materials varies in range from 0 to 1. A similarity value of the similarity matrix takes zero if the any two materials are not ever included in the same cluster, under the partitioning method by the linear regression-based clustering. On the other hand, the similarity value takes the maximum value

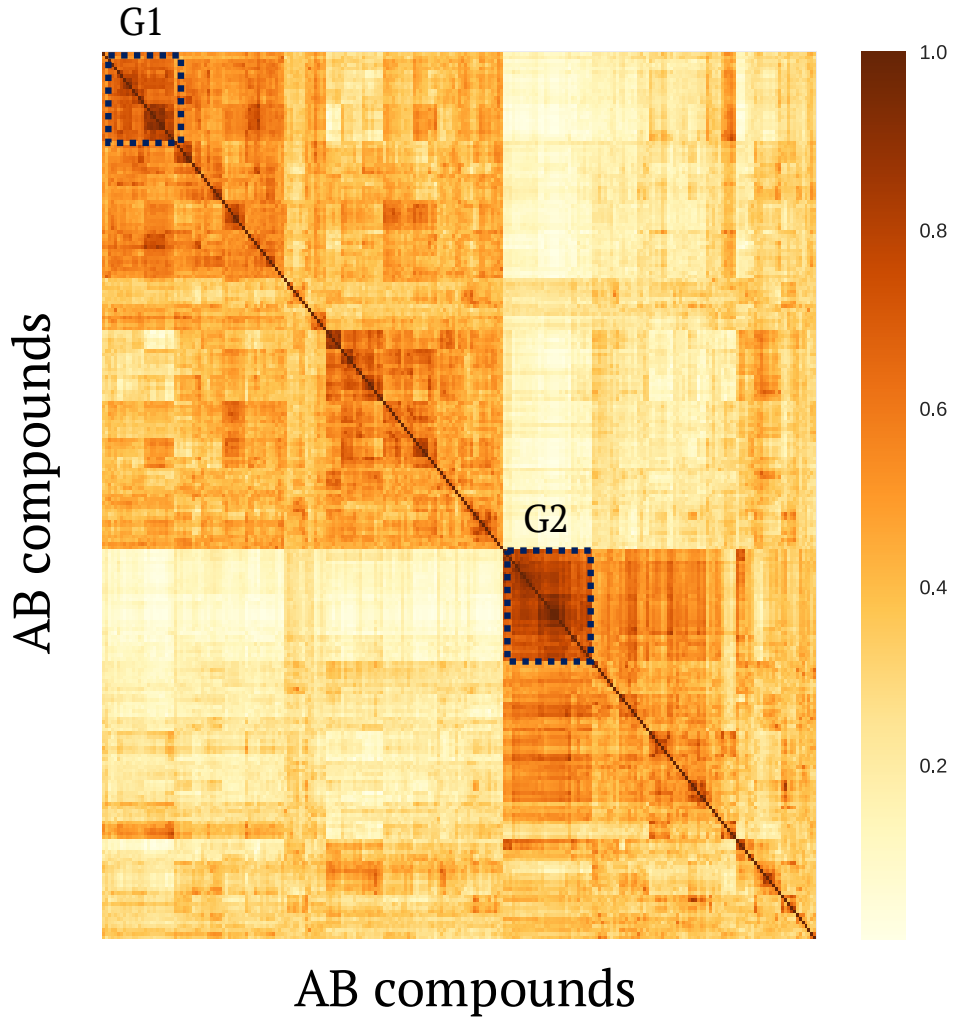
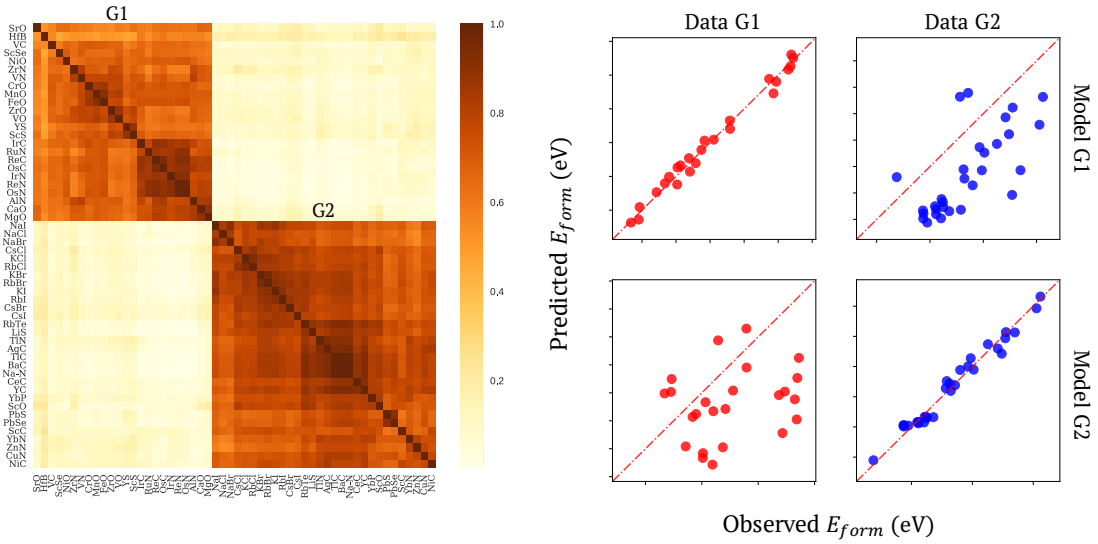


FIGURE 4.3: a) Affinity matrix between the $Fm\bar{3}m$ AB materials yielded by regression-based committee voting machine.

of one if the any two materials always showing in the same cluster. Further investigation, we could roughly divides all the materials in the data-set into two distinct groups. The detail information are represented by the upper left and bottom right of Fig.4.3.

Figure 4.4a shows an broaden image of the simiarity matrix for two groups G1 and G2 of high similarity materials region. It is clear to see that the affinities between materials within each of these two groups, G1 and G2, exceed 0.7. In other word, all materials of each group have high intragroup similarity. In contrast, the affinities between materials in different groups are smaller than 0.2, showing significant dissimilarity between G1 and G2. Further detailed investigation reveals that the materials in G1 are oxide, nitride, and carbide. The maximum common positive oxidation number of the A elements is greater than or equal to the maximum common negative oxidation number of the B elements for the compounds in this group. On the other hand, the materials in G2 are halides of alkaline metal, oxide, nitride, and carbide, for which the maximum common positive oxidation number of the A elements is less than or equal to the maximum common negative oxidation number of the B elements. Looking more into details of extracted results, the matrix shows only seven among 24 materials in group G1 have computed electronic structures



(A)

FIGURE 4.4: a) Broaden view of highly similar elements in G1 and G2 regions in affinity matrix. b) Confusion matrixes measuring linear similarities among materials in G1 and G2, as well as dissimilarities between models generated for materials in different groups.

with a non-zero band gap. In contrast, half of the compounds in G2 have computed electronic structures with a band gap. The collected results suggest that the bonding nature of compounds in G1 is different from that of compounds in G2.

The linearities between the target variable and the predicting variables for the two groups are summarized in Fig.4.4b. The diagonal images illustrate the linear correlations between the observed and predicted values of the target variables collected by using linear models of the predicting variables for the materials in the two groups. On the other hand, the off-diagonal images illustrate the linear correlations between the observed values and predicted values for the target variables collected using the linear models of the other groups. These results confirm the intra-group similarity and the dissimilarity among different groups.

To quantitatively evaluate the validity of the analysis process, we embedded the similarity measured by the committee machine into the regression of E_{form} of the $Fm\bar{3}m$ AB materials. For predicting the value of the target variable of an unknown material, instead of using the entire available data set, only one-third of the available materials having the highest similarity to the new material are selected. It should again be noted that the similarity between the materials in the data set and the new material can be determined without knowing the value of the target physical property, using the value predicted by ensemble averaging of the non-linear regression models.

Table 4.1 summarizes the PA in predicting E_{form} values of the $Fm\bar{3}m$ materials collected using several regression models with the designed predicting variables. The non-linear model collected using ensemble averaging of the best non-linear regression models, having an R^2 score of 0.972 (MAE: 0.117 eV), could be improved significantly to an R^2 score of 0.982 (MAE: 0.101 eV) regarding the information from the similarity measurement (Fig.4.2a). Therefore, the collected results provide significant evidence to support our hypothesis that the similarity voted by the committee machine reflects the similarity in the actuating mechanisms of the target material physical property.

TABLE 4.1: *PA* values for E_{form}, L_{const}, and T_c prediction problems. The results collected with and without using the similarity measure (SM) information are shown for comparison.

Prediction method		E_{form} (eV)		L_{const} (Å)		T_c (K)	
		without SM	with SM	without SM	with SM	without SM	with SM
GKR with all variables	R^2	0.929	0.954	0.982	0.986	0.893	0.929
	MAE	0.189	0.154	0.022	0.018	78.80	58.09
GKR with the best variable combination	R^2	0.967	0.978	0.989	0.992	0.968	0.988
	MAE	0.122	0.110	0.014	0.013	42.74	25.76
Ensemble of GKRs with top selected best variable combinations	R^2	0.972	0.982	0.991	0.992	0.974	0.991
	MAE	0.117	0.101	0.013	0.011	37.87	24.16

Experiment 2: Lattice parameter for body-centered cubic structure data set

In this experiment, we perform three times ten folds cross validation to evaluate all prediction models derived by all possible combinations of our seventeen designed variables. The total number of examined combinations are (2^{17} - 1 = 131,071). In these combinations, we finally found 60,568 variable combinations for deriving GKR models with R^2 scores exceeding 0.90 (Fig.4.2). Among them, there are 57 variable combinations yielding regression models with R^2 scores exceeding 0.9895. The highest *PA* for this experiment is 0.989 (MAE: 0.014 Å), which is collected using the combination $\{\rho, \ell_A, r_{covB}, m_A, m_B, \rho_B, n_{eB}\}$. We could obtain a better *PA* with an R^2 score of 0.991 (MAE: 0.013 Å) by taking ensemble averaging of GKR models which derived from the 57 selected variable combinations. This result is a considerable improvement in comparison with the maximum *PA* (R^2 score: 0.90) of the support vector regression technique with the feature selection strategy mentioned in Takahashi et al., 2017.

In the regression-based clustering analysis, the 57 selected variable combinations accompanied by 1000 initial randomized states for each combination are used to search for the most probable clustering results to construct the committee machine. The affinity matrix collected for all materials is shown in Fig.4.5a, after rearrangement by a hierarchical clustering algorithm Everitt et al., 2011. By utilizing this similarity, we could roughly divide all materials into three groups: G1, G2, and G3. Further investigation revealed that most materials in G1 are constructed from two heavy transition metals. In contrast, the materials in G2 and G3 are constructed from a metal and a non-metal element, *e.g.* oxide and nitride. For a given A element, the L_{const} of the materials in G1 increases with the atomic number of the B element. On the other hand, the L_{const} of the materials in G2 remains constant for the materials sharing the same A element. Further, the L_{const} for the materials in group G3 mainly depends on the electronegativity difference between the constituent elements A and B. Note that the materials in these three groups are visualized in detail in the Supplemental Materials. The linearities between the observed and predicting variables for these groups are shown in Fig.4.5b.

For predicting the L_{const} of a new material, we use the same strategy as that explained in the previous experiment. Table 4.1 summarizes the *PA* values collected in our experiments. The non-linear model collected using ensemble averaging of the

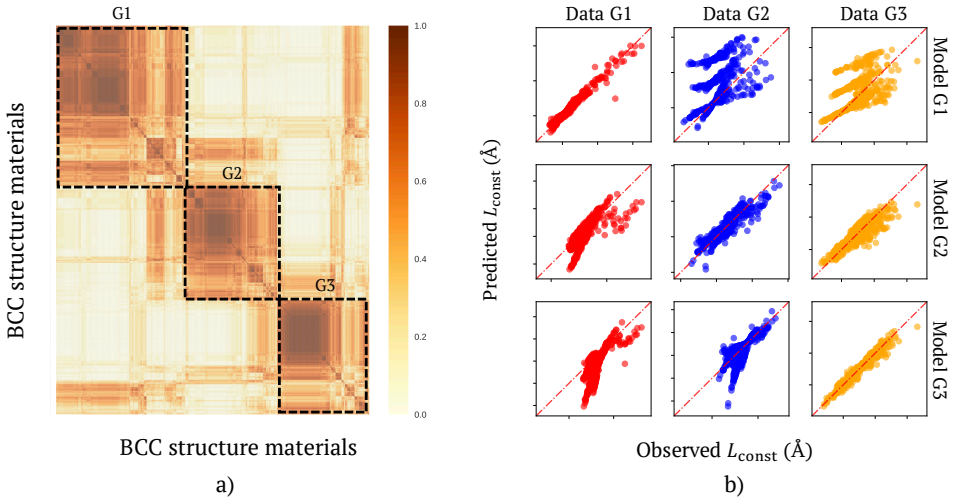


FIGURE 4.5: a) Similarity matrix between materials for L_{const} prediction problem yielded by regression-based committee voting machine. This similarity matrix can be approximated as three disjoint groups of materials denoted by G1, G2, and G3. b) Confusion matrixes measuring linear similarities among materials in each group, as well as dissimilarities between models generated for materials in different groups.

best 57 non-linear regression models and having an R^2 score of 0.991 (MAE: 0.013 Å) could be marginally improved to an R^2 score of 0.992 (MAE: 0.011 Å) by including information from the similarity measurement (Fig.4.2b).

Experiment 3: Curie temperature of Rare earth Transition metal magnetic data set

In this experiment, we perform leave one out cross validation to evaluate all prediction models derived by all possible combinations of our designed variables. The total number of examined combinations are $2^{21} - 1 = 2,097,151$. In this result, we finally found 84,870 variable combinations in which the deriving GKR models are all shown R^2 scores larger than the threshold 0.90 (Fig.4.2). In this result, there are fifty nine variable combinations that all derive Gaussian kernel ridge regressison models accompanied with R^2 scores over 0.95. Those combinations of designed variables are collected to apply into the next analysis step. The maximum value PA in this experiment is 0.968 (MAE: 42.74 K), under the use of the variable combination $\{C_R, Z_R, Z_T, \chi_T, r_{covT}, L_{3d}, J_{3d}\}$. Furthermore, we obtain a prediction model associated with prediction ability higher than the PA with R^2 score at 0.974 (MAE: 37.87 K), by taking average of the top 59 highest prediction accuracy models.

In applying linear regression-based clustering analysis, there are in total 59 variable combinations deriving highest prediction accuracy which used to search for the most probable clustering results. Each case has been applied with 1000 random initial states to construct the committee machine that votes for the similarity between the alloys. The collected affinity matrix for all the alloys is shown in Fig.4.6. An broaden view of the three groups of alloys having high similarity (denoted G1, G2, and G3) is shown in Fig.4.7–left. Further investigation revealed that G1 includes Mn- and Co-based alloys with high T_c, *e.g.* $Mn_{23}Pr_6$ (448 K), $Mn_{23}Sm_6$ (450 K), Co_5Pr (931 K), and Co_5Nd (910 K). Other low-T_c Co-based alloys, *e.g.* Co_2Pr (45 K) and Co_2Nd (108 K), are counted as having higher similarity with Ni-based alloys in G3, *e.g.* Ni_5Nd (7 K) and Ni_2Ho (16 K). In contrast, G2 includes all the Fe-based

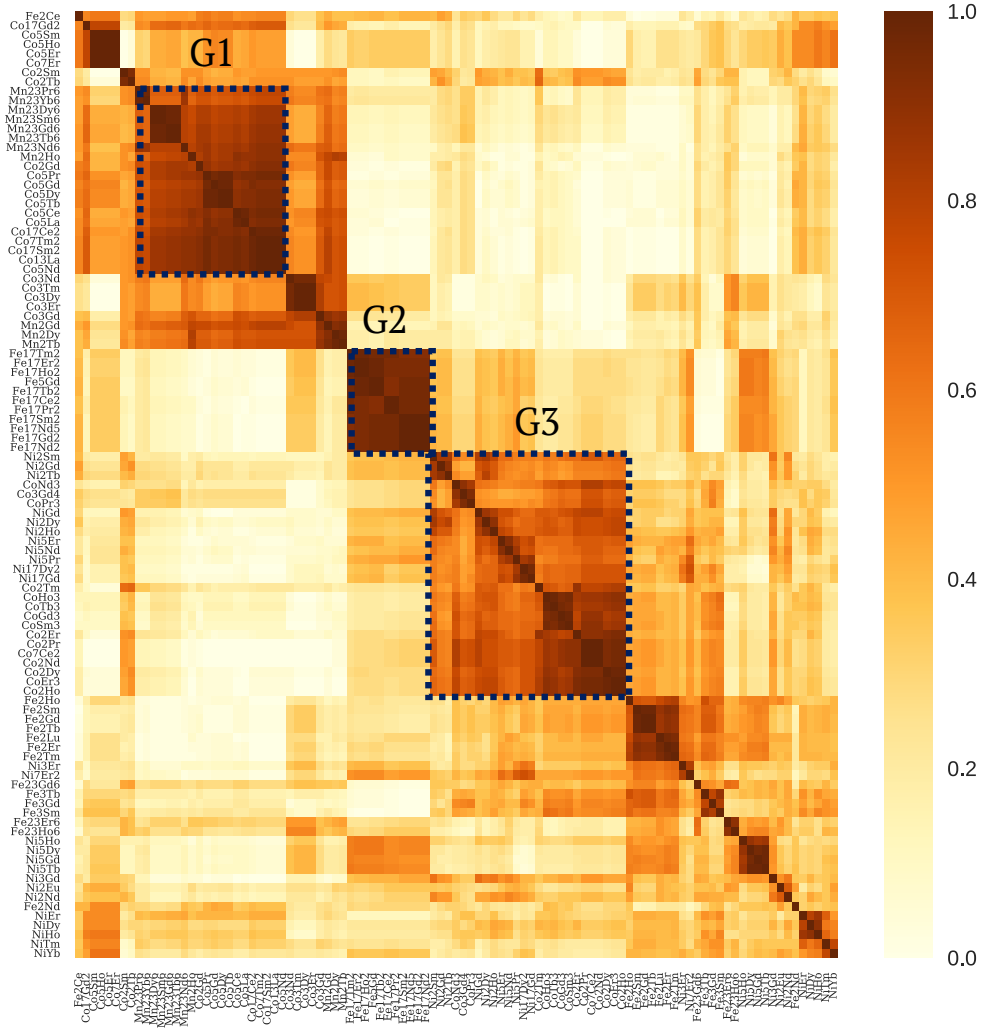


FIGURE 4.6: Similarity matrix between the rare-earth–transition metal alloys yielded by regression-based committee voting machine.

$Fe_{17}RE_2$ alloys, where RE shows different rare-earth metals. To confirm the value of our similarity measure, Fig.4.7–right shows the linearities between the observed and predicting variables for these groups, as well as the dissimilarities among these groups.

In the next analysis step, we used the collected similarity measure for predicting T_c for a new material by using the same strategy used in the two previous experiments. The non-linear model collected using ensemble averaging of the best non-linear regression models and having an R^2 score of 0.974 (MAE: 37.87 K) could be improved significantly to attain an R^2 score of 0.991 (MAE: 24.16 K) by utilizing the information from the similarity measurement (Fig.4.2c and Table 4.1). The collected results provide significant evidence to support our hypothesis that the similarity voted for by the committee machine indicates the similarity in the actuating mechanisms of the T_c of the binary alloys.

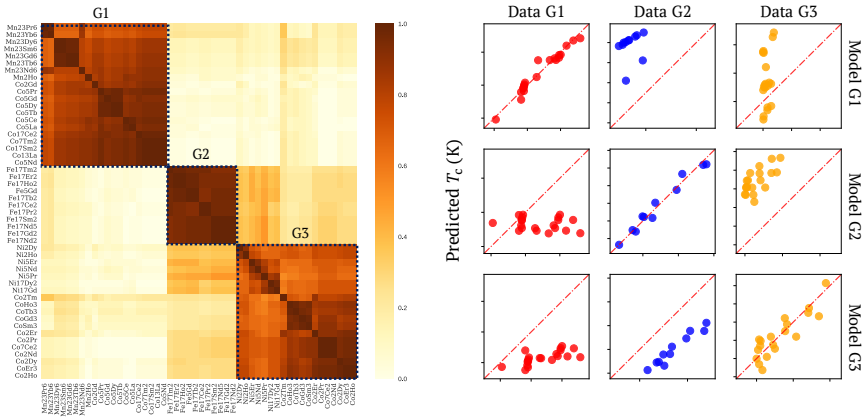


FIGURE 4.7: Left: Broaden view of highly similar elements in G1, G2, and G3 regions in similarity matrix. Right: Confusion matrixes measuring linear similarities among alloys in each group as well as dissimilarities between models generated for alloys in different groups.

4.2.3 Conclusion

In this work, we developed a method to measure the similarities between alloys, focusing on specific physical properties, to describe and interpret the actual mechanism underlying a physical phenomenon in a given problem. The proposed method consists of three steps: variable evaluation based on non-linear regression, linear regression-based clustering, and similarity measurement with a committee machine constructed from the clustering result. We applied our proposed method to three data sets of crystal materials that represented by key atomic predicting variables. Three different physical target properties: the formation energy, lattice parameter, and Curie temperature are used to examine our considered alloys. The extracted results show that rational and meaningful group structures can be synthesized by utilizing our proposed approach. The similarity measure information significantly improve the prediction accuracy for the prediction ability of all experiments.The ensembling method applied to top kernel ridge prediction models, the R^2 score significantly improve from 0.972 to 0.982 to predict the formation energy; improve from 0.974 to 0.991 in predicting the Curie temperature. There is slightly increasing in the prediction accuracy for work of predicting the lattice constant. Through these results, our proposed data analysis could be seem as a systematic method to help researcher go further in interpreting and understanding different physical phenomena by recognizing similarity patterns hidden inside the data set.

4.3 Committee voting machine for dissimilarity measurement

4.3.1 Dissimilarity voting machine

In this work, we use the information from the bagging experiment to vote for the dissimilarity among data instances. To perform the dissimilarity voting procedure, the very first need is to define a threshold in prediction error δ_{thres} for all prediction models $\hat{f}$ learned from a data set $\mathcal{D}_{bagg}$ (Eq. 3.9), which satisfies $|\hat{f}(\boldsymbol{x}_i) - y_i| < \delta_{thres}$, are collected. The second need of the method is to define the condition of considering neighbors in description space k_{thres}, for all pairs of $\boldsymbol{x}_i$ in $\mathcal{D}$ and $\boldsymbol{x}_j$ in the corresponding $\mathcal{D}_{test}$. Then, a vote $ds(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to the dissimilarity state between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as following:

$$ds(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} 1, & \text{if } c_i = 0 \; and \; k(\boldsymbol{x}_i, \boldsymbol{x}_j) < k_{thres} \\ 0, & otherwise \end{cases}. \tag{4.2}$$

In this voting machine, for each alloy, we pay more attention on its relationship with the neighborhood alloys (in the data set) in the description space. If the neighbor alloys are omitted or in other word, it does not contribute to the predictive models that accurately predict the target variable value of the concerning alloys, those neighborhood alloys are considered dissimilar to our concerned alloys.

The pseudo code for the bagging-based dissimilarity voting algorithm is summarized as following:

Algorithm 1: Bagging-based dissimilarity voting algorithm

Data:

Dataset: $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \ldots (\boldsymbol{x}_p, y_p)\}$

Base learning: $\hat{f}$

Number of base learners: T

Parameters: $k_{thres}, \delta_{thres}$

Result: Dissimilarity matrix, $\boldsymbol{dS}$

```
begin
    for t ← 1 to T by 1 do
        h_t = f̂(D, D_bagg)
    H(x) = Σ_{t=1}^{T} 𝕀 (h_t(x_*) ≤ δ_thres)
    dS = 0 with dS = [ds_ij]_{p×p}
    foreach h_t ∈ H do
        forall k(x_i, x_*t) ≤ k_thres do
            if c_i = 0 then
                ds_ij += 1 ∀x_j ∈ h_t
return dS.
```

4.3.2 Experiments

Experiment 1: Prototype models

For illustrate the effect of applying the ensemble–bagging prediction model in investigating the structural insight data sets, we present the results of applying the model to two-dimensional simulation data. The prototype data set contains seventy instances with a one-dimensional descriptive variable, x, and target variable, y, as depicted in Figure 4.8b. The bagging prediction model includes one million random samplings, with the sampling size is 35 % of the total number of data instances. Details about setting parameters are described in Table **??** in Appendix **??**.

Figure 4.8a shows the distributions of the predicted values, $\hat{y}$, collected using the bagging model. It is obvious that, for x values lesser than -0.4, the $\hat{y}$ distributions include a single distribution centered around 0.1. On the other hand, for x values greater than -0.4, almost all the $\hat{y}$ distributions can be considered to be a mixture of two main Gaussians, whose mean are approximately 0.2 and -0.1, respectively.

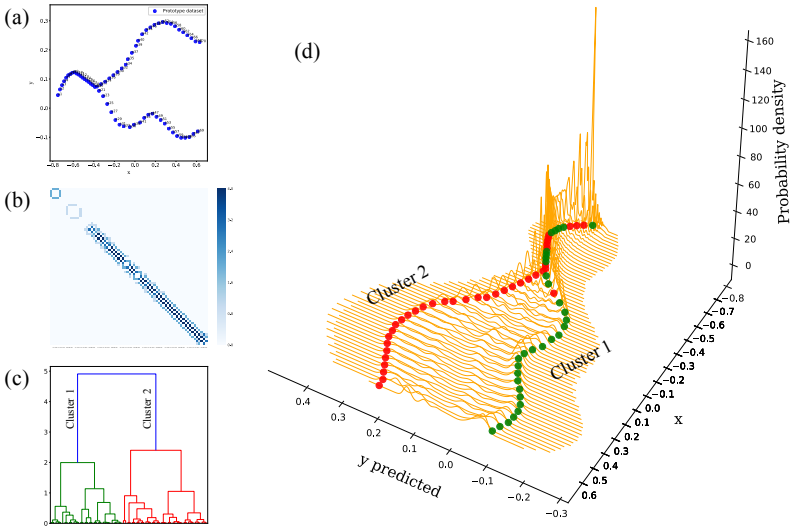


FIGURE 4.8: a) Visualization of the prototype data with the one-dimensional predictor variable, x, and target variable, y. b) Dissimilarity voting matrix with colored cells show the dissimilarity pairs of materials. c) Hierarchical clustering graph is constructed by embedding information of dissimilarity voting matrix. d) Distribution of the predicted values y along the x axis, applying the bagging model (orange lines) and observed data (green and red points), which are clustered using hierarchical clustering technique and information from dissimilarity voting results.

These distribution components of the predicted value, $\hat{y}$ reflect the actual shape of the designed data, shown by colored points.

Figure 4.8c shows the dissimilarity (70×70) matrix collected by our developed dissimilarity voting machine. In the matrix, dark blue cells represent dissimilarity pairs of data instances. Zero value cells show no dissimilarity information between corresponding pairs. For convenience, the ordered data instances shown in the matrix are sorted by the x values. Details about how the voting machine works to detect dissimilarity effect are shown by zero contribution example profiles in Appendix **??** section **??**.

There are two noticeable points extracted from this figure. First, the upper left of the matrix shows a large bright region or the region of non-dissimilarity among instances. It is consistent with the monotone and smoothly changes for x values lesser than -0.4. Second, for x values larger than -0.4, one can notice that any data instances in this region are dissimilar with their two closest neighbors and similar to the next ones. Once again, this extracted information shows consistency with the actual distribution of the designed data set.

By transferring the extracted information from the dissimilarity matrix to hierarchical clustering Murtagh and Contreras, 2011, one can acquire the clustering outcome as shown in figure 4.8c. The dissimilarity information from the voting machine bring the ability of identification the data set contains a mixture of two main groups, labeled by green and red color in the figure. These two groups successfully reassemble the distribution of two branches in the bifurcated region of the data set. To summary, through analyzing result extracted from the prototype data set, the dissimilarity voting machine based on the ensemble-bagging algorithm is shown the ability to unveil the mixture regressions/phenomena regarding a specific target property.

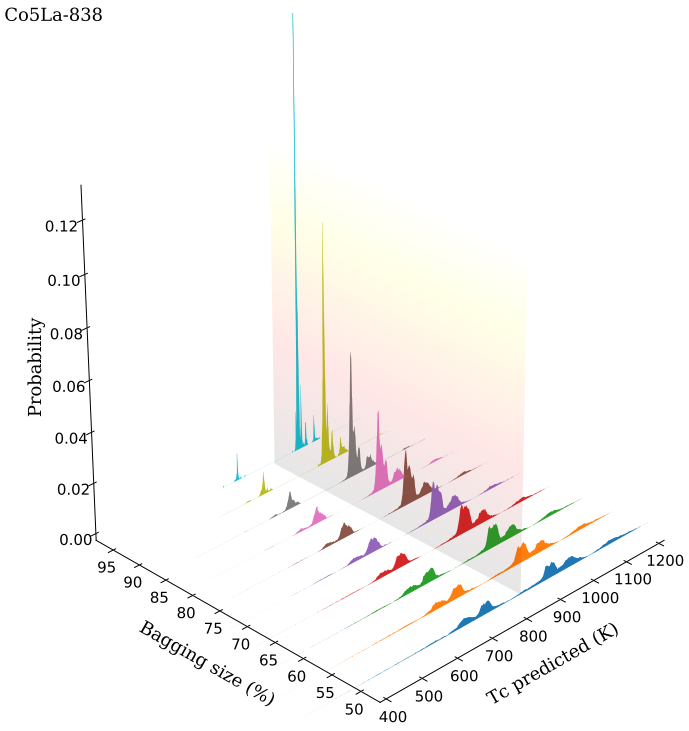


FIGURE 4.9: T_C predicted-value distribution of Co_5La for different bagging sizes. The constant plane shows the position of the observed value.

Experiment 2: Curie temperature of rare-earth–transition metal alloys data sett

For the pre-designed descriptive variables, the maximum prediction ability with an R^2 score of 0.967 ± 0.004 is achieved by a model derived from the variable combinations, $\{\chi_R, \chi_T, J_{4f}\left(1-g_j\right), Z_T, r_{covT}, IP_T, S_{3d}, L_{3d}, J_{3d}, C_R\}$. The high prediction accuracy level of this model shows that it is possible to accurately predict the T_C values of rare-earth transition alloys with these designed variables. In other words, under the use of this variable combination, the regression function for predicting T_C could be seem approximately as single function. However, there are with a number of probable unknown anomaly alloys in the model with higher prediction error comparing to the others. Our designed dissimilarity voting machine could help to address this problem.

In the following, we analyze the predicted value distribution of the Curie temperature T_C. Almost predicted value distribution for all alloys associate with a single Gaussian function distribution. However, there are a number of alloys which associated with its non-ordinary distributions. In the that are a mixture of Gaussians. For instance, figure 4.9 shows the T_C predicted-value distribution of Co_5La (with observed T_C of 838 K) for bagging sizes varying from 50–95 percent of the total data set. It is clear to see that, the distribution of predicted values is consistent for all setting the bagging size in the subset sampling selection. Even by varying the size, the corresponding Gaussian distributions whose peaks at 686 K, 739 K, 925 K, and 980 K remained stable. Figure 4.10 displays an broaden view of the distribution for a ensemble-bagging size of 65 percent of the total data set. In looking into detail result, on applying the Gaussian mixture model, it is clear to see that this distribution is a mixture of seven Gaussian distributions. All distribution components associate with

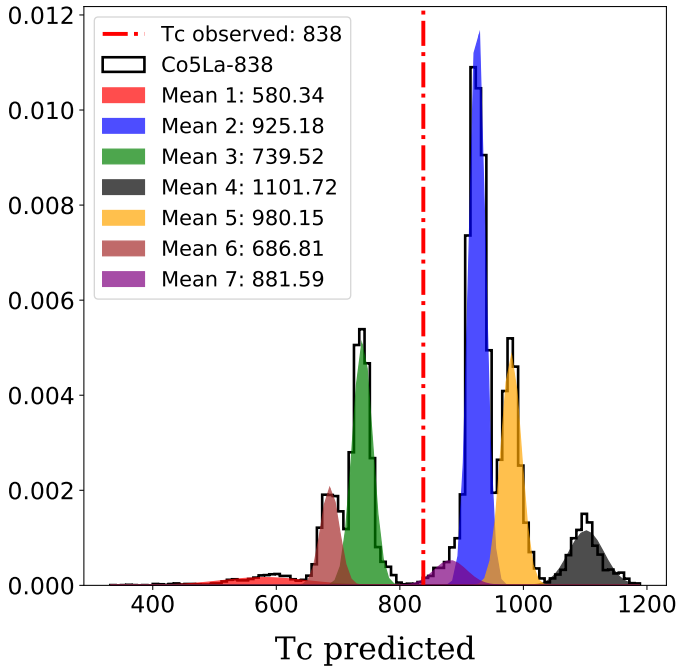


FIGURE 4.10: T_C predicted-value distribution of Co_5La with a bagging size of 65 % of the total data instances in the dataset. The distribution is a mixture of seven Gaussian components. The red dashed lines shows the positions of the observed values.

means at 580.34 K, 686.81 K, 739.52 K, 881.59 K, 925.18 K, 980.15 K, and 1101.72 K, respectively. This indicates the appearance of a mixture of nonlinear functions in the structural insight data set. Further investigations show the significance of appearance of these functions.

Figure 4.11 shows the contribution of each training alloy to the target alloys, Co_5La. The color bar encodes contribution values of all materials in the data set in concerning to Co_5La. It is clear to recognize the zero center symmetric distribution with colors in white (zero contributions)of any materials do not appear in training set. The vertical axis shows those sorted by the predicted T_C value, i.e., the summation of all the contribution rows. The horizontal axis shows materials with an ascending order of the L_1 distance to the target material. The top five closest to Co_5La are: Co_5Ce (662 K), $Co_{13}La$ (1298 K), $Co_{17}Ce_2$ (1090 K), Co_5Pr (931 K),and Co_7Ce_2 (50 K). This figure also shows that models with the closest predicted value of T_C (purple distribution with a mean of 881.59 K) are constructed from a combination of instances, $\mathcal{D}_{bagg}$, with no contribution from Co_5Ce, $Co_{13}La$, $Co_{17}Ce_2$, and Co_7Ce_2. Only Co_5Pr shows significant contribution to Co_5La. Details about zero-contribution counting profiles are shown in Figure **??** in Appendix **??** .

For comparing, the two nearest neighbor models of the purple model, namely, the distribution in green associated with mean 739 K, and the distribution in blue associated with mean 925.18 K are used to further analyzing. For the blue distribution, top five nearest neighbors are considered as contributors. However, in considering to the green distribution, the contribution of $Co_{13}La$ alloys is omitted. The results obtained by analyzing contributions of the training alloys as shown above, provide meaningful clues on the actual physical meaning. The three alloys, Co_5Ce,

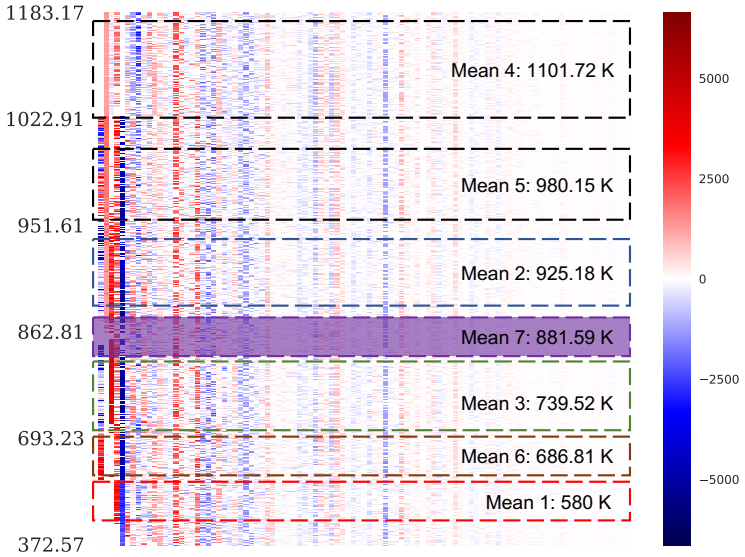


FIGURE 4.11: Heat map visualization depicting the contribution of the training alloys to the target alloys under the different prediction models in Figure b). The horizontal axis shows training materials sorted by the L1 distance to the target material on the description space. The vertical axis shows the predicted T_C value with sorting order, i.e., the summation of all the contributions.

$Co_{13}La$, and $Co_{17}Ce_2$, should not be considered highly "similar" to Co_5La with respect to T_C, even though they have the same constituent T-metal and the same constituent R-metals or are positioned next to each other on the periodic table (Z_{La}=57 and Z_{Ce}=58). In other words, the distance between these materials with respect to the rare-earth element difference should not be close, as measured by the three R predictive variables, χR, $J_{4f}(1-g_j)$, and C_R. As the concentration of the R-metal efficiently indicates the change in T_C values among those sharing the same R and T alloys, e.g., Co_5La vs $Co_{13}La$, these dissimilarity results show that the other two R variables do not capture the real mechanism of T_C.

From the clustering result, the collected Curie temperature data set could be divided into four main groups. These groups are classified basing on the transition metal elements: Cobalt based, Iron based, Manganese based, and Nickel based alloys. In our defined k_{thres} neighbor regions, the number of dissimilarity values is not identical for all the groups of alloys. Here, we analyze the cobalt-based and iron-based material groups. In the cobalt-based group, we can notice that Co_5Ce does not receive contributions from the other alloys of the group Co_5R. The dissimilarity between the Co_5Ce and the Co_5La alloy is shown in the previous analysis. Here, the dissimilarity can be observed more distinctly. Compared to the other Co_5R alloys, Co_5Ce has a T_C of 662 K, which is considerably lower than those of Co_5La at 838 K, Co_5Pr at 931 K, Co_5Nd at 910 K, and Co_5Sm at 1016 K. In this family, except for Co_5Ce, an increase in the atomic number of the rare-earth element correlates to an increasing T_C value. Figure 4.12 shows the hierarchical clustering result collected by utilizing the information from the dissimilarity voting machine. It is obvious that Co_5Ce is isolated from other Co-based alloys. We can also confirm the anomalousness of Co_5Ce by comparing Co_5Ce with other Co-based alloys (Fig. 4.13, the compounds surrounded by red line). This result confirms the significance of our method of dissimilarity measurement.

In the group of iron-based alloys, Figure. 4.12 shows that Fe_5Gd appears with a

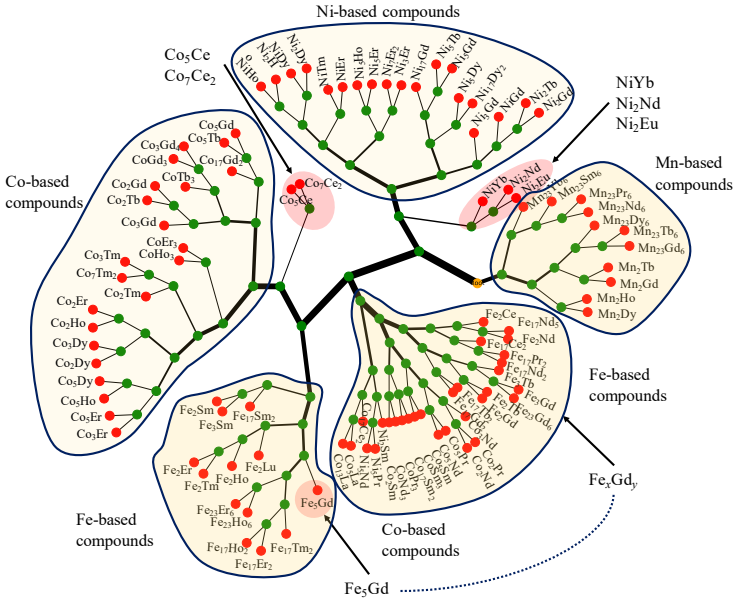


FIGURE 4.12: Hierarchical clustering model by utilizing information from dissimilarity matrix.

large number of dissimilarity values compared to other Fe-based alloys – especially the Fe_xGd_y group, indicating that Fe_5Gd is out of trend with its nearest neighbors. From Figure 4.13 (the compounds surrounded by blue line), it can be shown that, for an increasing concentration of rare-earth elements C_R, $Fe_{17}Gd_2$, Fe_5Gd, $Fe_{23}Gd_6$, Fe_3Gd, and Fe_2Gd, the T_C values are 479 K, 465 K, 659 K, 725 K, and 814 K, respectively. It is clear that, Fe_5Gd does not follow the general trend of Fe_xGd_y groups. Once again, the results illustrated that the information collected by the dissimilarity voting machine is potentially applied as a useful method for detecting anomalies.

4.3.3 Conclusion

In this work, a method for dissimilarity extracting between data instances by respecting to a given target variable is propose. The model is initially constructed by ensemble-bagging method with Kernel ridge regression as the predicting model (Chapter 3); multiple random subset sampling of the materials is performed to generate prediction models and corresponding contributions of the reference training materials in detail. The predicted value distribution is analyzed under the use of Gaussian mixture clustering method. The reference training materials contributed to the prediction model that accurately predicts the physical property value of a specific material, are considered to be similar to that material, or vice versa. To evaluate the proposed model, a prototype data set is used to show the intuitively meaning of dissimilarity between data instances. Next, the proposed model is applied for analyzing the Curie temperature (T_C) prediction of the binary $3d$ transition metal - $4f$ rare-earth binary alloys problem. As the results, our propose model shows a number of meaningful results, in considering to the physical meaning. To conclude, the proposed dissimilarity voting model could be considered as a potential tool for obtaining a deeper understanding of the data's structure, under the consideration to a given target property.

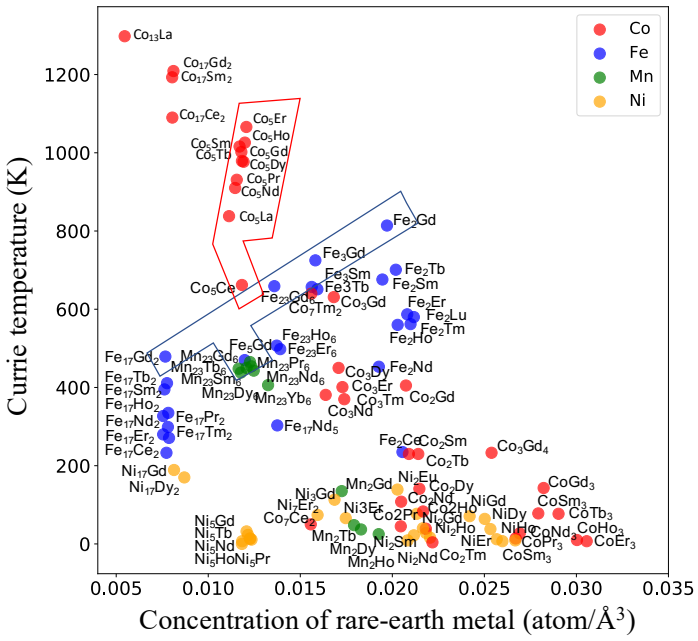


FIGURE 4.13: Relationship between T_C and the concentration of rare-earth element variable, C_R. The lines in red and blue show anomalies of Co_5Ce and Fe_5Gd

4.4 Combining evidence on similarity with Dempster–Shafer theory

Canonical similarity measurements in machine learning determine relationships among data objects over the description space to achieve efficient inferences. In contrast, in human recognition and other scientific fields, the similarity between two objects depends on whether they follow a common mechanism/function. In this Chapter, we propose an approach for measuring similarity–dissimilarity among data objects with respect to a given target variable by integrating evidences on their relationships. Multiple random subset sampling of regression functions is used to generate evidences of a common mechanism/function, which indicates similarity . The collected evidences are then combined within the framework of the Dempster–Shafer theory. The approach is evaluated using two prototype datasets, a simulated motorcycle accident with head acceleration measurements, and materials science data of physical properties of magnetic binary alloys, and the results show that similarity–dissimilarity can be effectively measured. Furthermore, the approach is applicable to unveiling mixtures of regression models, finding hidden laws, or detecting anomalies. We show that knowledge on each domain problem can be extended using the proposed similarity and dissimilarity information.

4.4.1 Similarity–dissimilarity evidence modeling

We consider a dataset $\mathcal{D}$ of p data instances. Assume that an instance with index i is described by an n-dimensional descriptive variable vector, $\boldsymbol{x}_i = (x_i^1, x_i^2, \ldots, x_i^n) \in \mathbb{R}^n$. The dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \ldots (\boldsymbol{x}_p, y_p)\}$ is then represented using a $(p \times (n+1))$

matrix. The target variable values of all data instances in the dataset are stored as a p-dimensional target vector $\boldsymbol{y} = (y_1, y_2 \dots y_p) \in \mathbb{R}^p$.

Similarity–dissimilarity modeling by mass function

We formulate a method to measure similarity among a pair of data points considering reference functions constructed using other points. We start with a definition of mass function. **definition** Let $\mathcal{X}$ be the universe set representing similarity states of any two data points $A = (\boldsymbol{x}_A, y_A)$ and $B = (\boldsymbol{x}_B, y_B)$, with reference to a given function f_r, $\mathcal{X} = \left\{s_{AB}^{f_r}, ds_{AB}^{f_r}\right\}$. All possible combinations states from $\mathcal{X}$ are given as:

$$2^{\mathcal{X}} = \left\{\emptyset, \left\{s_{AB}^{f_r}\right\}, \left\{ds_{AB}^{f_r}\right\}, \left\{s_{AB}^{f_r}, ds_{AB}^{f_r}\right\}\right\} \tag{4.3}$$

where $\left\{s_{AB}^{f_r}\right\}$, $\left\{ds_{AB}^{f_r}\right\}$ and $\left\{s_{AB}^{f_r}, ds_{AB}^{f_r}\right\}$ denote similarity, dissimilarity, and ambiguous states, respectively, between A and B. In other words, $\left\{s_{AB}^{f_r}, ds_{AB}^{f_r}\right\}$ shows *"unknown"* information about similarity and dissimilarity.

Using the bootstrap-aggregating algorithm Baldi and Sadowski, 2014; Hastie Trevor, 2009; Dietterich, 2000, we perform random subset samplings of the data instances to generate multiple reference function f_r models. For each sampling, we have two datasets: reference dataset, $\mathcal{D}_{ref}$, and evaluation dataset, $\mathcal{D}_{eval}$. These two datasets satisfy $\mathcal{D}_{ref} \cap \mathcal{D}_{eval} = \emptyset$ and $\mathcal{D}_{ref} \cup \mathcal{D}_{eval} = \mathcal{D}$. The reference function, f_r is regressed by a Gaussian process using the mean and covariance functions among data points.

According to the Dempster–Shafer theory, a mass function m to each element E in $2^{\mathcal{X}}$ is formulated:

$$m : 2^{\mathcal{X}} \to [0, 1] \text{ with } m(\emptyset) = 0 \text{ and } \sum_{E \in 2^{\mathcal{X}}} m(E) = 1 \tag{4.4}$$

Different sources of evidence could be used simultaneously to model m. In this paper, we propose two evidence sources with one from error-based reasoning and another adopted from a local observation in a previous work on a dissimilarity voting machine Nguyen et al., 2019.

Error-based source of evidences

By defining a function f_r as a reference to determine the degree of similarity, deviations from observation points to f_r are used as a source of evidence. For each f_r, the mass function m for all pairs of data points is modeled as follows.

Definition 1 The mass function value, $m\left(\left\{s_{AB}^{f_r}\right\}\right)$, of the similarity state between A and B under reference function f_r is defined by a likelihood product as follows :

$$m\left(\left\{s_{AB}^{f_r}\right\}\right) = p(A|f_r)p(B|f_r) \tag{4.5}$$

where $p(A|f_r)$ and $p(B|f_r)$ show the likelihood of observing points A and B for a given f_r, respectively.

Definition 2 The mass function value, $m\left(\left\{ds_{AB}^{f_r}\right\}\right)$, of the dissimilarity state between A and B under reference function f_r is defined relative to the deviation of

predictive distribution at query points σ_A and σ_B

$$m\left(\left\{ds_{AB}^{f_r}\right\}\right) = p(A|f_r)p(B|f_r)\mathbb{1}_{dev_A^{fr}>\sigma_A \vee dev_B^{fr}>\sigma_B} \tag{4.6}$$

where $dev_{A,B}^{f_r} = |y_{A,B} - \hat{f}_{ref}(\boldsymbol{x}_{A,B})|$ is a deviation from true values of points $AandB$ to predicted values from f_r, respectively. Regarding the indicator function, $\mathbb{1}_{cond} = 0$ if $cond$ is False and $\mathbb{1}_{cond} = 1$ if $cond$ is True. Finally, the mass function for the ambiguous state is a complement value to the condition in Equation 4.4.

For all pairwise instances of A and B in $\mathcal{D}_{eval}$, the mass function evidences, m of similarity states are collected, as described in Section 4.4.1. In an error-based experimental setting, without losing any generality, $p(A|f_r)$ (or $p(B|f_r)$) is modeled by:

$$p(A|f_r) = p\left(y_A|\boldsymbol{x}_A, \mathcal{D}_{ref}\right) = 2\int_{dev_A^{f_r}}^{+\infty} \mathcal{N}\left(x|0,\sigma_A\right) dx \tag{4.7}$$

In other words, the probability of observing point A with the reference function f_r is a normal distribution $\mathcal{N}$ with its mean as a predicted value $\hat{f}_{ref}$ and its predictive variance, σ_A. This method of modeling $p(A|f_r)$ ensures that mass function values m satisfy all conditions in Equation 4.4. Details of proof are shown in Supplemental material.

Local-based source of evidences

We adopted an existing dissimilarity voting machine from Nguyen et al., 2019. Rather than using the likelihood measurement to reference functions, as in the previous section, appearances/absences of the reference data points in D_{ref} are counted as source evidences for dissimilarity information with respect to other data points in D_{eval}. The dissimilarity value between points A and B incrementally increases if two following conditions are satisfied: (1) A and B are neighbors in the limit of a predefined number of neighbors k_{thres} and (2) A is not in D_{ref}, for which the corresponding $p(B|f_r) > \delta_{thres}$ with a predefined parameter δ_{thres} and vice versa. Converting to the Dempster–Shafer theory, the mass function between these two points A and B is defined as:

$$m\left(\left\{s_{AB}^{f_r}\right\}\right) = 0\,; m\left(\left\{ds_{AB}^{f_r}\right\}\right) = 1-u\,; m\left(\left\{s_{AB}^{f_r}\right\}, \left\{ds_{AB}^{f_r}\right\}\right) = u \tag{4.8}$$

The parameter u heuristically indicates the ambiguous level of the evidence. In all of experiments in this work, δ_{thres} is set as 0.5 and u is set as 0.9.

4.4.2 Dempster's rule in combining evidences

According to the Dempster–Shafer theory, with multiple mass functions $\{m_1, m_2 \ldots m_n\}$ either collected from single or multiple sources of evidence, a combining function from is calculated as follows:

$$\left(m_1 \oplus m_2 \oplus \ldots m_n \oplus\right)(E) = \frac{\sum\limits_{X_1 \cap X_2 \cap \cdots \cap X_n = E} m_1(X_1) \cdot m_2(X_2) \cdot \cdots \cdot m_n(X_n))}{\sum\limits_{X_1 \cap X_2 \cap \cdots \cap X_n \neq \emptyset} m_1(X_1) \cdot m_2(X_2) \cdot \cdots \cdot m_n(X_n))} \tag{4.9}$$

Finally, by combining multiple evidences, the proposed method estimates similarity–dissimilarity among any pairwise instance of data points. The similarity information

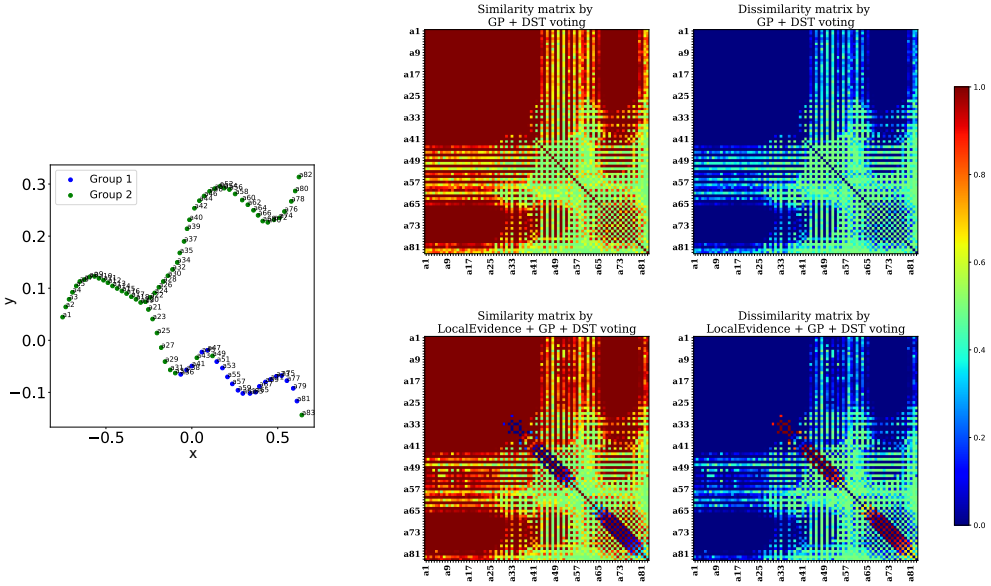


FIGURE 4.14: Left: source data and hierarchical clustering result collected using the extracted dissimilarity matrix. Two bifurcated branches are unveiled. Right: similarity and dissimilarity matrices constructed from combining error-based evidences (upper panel), as well as error-based and local-based evidences (lower panel).

could be considered a new layer of data extracted from the original data. The collected dissimilarity matrix is applicable as a distance matrix for use in clustering methods as hierarchical clustering or directly applicable to the multidimensional scaling method Kruskal, 1964; Borg, 1997; Mead, 1992.

4.4.3 Contribution

Our work contributes to the field of study in the following aspects: (1) we propose a concept of similarity measurement between data instances considering the appearance of combinations of other data instances; (2) we model the mentioned concept in the framework of the Dempster–Shafer theory by designing appropriate mass function and two sources of similarity evidences (3) through experiments, the extracted similarity-dissimilarity information shows a great potential for application to various purposes (unveiling mixtures of regression models, finding hidden laws or detecting anomalies).

4.4.4 Experiments

Experiment 1: Bifurcate data

We simulate a dataset containing 83 data points with a one-dimensional descriptive variable x and a target variable y, as shown in Figure 4.14 (left). Each data point is denoted by index a_i with i varying from 1 to 83 corresponding to increasing x values. The dataset is designed as a mixture of three main functions. In the range of x lesser than -0.2, the function $y = f(x)$ is monotonic and centered at 0.1. In the range of x greater than -0.2, the function f is bifurcated with a branch fluctuation increasing from 0.1 to 0.3 and the other variation decreasing from 0.1 to -0.15.

Figure 4.14 (right) shows the similarity and dissimilarity matrix collected by applying the Dempster's rule in combining evidences. Red cells in the similarity matrices show pairs associated with high similarity values and vice versa. Two upper and two lower matrices show results collected from the error-based evidences and from both the error-based and local-based evidences, respectively. In all matrices, similarity and dissimilarity values among data points from a_1 to a_{27} are homogeneous. In other words, these data points lie on the same function. Otherwise, data points from a_{28} to a_{83} show high dissimilarity values to its two closest neighbors, and high similarity values to the second closest neighbors, etc. This leads to the appearance of checkerboard patterns in these matrices. By adding the local evidences, the dissimilarity information about neighbors is emphasized, i.e., addition of blue/red checkerboard patterns in the lower similarity/dissimilarity matrix, respectively. The patterns are consistent with the arrangement of the source data.

Figure 4.14 shows a clustering result using the collected dissimilarity matrix as a distance matrix to the hierarchical clustering algorithm. Two bifurcate branches are separated with maximum resolution at data instance a_{31} with the use of local evidences. Table 4.2 shows comparison of prediction accuracy through coefficient determination R^2 score Kvalseth, 1985 and mean absolute error MAE of single model and unveiled mixture models in this dataset.

Experiment 2: Motorcycle data

We use a motorcycle dataset from "Motor cycle dataset", which are derived from a study on the effect of protective helmets in motorcycle accidents G., R., and F., 1981; Silverman, 1985. The dataset contains 133 measurements of head acceleration (in g) at the time (in ms) after impact in simulated motorcycle accidents. This dataset is widely used to demonstrate the effect of various methods, such as in Rasmussen and Ghahramani, 2002; Meeds and Osindero, 2006; Souza and Heckman, 2014; Silverman, 1985. This dataset is initially considered as non-stationary and input-dependent noise Rasmussen and Ghahramani, 2002, which is shown through the time-dependency variance of acceleration in Figure 4.15 (left). However, since the data is collected from multiple accident sources, the existence of multiple generated mechanisms rather than an individual mechanism can be assumed. These hidden mechanisms could be overlapped in certain time stamps and show distinguishable features in other time stamps. For this reason, we apply the combination of evidences using the Dempster–Shafer theory for screening the similarity among any pairwise measurement events rather than a hard assumption of the existence of a fixed number of regression lines in the dataset.

Figure 4.15 (right) shows the similarity and dissimilarity matrices among data points. The order of data points in these matrices follows the order of the accident time. In general, two models mostly overlap in the initial phase (for $t < 15ms$) and the ending phase (for $t > 45ms$) of accident time. High similarity values and no dissimilarity values among all data points in the initial and ending phase indicate the overlapping of models. For the middle phase of accident, $15ms \leq t \leq 45ms$, dissimilar patterns could be observed, i.e., blue and red patterns in the similarity and dissimilarity matrices, respectively.

Using the extracted dissimilarity matrix in the hierarchical clustering algorithm, we recognized two mechanisms, denoted by blue and red accordingly in Figure 4.15. The point with the largest difference between two mechanisms is focused on the highest variance region, within the time range $30ms \leq t \leq 40ms$. As the first group,

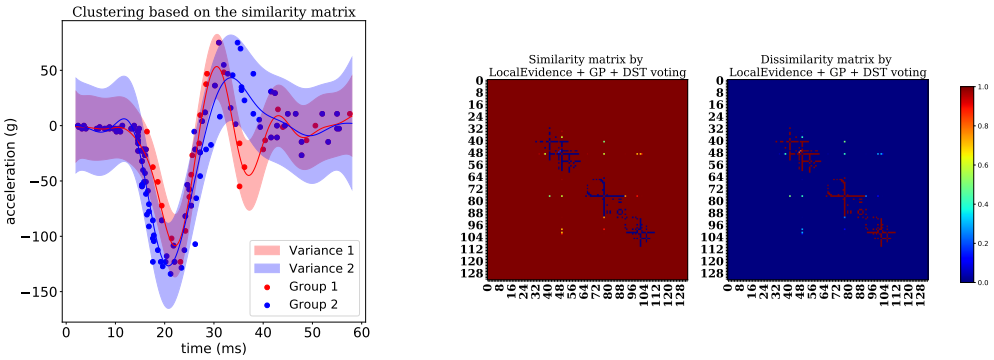


FIGURE 4.15: Left: Original data points and two hidden mechanisms of motorcycle accident types. For the middle phase of accident, $15ms \leq t \leq 45ms$, the first type (red) is associated with 1.5 damping oscillation period and the second type (blue) is associated with only one damping oscillation period. Right: Similarity and dissimilarity matrix constructed from combining error-based and local-based evidences.

TABLE 4.2: Prediction accuracy of unveiling mixture of regression

	All		Mixture of regression	
Name	R^2	MAE	R^2	MAE
Bifurcate data	-0.029 ± 0.02	0.113 ± 0.001	0.60 ± 0.023 0.98 ± 0.02	0.04 ± 0.001 0.002 ± 0.001
Motorcycle data	0.759 ± 0.014	17.54 ± 0.42	0.823 ± 0.003 0.845 ± 0.002	10.85 ± 0.19 14.70 ± 0.21

the red regression line shows an accident type with 1.5 period of acceleration oscillation, and as the second group, the blue regression line shows another accident type with one period of acceleration oscillation. In fact, the observation data reflect a damping oscillation of helmets in accidents. Therefore, the extracted regression lines with cosine forms and graduated decrease in amplitude show reasonable meaning compared with previous works Lázaro-Gredilla, Vaerenbergh, and Lawrence, 2012; Souza and Heckman, 2014. Table 4.2 shows that the prediction accuracy of the unveiled mixture models could be improved over the single model using this dataset.

Experiment 3: Noisy data

We also apply the similarity voting machine to two dimensional synthesized noisy datasets. The dataset contains 100 data points generated from a cosine function and a number of random background data points. The level of randomness, bg is sampled from 10% to 90% of total number of the cosine lines.

Figure 4.17 shows the source datasets (upper panel) with the noise level for three cases $bg - 30\%, 50\%, and 70\%$. Owing to space limitations, Figure 4.16 shows only the similarity and dissimilarity matrices constructed from combining the error-based and local-based evidences with the Dempster–Shafer theory applied to $bg - 50\%$ noisy dataset. In these matrices, pattern points with prefix a are all similar. In contrast, noise background points with prefix bg show high dissimilarity among all other data points.

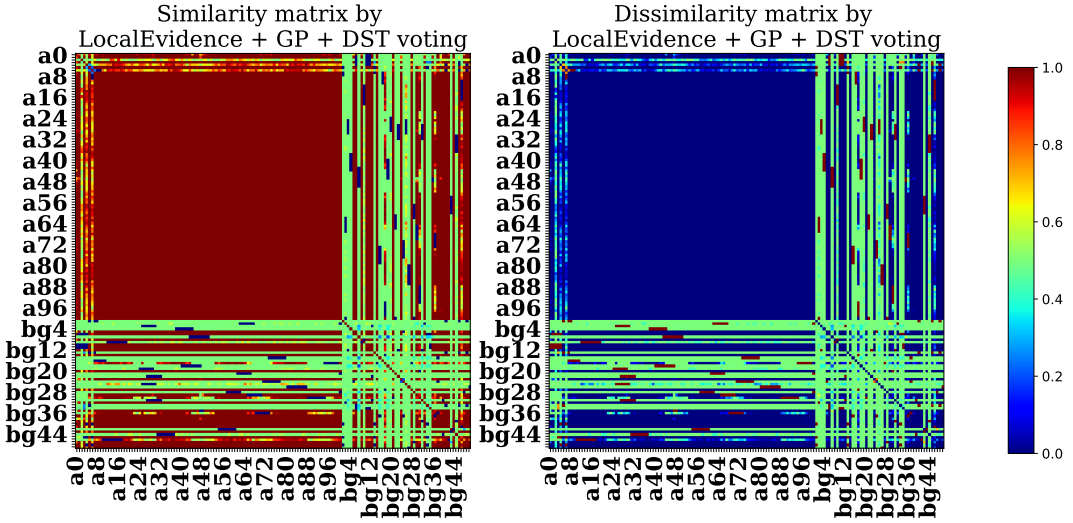


FIGURE 4.16: Similarity and dissimilarity matrix constructed from combining the error-based and local-based evidences to a noisy dataset with $bg - 50\%$. The pattern and background points are labeled with prefixes a and bg, respectively.

TABLE 4.3: Evaluation of noise removal ability using similarity–dissimilarity information

	Noise level	20%	30%	40%	50%	60%	70%	80%	90%
Pattern	Precision	0.95	0.90	0.85	0.78	0.76	0.72	0.66	0.60
	Recall	1.0	1.0	1.0	0.98	1.0	1.0	0.95	0.89
	f1-score	0.98	0.95	0.92	0.87	0.86	0.84	0.78	0.71
Background	Precision	1.0	1.0	1.0	0.92	1.0	1.00	0.86	0.73
	Recall	0.75	0.63	0.57	0.44	0.47	0.44	0.4	0.33
	f1-score	0.86	0.78	0.73	0.59	0.64	0.61	0.55	0.46

The lower panel in Figure 4.17 shows data points projecting the extracted dissimilarity matrix to an abstract 2D dimension through the multidimensional scaling algorithm Kruskal, 1964; Borg, 1997; Mead, 1992, respect to noisy level $bg - 30\%, 50\%, 70\%$. Overall, the cosine pattern and random noise data points are clearly separated into distinct groups. With increased noisy level, more background data points become close into the group of pattern points. The predicted labels of the pattern and background are assigned according to labels of dominant elements in each group. Table 4.3 shows the order of precision and recall in estimating the patterns and noise background. With increasing noise level from 20% to 90%, the precision of identifying the cosine patterns decreases from 0.95 to 0.60 gradually. On the other hand, the precision of background identification remains accurate up to the noise level of 80%. Regarding the recall score, the lowest value for the cosine pattern group is 0.89 with 90% noise, and the highest value of the background group is 0.75 with 20% noise. According to the table, our model uncovered at least 89% points of the cosine pattern. In conclusion, the extracted similarity information show potential in identifying outliers for removing noise in datasets.

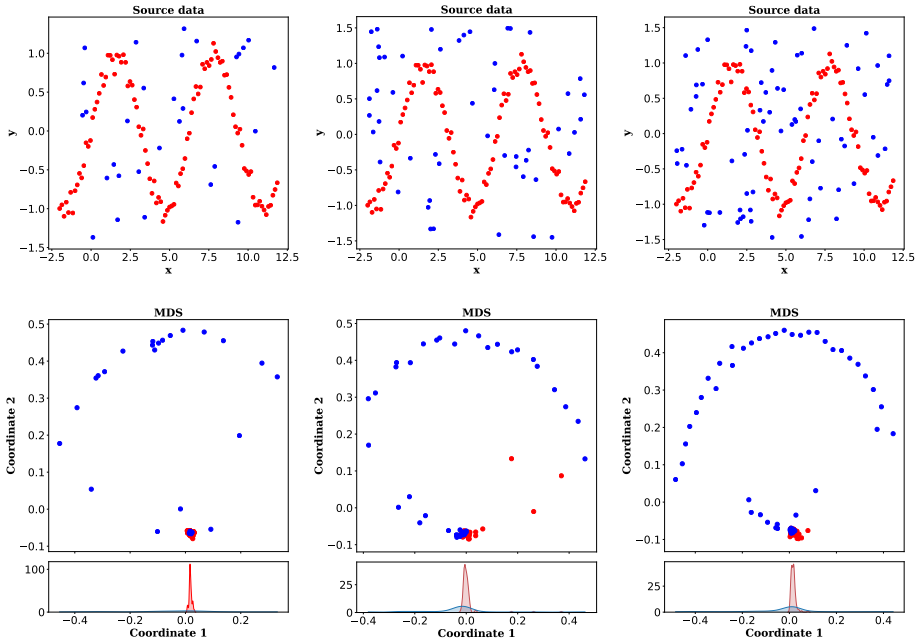


FIGURE 4.17: From left to right, upper panel: Synthesized noisy data with $bg - 30\%, 50\%, 70\%$ respectively with cosine patterns (red) and random noise points (blue). Lower panel: projection of extracted similarity matrices into a new dimensional space using the multidimensional scaling method and distribution on the first coordinate with red and blue colors consistent with source data.

Experiment 4: Curie temperature of rare-earth–transition metal alloys data set

We use the Curie temperature data sets of binary transition-rare earth binary compounds as a test bed for this experiment. A variable combination, $\chi_R, \chi_T, J_{4f}\left(1 - g_j\right)$, Z_T, r_{covT}, IP_T, S_{3d}, L_{3d}, J_{3d}, C_R , and parameters $\alpha = 5.12 * 10^{-5}, \gamma = 0.461$ from Nguyen et al., 2019 are used to build the Gaussian process model as the main configuration for our work. The model offers a high prediction accuracy with R^2 score of 0.942 ± 0.005 and MAE of 25.8 ± 1.9(K).

Figure 4.18 shows the similarity matrix with almost high similarity values among data points (red cells). In other words, almost all data points lie on a relatively smooth function, which is consistent with the high prediction accuracy of the model. However, five points are dissimilar to the others (blue cells): $Fe_{17}Nd_5$, Fe_5Gd, Co_7Er, Co_5Sm, Ni_2Eu.

Further investigation of the map of compounds and the Curie temperature, Figure 4.19, would help in better understanding these compounds. In this figure, from left to right, all compounds are visualized under transition metal based criteria for manganese, iron, cobalt, and nickel based families. Five extracted compounds are labeled, and compounds with the same structural group are indicated by the same marker style. Firstly, compounds $Fe_{17}Nd_5$, Fe_5Gd, andCo_7Er appear in the data space without any neighbors having the same structural group. In other words, the descriptive space surrounding these compounds is relatively sparse. Secondly, for all families, gadolinium based compounds are all in the local maximum. However, Co_5Sm, Ni_2Eu with its local maxima are not bound by this rule. From the observation, we claim that the similarity information extracted using our method is reasonable, and the method can be applied to future studies in materials science to accelerate the finding of outliers — new hidden knowledge in the dataset.

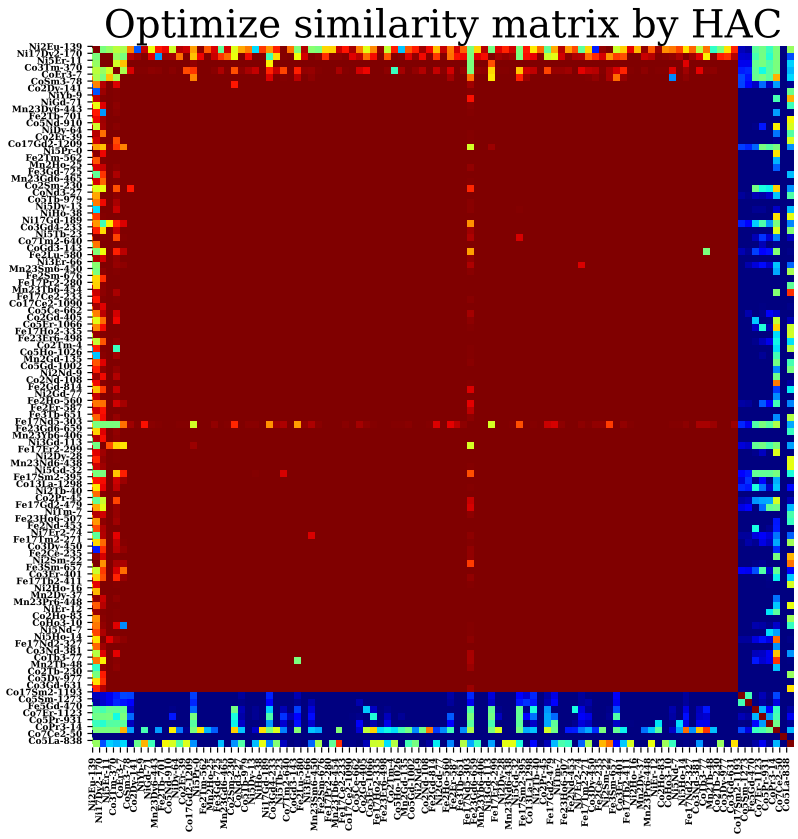


FIGURE 4.18: Similarity matrix extracted through the combination of error-based evidences with the Dempster–Shafer theory

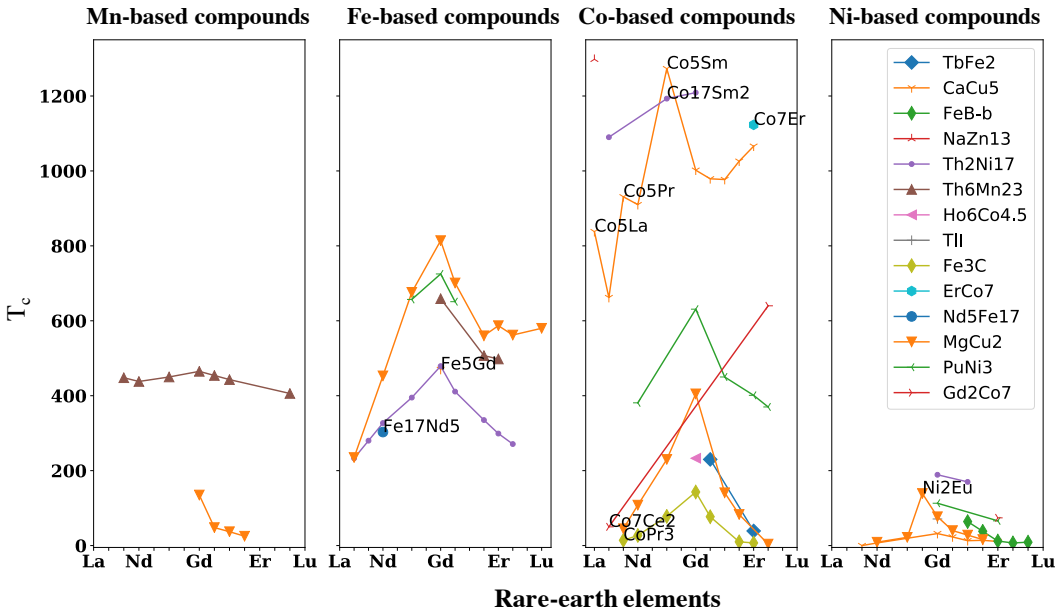


FIGURE 4.19: Outlier compounds (with labels) extracted from the similarity matrix.

4.4.5 Conclusion

In this work, the similarity concept is modeled by integrating evidence about their relationships according to the Dempster's rule. Multiple sampling of regressions is used to generate evidence about similarity, in which data points on the same function are considered to be similar. The extracted information from similarity–dissimilarity matrices appear to be useful in multiple tasks, such as unveiling hidden laws, identifying mixtures of regression models or detecting anomalies. The method is evaluated using two prototype data sets, simulated accident data and materials science data considering the Curie temperature. The collected results extend existing knowledge on problems related to the applied data sets.

Chapter 5

Contributions and limitations of the thesis

The thesis was conducted by proposing the use of machine learning methods in doing research in materials science. The central topic of the thesis is the regression-based similarity concept, in which the similarity among materials is measured with considering appearance of the function (linear or non-linear) created by other materials. By proposing this concept, the thesis provides a unified viewpoint for various scientific domains: cognitive science, psychology, materials science, and explainable machine learning.

An overview about effect of breakthrough discovery in science, phylosophical viewpoint about similarity are discussed in 1. Chapter 2 represents viewpoints of materials scientist in applying machine learning algorithms and machine learning experts with discussion about similarity modeling by machine and similarity understanding by human. The central meaning of similarity is represented in machine learning language by regression function and discussed in detail in3. Three developed similarity methods, heuristic similarity–dissimilarity voting machines and combining similarity evidence by Dempster-Shafer theory shown in 4. Other works related to grouping similar behavior catalyst domains shown in Chapter B.

Hereby, the thesis shows several contributions as follows:

The thesis brings the similarity terminology frequently used in convention human cognition, especially in physics to developed machine learning models. It helps the extracted results by the models are highly interpretable the meaning rather than canonical models.

Developed voting machines, e.g for similarity or dissimilarity measurement by using appropriated parametric and non-parametric models. These models could be used in the future research simutanously for better understanding about the real structure of data.

The method of utilizing Dempster-Shafer in combining pieces of evidence about the similarity and dissimilarity among data instances open a new view of connecting theory of evidence to similarity measurement. The method provides a systematic viewpoint of measuring similarity just by designing different source of evidence.

With the performance in using similarity measurement machine, the thesis shows that the extracted similarity information could be seem as a new layer of data. This kind of data shows convenient to use for multiple purposes as identifying mixture models, reducing noise or detecting outliers.

Beside that, the thesis exists a number of limitations, which are consider for future improvements.

Similarity and dissimilarity voting machine are heuristic models. Even though all machine learning models are inductive reasoning with pre-defined assumptions,

the model's assumptions and implementations should be revised with higher logical inference. For example, the assumption about non-contribution neighbor objects counted as dissimilarity information could seem as a corollary of similarity definition and it is not a completeness statement.

Even successful in showing meaning information for prototype model, all developed similarity measure required a screening step of variable evaluation firstly to select a variable combination set that relatively derive high prediction accuracy model. The model to examine similarity information from arbitrary design of description variable combination has not yet developed.

In the use of canonical machine learning models as Gaussian mixture model of clustering, the design, selection and interpretable of important variables has been relied mostly by human experts.

Chapter 6

Publication list

List of published/accepted publications

1. **Nguyen Duong Nguyen**, Pham Tien Lam, Nguyen Viet Cuong, Ho Tuan Dung, Tran Truyen, Takahashi K and Dam Hieu Chi, Committee machine that votes for similarity between materials, IUCrJ, 5, 830–840, 2018

2. Tien-Lam Pham, **Nguyen-Duong Nguyen**, Van-Doan Nguyen, Hiori Kino, Takashi Miyake, and Hieu-Chi Dam, Learning structure-property relationship in crystalline materials: A study of lanthanide transition metal alloys, The Journal of Chemical Physics, 148, 20, 204106, 10, 2018

3. **Nguyen Duong Nguyen** and Dam Hieu Chi, A regression-based model evaluation of the Curie temperature of transition-metal rare-earth compounds, Journal of Physics: Conference Series (JPCS), (accepted)

4. Makoto Hirose, Nozomu Ishiguro, Kei Shimomura, **Duong Nguyen Nguyen**, Hirosuke Matsui, Hieu Dam, Mizuki Tada, and Yukio Takahashi, Oxygen-diffusion-driven oxidation behavior and tracking areas visualized by X-ray spectro-ptychography with unsupervised learning, Communications Chemistry, volume 2, Article number: 50 (2019)

5. **Nguyen, Duong-Nguyen**; Pham, Tien-Lam; Nguyen, Viet-Cuong; Kino, Hiori; Miyake, Takashi; DAM, Hieu-Chi, Ensemble learning reveals dissimilarity between rare-earth transition binary alloys with respect to the Curie temperature, Journal of Physics: Materials, Special Issue Article

6. Yuanyuan Tan, Hirosuke Matsui, Nozomu Ishiguro, Tomoya Uruga, **Duong-Nguyen Nguyen**, Oki Sekizawa, Tomohiro Sakata, Naoyuki Maejima, Kotaro Higashi, Hieu Chi Dam, and Mizuki Tada Three-dimensional Catalyst Degradation Maps of Pt3Co/C Cathode Catalyst in Polymer Electrolyte Fuel Cell, Journal of Physical Chemistry C, 123, 18844-18853, 2019

List of submission paper

1. Minh-Quyet Ha, **Duong-Nguyen Nguyen**, Van-Doan Nguyen, Truyen Tran, Hieu-Chi Dam, Combining Evidence on Similarity with Dempster–Shafer theory, (in submission)

List of presentations

1. **Duong-Nguyen Nguyen**, Tien-Lam Pham, Viet-Cuong Nguyen, Hiori Kino, Takashi Miyake, Hieu-Chi Dam, Ensemble learning reveals insights into the

mechanism of physical properties of materials, Computational Sciences Workshop (CSW), organized by Computational Design of Advanced Functional Materials (CD-FMat) and The National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Jan. 16 - Jan. 19, 2019, (poster)

2. **Duong-Nguyen Nguyen**, Tien-Lam Pham, Viet-Cuong Nguyen, Hieu-Chi Dam, Committee machine that votes for similarity between materials, XXX International Conference on Computational Physics (CCP2018), University of California, Davis, July 29 - August 2, 2018 (presentation)

3. **Duong-Nguyen Nguyen**, Tien-Lam Pham, Viet-Cuong Nguyen, Hieu-Chi Dam, Committee machine that votes for similarity between materials, Asian Consortium on Computational Materials Science Theme Meeting on "Multi-scale Modelling of Materials for Sustainable Development" (ACCMS-TM 2018) Sept 7th to Sept 9th, 2018 Hanoi, Vietnam. (presentation)

Appendix A

Appendix for combining similarity evidence work

Proof:

The mass function values need to satisfy non-negative and normalization conditions shown in equation 2. We define the mass function values for similarity and dissimilarity states in equations 3, 4. Therefore, we prove the non-negative condition of the ambiguous state as follows.

If $\mathbb{1}_{dev_A^{f_r}>\sigma_A \vee dev_B^{f_r}>\sigma_B} = 0$, or equivalent, $dev_A^{f_r} \leq \sigma_A$ and $dev_B^{f_r} \leq \sigma_B$. From equation 5, we have $p(A|f_r) \leq 1$ and $p(B|f_r) \leq 1$. From equations 3 and 4, the mass function values of similarity state, $m\left(\left\{s_{AB}^{f_r}\right\}\right) = p(A|f_r)p(B|f_r) \leq 1$ and the mass function values of dissimilarity state $m\left(\left\{ds_{AB}^{f_r}\right\}\right) = 0$. Therefore, the mass function values of ambiguous state $m\left(\left\{s_{AB}^{f_r}, ds_{AB}^{f_r}\right\}\right) \geq 0$.

If $\mathbb{1}_{dev_A^{f_r}>\sigma_A \vee dev_B^{f_r}>\sigma_B} = 1$ or equivalent, either $dev_A^{f_r} > \sigma_A$ or $dev_B^{f_r} > \sigma_B$, or both are true. Without loss of generality, we consider the case $dev_A^{f_r} > \sigma_A$. From equation 3 and 4, the mass function values of similarity state $m\left(\left\{s_{AB}^{f_r}\right\}\right) = p(A|f_r)p(B|f_r) < 2\int_{\sigma_A^{f_r}}^{+\infty} \mathcal{N}(x|0,\sigma_A)\,dx \approx 0.32$. In this case, $m\left(\left\{ds_{AB}^{f_r}\right\}\right) = m\left(\left\{s_{AB}^{f_r}\right\}\right) < 0.32$. Therefore, the mass function values of ambiguous state, $m\left(\left\{s_{AB}^{f_r}, ds_{AB}^{f_r}\right\}\right) = 1 - 2m\left(\left\{s_{AB}^{f_r}\right\}\right) > 0.36$.

Appendix B

Joint distribution context in grouping similarity

B.1 Introduction

In this chapter, we show another connection from intuitive similarity meaning in human cognition to canonical machine learning algorithms. In the previous chapter, 4, the central representation of the context in similarity terminology is the relation function (regression). However, in this chapter, the context meaning is the appearance of Gaussian distribution in some representation space.

A cartoon in B.1 is used to better represent the idea. The left sub-figure in figure B.1 shows three data points A, B and C on two dimensional space x_1 and x_2. The conventional distances, e.g Euclidean among three points denoted with d_1 as distance measurement between A and B; d_2 as distance measurement between A and C. As an illustration in the figure, $d_1 < d_2$. The very first intuitive conversion to similarity meaning is that A is similar to B than C. However, in considering to other data points (blue), one easily notices that A and C should be all generated by a simple Gaussian distribution and B is generated by another Gaussian distribution. In that case, we often claim that A is similar to C than B. The method to identify these distributions could be processed by a simple machine learning algorithm called Gaussian mixture model Murphy, 2012a.

The most difficult parts here is shown through the work of define data instances in each problem, enrich the appropriated features then finding the potential space that exist a mixture of effects. In the following, two case studies regarding analyzing chemical imaging experiment are used to show how these data mining works effect materials science.

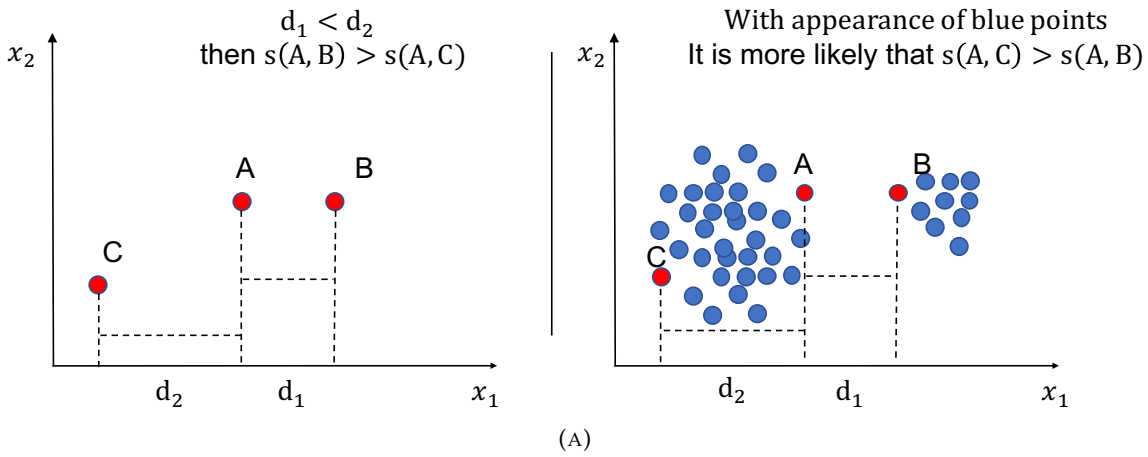
B.2 The first case study: Gaussian mixture model in unveiling oxygen diffusion track

B.2.1 Introduction

This work was published in Communications Chemistry, Hirose et al., 2019.

B.2.2 Oxygen storage and release mechanism

Three-way exhaust catalysis is a primary reacting process for all most all automobile systems G. et al., 2008; S., L., and J., 2013. The oxides with mixing cerium component have been used in very common to support exhausting process of catalysts. The



(A)

FIGURE B.1: Similarity between data points A, B, C in considering distribution of other points.

oxygen storage and release capacity (OSC) of cerium-containing mixed oxides cooperated with the reversible/unpredictable oxidation and reduction of Ce^{3+} and Ce^{4+} ions. These process is potential to use for widening of our understand of operation of three-way catalysts C. and Y., 1984; J., P., and Graziani, 1999; S, 2004. In particular, $Ce_2Zr_2O_x$ (denoted CZ-x, where $x = 7 - 8$) solid solutions with an ordered arrangement of Ce and Zr atoms exhibit remarkable OSCs A. et al., 2002; Urban et al., 2017. The dynamic structural changes of CZ-x compounds with oxygen diffusion in the bulk during the redox reaction have been investigated using X-ray diffraction Sasaki et al., 2004; SASAKI et al., 2003, neutron diffraction Achary et al., 2009, time-resolved X-ray absorption fine structure (XAFS) "Origin and Dynamics of Oxygen Storage/Release in a Pt/Ordered CeO2–ZrO2 Catalyst Studied by Time-Resolved XAFS Analysis", and theoretical calculations F et al., 2009; A. et al., 2017. However, the reversible oxygen storage and release processes erase the oxygen diffusion track in the bulk of the CZ-x particles, and consequently the details of the oxygen storage pathways in the CZ-x particles remain unclear.

B.2.3 Experiment setting

Detailed information about experimental setting and data sample preparation are described in our published manuscript: Makoto Hirose, Nozomu Ishiguro, Kei Shimomura, **Duong Nguyen Nguyen**, Hirosuke Matsui, Hieu Dam, Mizuki Tada, and Yukio Takahashi, Oxygen-diffusion-driven oxidation behavior and tracking areas visualized by X-ray spectro-ptychography with unsupervised learning, Communications Chemistry, volume 2, Article number: 50 (2019)

B.2.4 Problem setting in data science

Data instance definition and feature enrichment

The result of 3D HXSP imaging process is shown through a 3D nano-scale of Cerium valence value map of six solid particles with $452 \times 450 \times 136$ voxels in total. From Data science viewpoint, one needs to clarify objects: data instances representation variable and target variable.

About the data point definition, we have three possible cases to designate data instance:

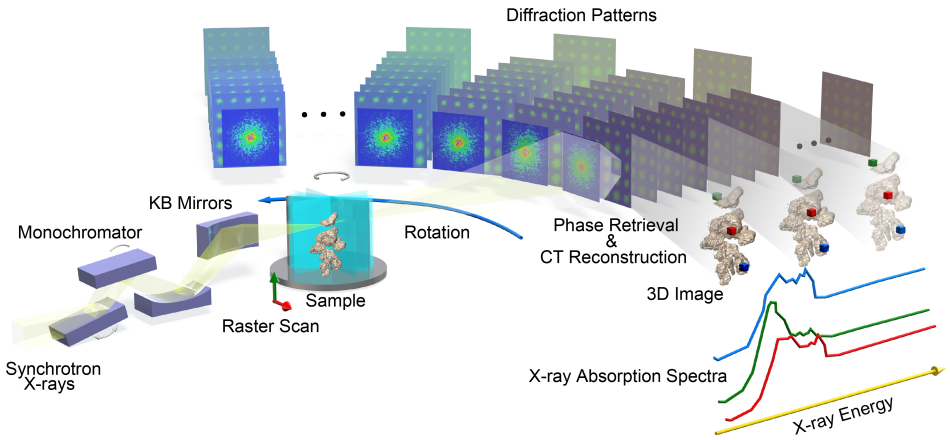


FIGURE B.2: Schematic representation of 3D HXSP. Synchrotron X-rays are monochromatized using a Si(111) double-crystal monochromator. The monochromatic X-rays are two-dimensionally focused using a pair of KB mirrors. A sample placed at the focal plane is laterally scanned across the illumination field. Coherent X-ray diffraction patterns are collected as a function of both the incident X-ray energy and angle. The projected amplitude and phase images at each angle and each energy are reconstructed, followed by 3D image reconstruction

- The whole 3D image as an individual data instance. Since there is only a single of the 3D image, if we choose the whole image, we cannot perform any operator.
- Each 2D images as an individual data instance, i.e there are 136 data instances, and each instance's dimension is 452×450. If that is the case, before discussing about the method we represent each 2D image, the expected similarity measurement operator will measure how this image is identical or dissimilar to others. However, these images represent Cerium value distribution for each layer of distinguishing particles. It seems difficult to interpret the chemical meaning of comparing one slide of particles to another slide of particles.
- Each particle as an individual data instance, i.e there are six data instances. From the experiment setting, all particles have no correlation for any pairwise. In other words, the expected similarity among particles potentially shows comparison about distribution over geometrical shape as Ce valence distribution on the surface / inside the bulk, volume, roughness/smoothness, etc. The expected in grouping similarity instances – similarity among data points is unreasonable.
- Each voxel as an individual data instance, i.e there are 27,662,400 Ce valence data instances. In this circumstance, one can perform clustering methods to grouping potential voxels with expected hidden "correlation" among them.

Finally we consider each data voxel as a data instance. The definition gives a data set of 27,662,400 Ce valence data instances for the six CZ-x particles for mining of the patterns of Ce valence in the particles during the oxygen storage process. In the original description, any voxel $\mathbf{v}$ associates with seven dimensions vector representation.

$$\mathbf{v} = \{x, y, z, Val_{Ce}(x, y, z), m(x, y, z), sd(x, y, z), r_{surf}(x, y, z), \kappa(x, y, z)\} \tag{B.1}$$

Among these features, $Val_{Ce}(x, y, z)$, valence value of Cerium is the only one that represent chemical property of the catalyst oxidation process. Without any target property, the process used to find correlations between parameters in the visualized 3D maps is only unsupervised learning. To find the correlation–function between geometrical shape and Val_{Ce}, it neccessary to investigate the differential of the Val_{Ce} to local shape. For the reason, to characterize each voxel, we considered the surrounding binning of $42 \times 42 \times 42\text{nm}^3$ ($3 \times 3 \times 3$ voxels) in the 3D map and used the local mean ($m(x, y, z)$) and local standard deviation ($sd(x, y, z)$) of the Ce valence state in each binning domain as descriptors. The variable $m(x, y, z)$ corresponds to the degree of oxygen storage ($Ce^{3+} \rightarrow Ce^{4+}$) and $sd(x, y, z)$ corresponds to the variation of oxygen storage in the local domains. Using these descriptors, we observed a volcano-type correlation between $m(x, y, z)$ and $sd(x, y, z)$, as shown in Figure B.2. Finally, to estimate the effect of oxidation process which effect from the outer surface into the bulk of particles, feature $r_{surf}(x, y, z)$, the distance from any voxel to its closest surface and the curvature to each position $\kappa(x, y, z)$, if the voxel locates on the surface are determined.

Gaussian mixture model in finding similarity evidences

In this step, an approach of imaging process with 3D hard X-ray spectro ptychography (HXSP) coupled with unsupervised learning was proposed to achieve the 3D nanoscale chemical imaging of heterogeneous reaction events in bulk solid materials. The 3D HXSP method gives us a chance to realize not only the simple 3D nanoscale imaging of the structure but also a map of valence state inside individual Pt/CZ-x solid solution particles during the oxygen storage process.

As a starting assumption, we assume the observed images describing partly exhausted catalyst particles. Therefore, the oxidation storage and release process are also naturally deduced to perform through several distinct phases. In our assumption, all reaction phases represent different mechanism in comparing together. Accordingly, all of reaction phases left different evidences to the imaged catalyst particles. In the most simple understanding, the chemical evidences of oxygen storage phase that left inside imaged particles should be totally dissimilar with the completed releasing oxygen region.

Concerning to regions whose the same chemical reaction phase, people observe data points associated with the same distribution and vice versa; a group of data points that shows a best fitting to a single distribution is likely to be generated under the same mechanism. From these assumptions, we perform Gaussian mixture model target to group similar data voxels generated from the same Gaussian distribution function.

The most important question after setting the problem is from which variables we can observe the evidences of the reaction. The evidence could be found in a single or multiple dimension of observation. With the number of dimensions larger than three, it is more complicate to visualize and understand obtained results. Therefore, we investigate the possible mixture effects for any pairwise of variables shown in variable preparation .

For any pairwise of variables, a voxel with index i in $\mathcal{D}$ is described by 2D vector $\mathbf{x}_i = (a_i, b_i)$, therefore the dataset $\mathcal{D}$ is represented using a $(n \times 2)$ matrix. The distribution was approximated by a mixture models of K Gaussian distributions as following:

$$p(\mathbf{x}_i|\theta) = \sum_{(k=1)}^{K} \pi_k \mathcal{N}(\mu_k, \sigma), \tag{B.2}$$

where π, μ, σ were weights, centers, and coefficient matrices for the 2D Gaussians. For a given number of mixture components K, the estimation for parameter is conducted through an Expectation-Maximization algorithm with detailed explanation in Murphy, 2012a. To determine the number of mixture component, a maximizing Bayesian information criterion G., 1978 process is utilized by applying several different trial to randomize initial states.

Finally, unsupervised data mining with Gaussian mixture models of the visualized 3D nanoscale chemical maps then successfully revealed the concealed heterogeneous oxygen- diffusion-driven 3D nanoscale Ce oxidation tracking areas inside the individual mixed-oxide particles during the oxygen storage process. The pairwise variables $m(x,y,z) - sd(x,y,z)$ are used to extract the mixture chemical reaction phases. Details results are shown in B.2.5.

B.2.5 Results

Result 1: 3D HXSP nanoscale imaging of Ce valence state

Detailed information about ordinary imaging analysis are described in our published manuscript: Makoto Hirose, Nozomu Ishiguro, Kei Shimomura, **Duong Nguyen Nguyen**, Hirosuke Matsui, Hieu Dam, Mizuki Tada, and Yukio Takahashi, Oxygen-diffusion-driven oxidation behavior and tracking areas visualized by X-ray spectroptychography with unsupervised learning, Communications Chemistry, volume 2, Article number: 50 (2019)

Result 2: Unveiling four reaction phases in the 3D nanoscale valence map

The distribution in the 2D $m(x,y,z) - sd(x,y,z)$ plot could not be represented by an ordinary distribution function. One can easily to recognize that the distribution shows a mixture of separated components and each components are seem to represent for different physical mechanisms. For this reason, we model the distribution of $m(x,y,z) - sd(x,y,z)$ approximately by a mixture of K Gaussian distribution components. To accurately determine the position of each component, an expectation–maximization algorithm is used to estimate necessary parameters, i.e the mean, variance and weight of each Gaussian component. The Bayesian information criterion is minimize in searching for the best value of K with different a number of K trials.

It should be noted that the Gaussian mixture model, which approximate the 2D plot of (m, sd) breaking down the actual distribution of the total observation data set $\mathcal{D}$. The entire distribution is then a mixture of four Gaussian components denoted G_1, G_2, G_3, and G_4 with respect to the set of the following centers $\mu = \{(3.43, 0.143), (3.61, 0.265), (3.85, 0.224), and (3.97, 0.063)\}$ and covariance matrices $\Sigma = [(0.02766, -0.00078), (-0.00078, 0.00221)], [(-0.02076, -0.00351), (-0.00351, 0.00837)], [(0.00478, -0.00441), (-0.00441, 0.00729)]$, and $[(0.00057, -0.00114), (-0.00114, 0.00252)]$, respectively (Fig. B.5a).

Taking into account that the horizontal axis $m(x,y,z)$ represent to the degree of oxygen storage, the distribution component in red G_1 corresponds to the domains where oxygen storage did not proceed for the most part but brought about a distribution of Cerium valence states with the highest probability around the CZ-7.5

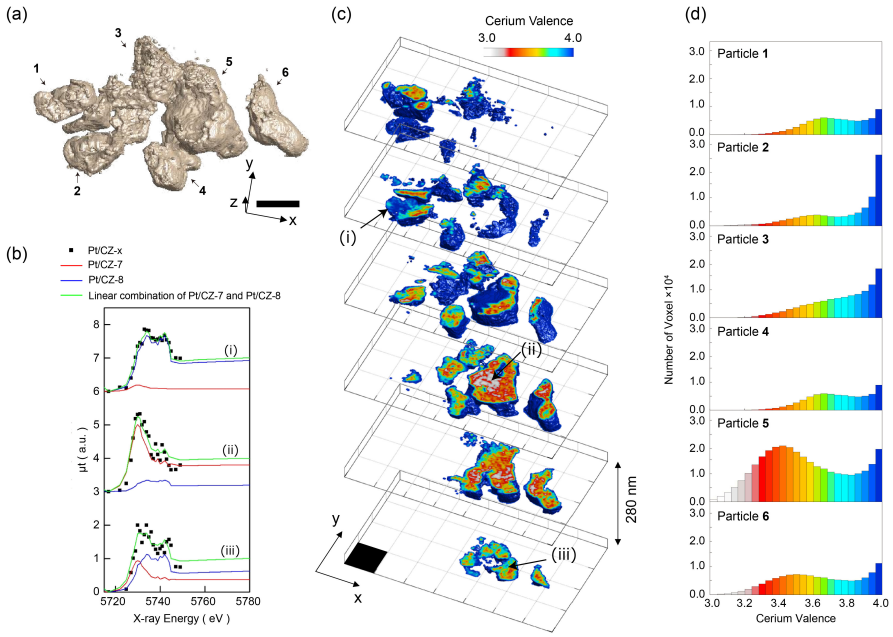
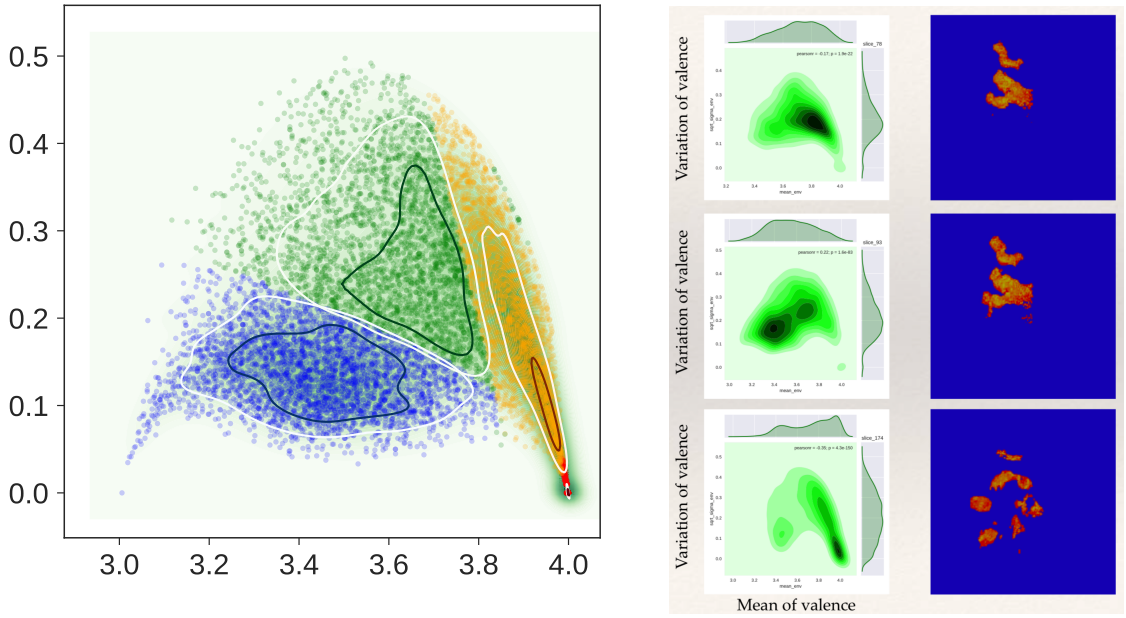


FIGURE B.3: 3D mapping and XAFS analysis. a Isosurface rendering of the reconstructed 3D phase map of partially oxidized Pt/CZ particles. The scale bar is 700 nm. b Three-dimensionally resolved XAFS spectra and their fitted spectra. The green spectra are the results of the linear combination of the XAFS spectra of Pt/CZ-7 (red) and Pt/CZ-8 (blue) normalized at the isosbestic point of 5.7697 keV. The black dots (i), (ii), and (iii) are the XAFS signals extracted from the $56 \times 56 \times 56\text{nm}^3$ volumes indicated at (i), (ii), and (iii) in c. c Series of slices of the 3D Ce valence image along the z direction. The black square represents $700x700\text{nm}^2$. d Ce valence distributions for the number of voxels of the particles labeled in a

phase during the oxygen storage process at 423 K for 1 h. The distribution component in orange G_2 corresponds to the domains where oxygen storage proceeded beyond the CZ-7.5 phase but accompany with a larger data variation. The distribution component in green G_3 corresponds to the domains where oxygen storage converged to the final state of CZ-8 containing Ce^{4+}. The distribution component in blue G_4 shows to the domains where oxygen storage was almost complete and displayed the smallest standard-deviation/variance to the neighbor. For these meaning, the 2D scattering plot of the mean Ce valence (m) and its standard deviation (*sd*) in the $42 \times 42 \times 42\text{nm}^3$ ($3 \times 3 \times 3$ voxels) domains of partially oxidized Pt/CZ-x particles has represented a volcano-type pattern as shown in Fig. B.5a. This type of pattern is related to the course of the oxygen storage process during Cerium oxidation from Ce^{3+} to Ce^{4+} via the G_1, G_2 and G_3 domains in the CZ-x particles concealed in the bulk. The Cerium oxidation in the G_1 domains proceeds in the CZ-7 phase with a pyrochlore structure, whereas the Cerium oxidation in the G_2 domains around $+3.61$ is regarded to occur in the disordered (mixed) phases accompanied by the transformation of the pyrochlore phase to the CZ-8 κ-phase with a fluorite structure and showing a larger *sd*. Further oxygen storage proceeds in the Cerium oxidation states above $+3.7$ and forms the G_3 domains with a maximum Cerium valence population around $+3.85$. The G_3 domains, considered nearly a fluorite phase, readily converge to the Ce^{4+} valence state in the G_4 domains and the final CZ-8 phase.

We represent the locations of the four groups (G_1, G_2, G_3, and G_4) in (x, y, z) real



(A) (B)

FIGURE B.4: (Scatter plot of mean Cerium valence (m) and its standard deviation (sd) for $42 \times 42 \times 42\text{nm}^3$ ($3 \times 3 \times 3$ voxels) domains of partially oxidized Pt/CZ-x particles, and classification of correlation trends using a Gaussian mixture clustering method. Figure (a): investigating mixture effect on whole data points, (b) investigation on different particles.

space and present a 3D map in Fig. B.5b, and cross-sectional images in Fig. B.5c. These figures shows apparently the dependence of the four groups's locations to the morphological characteristics of the particles. Any voxel regions belonging to G_4 (blue) are observed at the outermost surfaces of the particles. In the most contrast, any voxel regions belonging to G_1 (red) mostly locate in the particles's cores. Similar differences in the 3D Ce valence images of the particles were observed in Fig. B.3c. Recently, we reported five different types of correlations between the Ce density and Ce valence (positive, negative, quasi-constant to Ce density, quasi-constant to Ce valence, and no correlation) in 2D HXSP images M et al., 2018; the G_1 group can be related to the positive correlation between Ce density and Ce valence that was observed for Ce valences lower than +3.5 and near the centers of the particles, whereas the G_3 and G_4 groups can be related to the negative correlation that was observed for Ce valences of +3.5 to +4.0 and around the surfaces of the particles in the 2D images. However, the five different types of oxidation behavior in the local domains of the CZ-x particles described in the previous report M et al., 2018 were determined from the correlations in 2D images averaged over the entire depth direction of the local domains (along the optical axis), where the 2D data for the local domains may be merged with and obfuscated by the data at minor heterogeneous sites of the particles, such as boundaries, defects, and interfaces, although the 2D HXSP image analysis successfully revealed the 2D distribution of Ce oxidation states inside the catalyst particles26. The current 3D HXSP image rendering is the first report of the 3D visualization of the nanoscale oxidation tracking areas in CZ-x catalyst particles during the oxygen storage process.

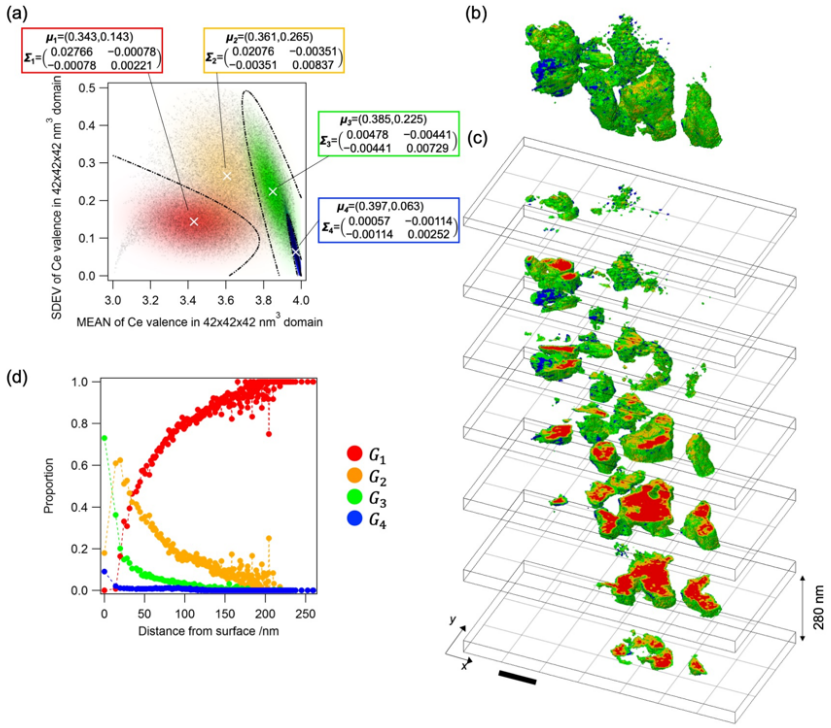


FIGURE B.5: The 3D map of Cerium valence by using unsupervised learning. (a) Joint distribution of mean Ce valence (m) and its standard deviation (sd) for $42 \times 42 \times 42\text{nm}^3$ ($3 \times 3 \times 3$ voxels) domains of partially oxidized Pt/CZ-x particles, and unveiling components by using a Gaussian mixture model. Red-G_1; orange-G_2; green-G_3; blue-G_4. The Gaussian centers are denoted by crosses in white for $G_1 - G_4$; $\mu_k, k = 1 - 4$: Gaussian center; Σ_k: covariance matrix. (b) The distribution in 3D rendering of the four component groups between m and sd for $42 \times 42 \times 42\text{nm}^3$ ($3 \times 3 \times 3$ voxels) regrions of partially oxidized Pt/CZ-x particles, and (c) series of slices showing the 3D distributions of the four correlation groups along the z direction. The scale bar for (b) and (c) is 700 nm. (d) The dependence of the proportion of each group in c respected to the distance from the particle surface.

In Figure B.5d, we investigate the dependence of these four unveiling groups to the distance from the surface. The group G_1 associated with limited oxidation does not appear with the depth smaller than 20 nm of the outer surfaces of the particles. However, its proportion greatly increase toward the bulk of the particles. The G_2 domains exhibited a maximum proportion at 20 ± 17 nm, which gradually decreased over 200 nm toward the bulk. Considering the stable surfaces of the fluorite/pyrochlore structures (e.g., (111) or (110)) P., R., and P, 2016; "Advanced electron microscopy investigation of Ceria–Zirconia-based catalysts", the maximum proportion (20 nm) of the group G_2 is equivalent to with approximate 40–50 oxygen vacancy sites in depth from the surface of the particles. The group G_3 with considerable oxidation are located at the surface regions of the particles and their fraction showed a rapid exponential decrease up to 25 nm followed by a gentle decrease over 100 nm toward the core. The group G_4 with a valence of $+4.0$ were located at the surface layer (smaller than 20 nm). Even though, it should be noted that these are a minor component and the surface region was composed of greater fractions of the G_3 and G_2 domains. These results demonstrate that achieving the complete oxidation

of $Ce3+$ to Ce^{4+} throughout the entire surface region is difficult during the oxygen storage process at 423K for1h.

Detailed information about physical interpretation are described in our published manuscript: Makoto Hirose, Nozomu Ishiguro, Kei Shimomura, **Duong Nguyen Nguyen**, Hirosuke Matsui, Hieu Dam, Mizuki Tada, and Yukio Takahashi, Oxygen-diffusion-driven oxidation behavior and tracking areas visualized by X-ray spectroptychography with unsupervised learning, Communications Chemistry, volume 2, Article number: 50 (2019)

B.2.6 Conclusion

The imaging experiment of 3D XSP in the hard X-ray region provide success tool for visualizing the 3D distribution of the oxidation states of Cerium in Pt/CZ-x exhaust catalyst particles with a 3D sampling pitch of 14 nm. Data mining method with Gaussian mixture model in analyzing these 3D images of the Cerium valence unveiled four main groups that represent to the morphological characteristics and ability of local chemical reaction of oxygen storage process. These analyses elucidate the oxidation pathways happening in the solid catalyst. The 3D HXSP imaging technique is expected to be an indispensable tool for determining reaction tracking areas and the relationships between the structure and function of heterogeneous functional materials. In particular, in next-generation synchrotron facilities where fluxes with much higher coherence will be achieved, the present approach will be applied to in situ 3D measurements. It is under a very high expectation for significant accelerating the progress in chemistry and materials science.

B.3 The second case study: Gaussian mixture model in understanding catalyst degradation process

B.3.1 Introduction

Catalyst degradation at the cathode in a membrane electrode assembly (MEA) remains a critical issue for practical operation of the polymer electrolyte fuel cell (PEFC). However, the PEFC wet system prevents the visualization of detailed events of the degradation of the cathode catalyst in a functioning cell. Operando spectro imaging (computer tomography by XANES) clearly visualized the 3D images of the morphology, the Pt and Co distributions, the Co / Pt atomic ratio and the Pt valence state of a Pt-Co / C cathodic catalyst in a PEFC MEA before and after the accelerated degradation test (ADT) PEFC for the first time. The visualized 3D images demonstrated the degradation behavior of the cathode catalyst with different ways of degradation of Pt and Co in the bimetallic catalyst. The infographic approach combining the spectro-imaging operation and the unsupervised 3D image learning has shown the degradation behavior of the catalyst with different ways of degradation of Pt and Co in the bimetallic catalyst and the local parts without degradation of the catalyst in the MEA.

B.3.2 Pt-Co catalysts in Polymer electrolyte fuel cells

Sustainable and fossil-free ways have long evolved in terms of energy production and creating a favorable environment for the future. Polymer electrolyte fuel cells (PEFC) show a great advantage in replacing current fossil fuels with clean energy, without carbon byproducts. However, current PEFC systems pose critical problems

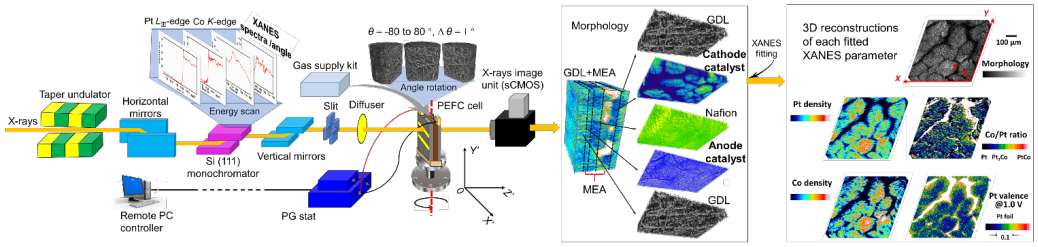


FIGURE B.6: Operating diagram of XANES-CT imaging under PEFC operating conditions and reconstituted 3D maps of the cathode catalyst layer in an MEA.

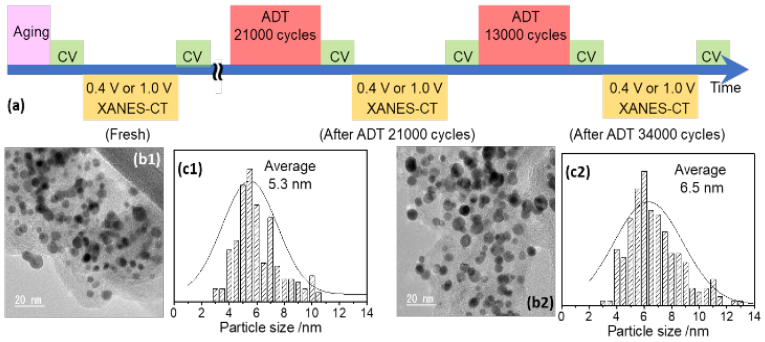


FIGURE B.7: (a) The experimental sequence of XANES-CT operando imaging and PEFC operation. (b) TEM images of cathodic catalysts striped from MEAs (1) before ADT and (2) after ADT 34,000 cycles. (c) Granulometric distribution of the cathodic catalysts analyzed from the TEM images

for the activity of the oxygen reduction reaction (ORR) and the durability of cathodic electrocatalysts.

The catalysts Pt-Co Stamenkovic et al., 2007; Greeley et al., 2009; Bordiga et al., 2013; Dai et al., 2017b exhibit good performances of activity and durability than those of other standard Pt / C catalysts. The Pt-Co catalysts are widely known for be useful for PEFC cathodic catalysts.

The Pt3Co / C catalysts are more durable than the Pt / C catalysts, but it is still impossible to completely eliminate the undesirable degradation of the cathode catalyst under PEFC operating conditions. Ishiguro et al., 2016; Nikkuni et al., 2015 Dissolution and aggregation of Pt in Pt-Co / C catalysts are more inhibited than Pt / C catalysts, but Co gradually dissolves bimetallic catalysts losing the effects of alloy . There are many reports on the preparation of durable cathodic catalysts with a bimetallic structure, but to our knowledge there are no reports on PEFC cathodic catalysts that completely inhibit cathode degradation.

B.3.3 Experiment setting

Detailed information about experimental setting and data sample preparation are described in our published manuscript: Yuanyuan Tan, Hirosuke Matsui, Nozomu Ishiguro, Tomoya Uruga, **Duong-Nguyen Nguyen**, Oki Sekizawa, Tomohiro Sakata, Naoyuki Maejima, Kotaro Higashi, Hieu Chi Dam, and Mizuki Tada Three-dimensional Catalyst Degradation Maps of Pt3Co/C Cathode Catalyst in Polymer Electrolyte Fuel Cell, Journal of Physical Chemistry C, 123, 18844-18853, 2019

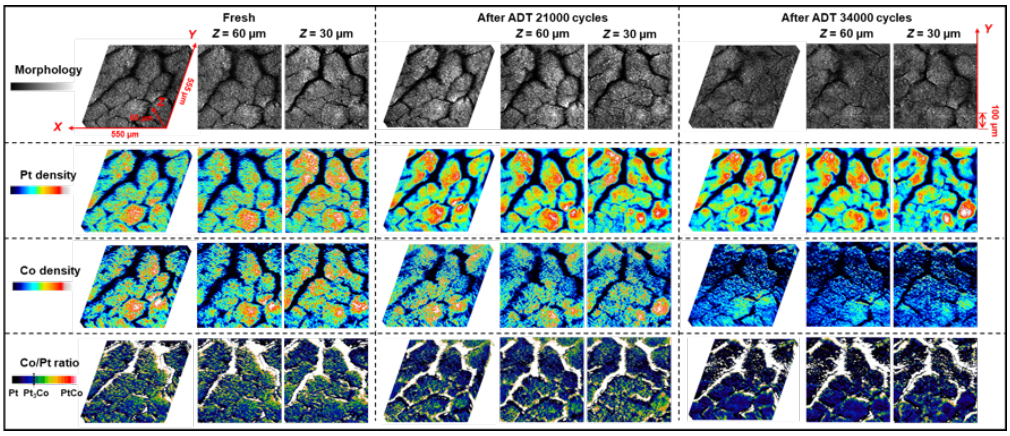


FIGURE B.8: 3D maps and cross-sectional images of the MEA with the Pt-Co/C cathode catalyst reconstructed by the operando PtL_{III}-edge and Co K-edge XANES-CT data. Morphology (μt at 11.497 keV before PtL_{III}-edge), Pt density (PtL_{III}-edge jump), Co density (Co K-edge jump), and Co/Pt ratio (1.17 × Co density/Pt density, calculated on the 3D images). Field of view: $X = 550, Y = 555,$ and$Z = 60$ m. Cross-sectional images: $Z = 60\mu m$ (interface between the cathode catalyst layer and the Nafion membrane), and $30\mu m$ (center of the cathode catalyst layer).

B.3.4 Problem setting in data science

From the viewpoint of data mining approach, the study of the degradation process of the Pt-Co cathode catalyst in MEA is equivalent to the problem of modeling distribution of descriptive variables – unsupervised learning Ghahramani, Luxburg, and U.; Ratsch, 2004 of this phenomena.

Firstly, we collect a dataset $\mathcal{D}$ contains $p = 550 \times 555 \times 60$ data instances that correspond to the voxels in the real device. Each data instance is described by $m = 15$ descriptive variables x which store information of the corresponding voxel. For each observation ADT states $t : (fresh, ADT21000, ADT34000)$. The descriptive variables **x** for a given data instance correspond to a voxel at (x,y,z) and time stamp t include: (1) Pt density ρ^t_{Pt}, (2) Co density ρ^t_{Co}, (3) Pt valence state at 1.0 V $val^t_{Pt-1.0}$, (4) Pt valence state at 0.4 V $val^t_{Pt-0.4}$, and (5) distance from the most closest surface $distance - from - surface^t$. In this study, $distance - from - surface^t$ for a given voxel is measured by Euclidean distance from the voxel to the closest voxel that has morphological value lower than 10^{-5}. To summary, any voxel **x** associates with seven dimensions vector representation.

$$\mathbf{x} = \{x^t, y^t, z^t, \rho^t_{Pt}, \rho^t_{Co}, val^t_{Pt-1.0}, val^t_{Pt-0.4}, distance - from - surface^t\} \tag{B.3}$$

In the problem, the cathode device of PEFC cell was captured by 60 slides of 2D cross-sectional images. Under this information, our assumption that the chemical behavior of catalyst particles depends on the order of cross-sectional images. Therefore, the work of grouping voxels into functional groups by Gaussian mixture model is conducted separately image by image.

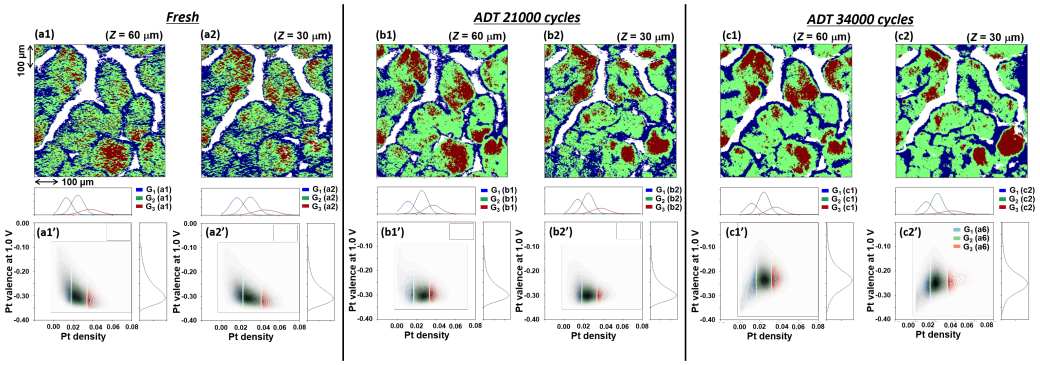


FIGURE B.9: (Bottom) Pearson diagrams of plate density and valence state from pt at 1.0 V to $Z = 60\mu m$ (1, Nation interface and cathodic catalyst layer) and at 30 μm (2: center of the cathode layer) catalyst layer). It has been suggested to divide the Gaussian model of each platinum density graph into three groups: G_1, G_2 and G_3 (in each package). (Top) Their distribution maps in the cross-sectional images (blue: G_1, green: G_2 and red: G_3). (a) New condition, (b) after ADT cycles 21000 and (c) after ADT 34000 cycles.

B.3.5 Results: Unsupervised learning of the visualized 3D maps of the Pt-Co catalyst in the MEA

The Pearson 2D plot of the Pt density and valence state of Pt at 1.0 V in Fig. 1 ref fig.PEFCFig4 indicated a correlation between the two parameters. The Pearson diagram had a straight shape, as shown in Figure ref fig.PEFCFig4 (a1 ′). We studied the Gaussian mixing model to estimate the number of mixing components in the plot, using a process of maximizing the score of Bayesian information criteria. It was found that the graph had three components (a1 ′) at platinum density (horizontal axis); G_1 (blue, low density of Pt), G_2 (green, average Pt density) and G_3 (red high platinum density). The quantities of the components are presented at the top of the Gaussian graphs (a1 ′) and the distribution maps of the three components are presented in figure ref fig.PEFCFig4 (a1). There were no significant differences between the two depths (Z = 60 and 30 mm), as shown in figure ref fig.PEFCFig4 (a). We also examined the Gaussian mixing models after the ADT cycles and three components (G_1, G_2 andG_3) were:: also suggested on the Pearson plots after 21000 and 34000 cycles of ADT (Figures ref fig.PEFCFig4 (b ′) and (c′)).

It should be noted that the forms of Pearson plots between the Pt density and the valence state of Pt at 1.0 V in Figure 4 have been largely modified by the ADT. The straight down form in the fresh state means that the dispersed portion (G_1 at low Pt density) was more oxidized than the aggregated portion (G_3 with high density of Pt). On the other han, the Pearson diagram with its shape after 21,000 cycles of TDA is almost flat (Figure ref fig.PEFCFig4 (b ′)) and left after 34,000 cycles of TDA (Figure ref fig. PEFCFig4 (c ′)). The left form means that G_1 with a low density of pt is smaller than G_3 with a high density of pt. In addition to the changes made to the shape of the plot, the average Pt valence is shifted upwards in the figure B.9 (c ′). These trends are closely correlate to the loss of the Co alloy effect to reduce the center of the PEP band (PEFC19, PEFC31, PEFC54, PEFC55, PEFC56, PEFC57) by the dissolution of Co of the Pt-Co catalyst by ADT. A similar analysis of the density of Co is presented in the figure B.10.

One could see estimation the geometric distance to the surface (cracks in the cathodic catalyst layer) using the visualized morphological image of the cathodic

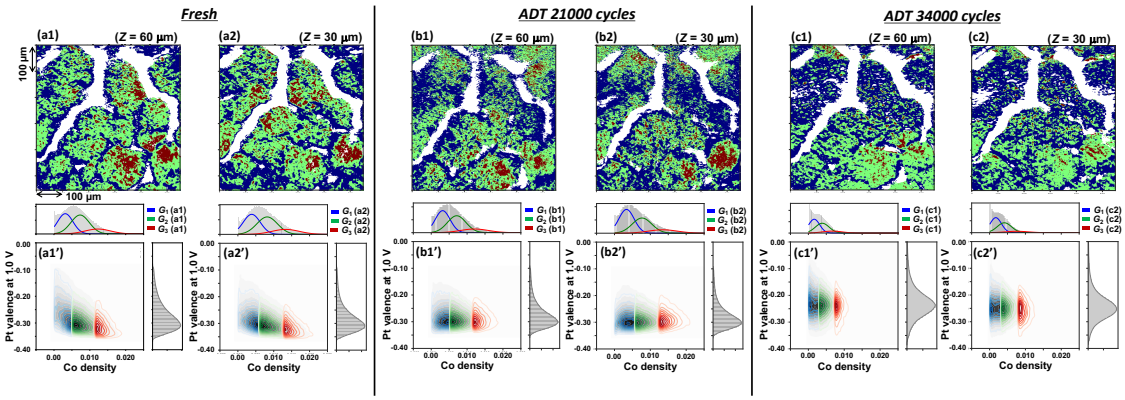


FIGURE B.10: (Bottom) Pearson plots of Co density and Pt valence state at 1.0 V at $Z = 60\mu m$ (1; interface of Nation and the cathode catalyst layer) and $30\mu m$ (2: center of the cathode catalyst layer). The Gaussian model of each Pt density plot was suggested that each Pt density plot was categorized to three groups of G_1, G_2, and G_3 (in each plot). (Top) Their distribution maps in the cross-sectional images (blue: G_1 , green: G_2, and red: G_3). (a) Fresh state, (b) after the ADT 21000 cycles, and (c) after the ADT 34000 cycles.

catalyst layer in Figure 1 B.8. The distance of the surface that measure from for a given voxel is defined by the Euclidean distance from the voxel to the nearest voxel having a morphological value less than 10^{-5} in the image of the morphology of the figure **??**. The figure B.11 shows the Pearson curves of the distance calculated from the surface (horizontal axis) and Δ_{Pt} or Δ_{Co}, which are defined as Pt or Co density differences between the two states (vertical axis) between the two ADT cycles (ADT $21,000 ADT$ 21,000 ADT $21,000 cycles ADT$ 34,000 cycles).

Similar unsupervised learning using the Gaussian mixture models show the present of four components (G_1, G_2, G_3, and G_4) in the Person plots between the distance from surface and ΔPt or ΔCo (Figure B.11 (a)). G_1 and G_2 were negative values, showing parts losing Pt or Co species by the ADT process, while G_4 (> 0) shows parts increasing Pt or Co species by the ADT process. G_3 around 0 means parts without changes in the Pt or Co quantity. The distributions of the four components (G_1, G_2, G_3, and G_4) are presented in Figure B.11 (b).

Note that components (G_1, G_2, and G_4) of Δ_{Pt} were localized from surface to bulk but G_3 (yellow) of Δ_{Pt} with negligible changes in the Pt density was remarkably localized at the surface part (distance from surface ≈ 0) as shown in Figure B.11 (a-Δ_{Pt}). All of results suggest that there are durable parts with negligible Pt loss around the crack structures in the cathode catalyst layer. The results suggest that the regulation of three-dimensional morphology inside the cathode catalyst layer is one of the key parameters to control the cathode catalyst degradation of the Pt-Co catalyst in the MEA. In contrast, all components of Δ_{Co} are randomly dispersed at all parts of the cathode catalyst layer (Figure B.11 (a-Δ_{Co1})).

B.3.6 Conclusion

The operando 3D XANES-CT imaging of the MEA under PEFC operating conditions and its unsupervised data mining successfully visualized the practical degradation of the Pt-Co/C catalyst in the MEA for the first time. The CT reconstruction of the structural parameters extracted from the Pt LIII-edge and Co K-edge XANES spectra provided clear 3D maps of metal locations and valence states of the cathode catalyst.

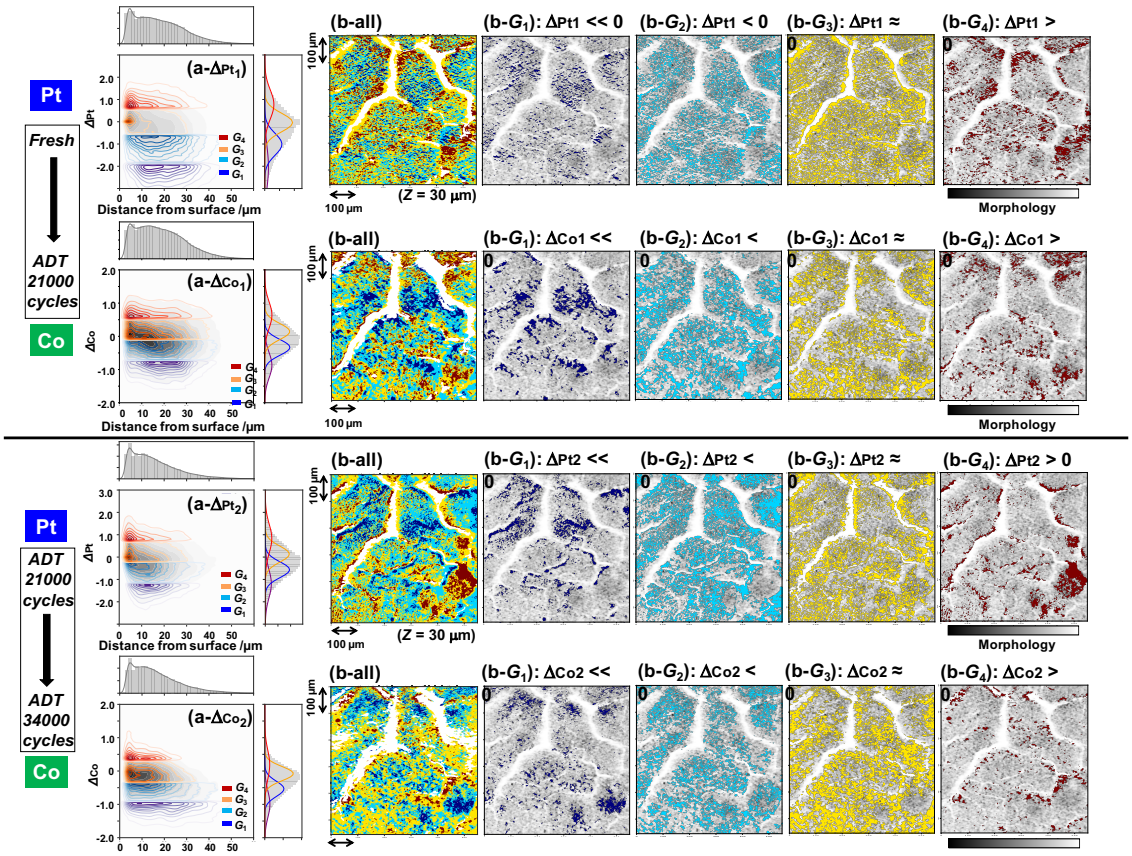


FIGURE B.11: (a) Pearson plots of the calculated distance from surface and Δ_{Pt} (difference in Pt density between two states) or Δ_{Co} (difference in Co density between two states) at $Z = 30\mu m$ (center of the cathode catalyst layer). Top: Fresh state $\rightarrow$ ADT 21000 cycles, bottom: ADT 21000 cycles $\rightarrow$ ADT 34000 cycles. The Gaussian model of each Δ_{Pt} or Δ_{Co} plot was suggested that each Δ_{Pt} or Δ_{Co} was categorized to four groups of G_1, G_2, G_3, and G_4 (in each plot). (b) Their distribution maps in the cross-sectional images overlaid on the morphology images (dark blue: G_1, sky: G_2, yellow: G_3, and red: G_4). (b-All) present all groups on the morphology images.

The different degradation manners of Pt and Co consequently brought about the heterogeneous degradation of the cathode catalyst in the MEA, highly depending on the structure (crack) of the carbon support. The infographic method combining the 3D chemical imaging and unsupervised learning would be a potential way to illustrate intrinsic events of experimental materials and devices.

B.4 Conclusion

In the aspect of applying data mining technics for finding meaningful results in Materials science, the works in this case study shows in detail how knowledge co-creation performed between human and machine. Grouping similar data points which generated by the same physical mechanism satisfies understanding and interpretation of Materials science community. The works: (1) define data instances, (2) enrich the appropriated features and (3) finding the potential space that exist a mixture of effects are the most critical points in requirement.

In the first case study, two extracted variables $m(x, y, z) and sd(x, y, z)$ are consequences of the need to investigate the behavior of mechanism – the function of observed data. The success of these features shows the way of setting similarity considering the mechanisms as the center is appropriate. If in Chapter 3, the sophisticated point relies on committee machine to identify similarity–dissimilarity relation, then in this chapter, the most sophisticated point relies on the design of these two variables.

Bibliography

A., Gupta et al. (2017). "Activation of oxygen in Ce2Zr2O7+x across pyrochlore to fluorite structural transformation: first-principles analysis". In: *J. Phys. Chem C* 121, pp. 1803–1808.

A., Suda et al. (2002). "Improvement of oxygen storage capacity of CeO2-ZrO2 solid solution by heat treatment in reducing atmosphere". In: *J. Ceram. Soc. Jpn* 110, pp. 126–130.

Achary, S. Nagabhusan et al. (2009). "Intercalation/deintercalation of oxygen: a sequential evolution of phases in Ce2O3/CeO2-ZrO2 pyrochlores". In: *Chem. Mater* 21, pp. 5848–5859. URL: https://doi.org/10.1021/cm902450q.

Ahluwalia, Rajesh K. et al. (2018). "Potential Dependence of Pt and Co Dissolution from Platinum-Cobalt Alloy PEFC Catalysts Using Time-Resolved Measurements". In: *J. Electrochem. Soc.* 165, F3024–F3035. DOI: doi:10.1149/2.0031806jes.

Almuallim, H. and T.G Dietterich (1991). "Learning with many irrelevant features". In: *The Ninth National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, pp. 547–552.

Anderson, A. B et al. (2005). In: *J. Phys. Chem. B* 109, p. 1198.

Angelo, Ziletti et al. (2018). "Insightful classification of crystal structures using deep learning". In: *Nature Communications* 1.9. DOI: 10.1038/s41467-018-05169-6. URL: https://doi.org/10.1038/s41467-018-05169-6.

Bach, Francis R. and Michael I. Jordan (2004). "Learning Spectral Clustering". In: *Advances in Neural Information Processing Systems 16*. Ed. by S. Thrun, L. K. Saul, and B. Schölkopf. MIT Press, pp. 305–312. URL: http://papers.nips.cc/paper/2388-learning-spectral-clustering.pdf.

Baldi, Pierre and Peter Sadowski (2014). "The dropout learning algorithm". In: *Artificial Intelligence* 210, pp. 78 –122. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2014.02.004. URL: http://www.sciencedirect.com/science/article/pii/S0004370214000216.

Behler, Jorg (2011). "Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations". In: *Physical Chemistry Chemical Physics* 13.40, pp. 17930–17955. DOI: 10.1039/C1CP21668F. URL: https://www.ncbi.nlm.nih.gov/pubmed/21915403.

Behler, Jorg and Michele Parrinello (2007). "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces". In: *Phys. Rev. Lett.* 98 (14), p. 146401. DOI: 10.1103/PhysRevLett.98.146401. URL: https://link.aps.org/doi/10.1103/PhysRevLett.98.146401.

Behler, Jörg (2016). "Perspective: Machine learning potentials for atomistic simulations". In: *The Journal of Chemical Physics* 145.17, p. 170901. DOI: 10.1063/1.4966192. eprint: https://doi.org/10.1063/1.4966192. URL: https://doi.org/10.1063/1.4966192.

BF, János Abonyi, ed. (2007). *Cluster Analysis for Data Mining and System Identification*. Springer.

Biesiada, Jacek and Wlodzisław Duch (2007). "Feature selection for high dimensional data - a Pearson redundancy based filter". In: *Computer Recognition Systems 2, Advances in Soft Computing, Springer, Berlin, Heidelberg* 45.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (2017). "Variational inference: a review for statisticians". In: *Journal of the American Statistical Association*. ISSN: 0162-1459.

Blum, Avrim L. and Pat Langley (1997). "Selection of relevant features and examples in machine learning". In: *Amficial Intelligence* 97.

Bordiga, Silvia et al. (2013). "Reactivity of Surface Species in Heterogeneous Catalysts Probed by In Situ X-ray Absorption Techniques". In: *Chemical Reviews*, p. 24396. URL: https://doi.org/10.1021/cr2000898.

Borg, I.; Groenen P. (1997). *Modern Multidimensional Scaling - Theory and Applications*. Springer Series in Statistics.

Botu, Venkatesh and Rampi Ramprasad (2014). "Adaptive machine learning framework to accelerate ab initio molecular dynamics". In: *Int. J. Quant. Chem.* 115.16, pp. 1074–1083.

Bu, Lingzheng et al. (2016). "Biaxially strained PtPb/Pt core/shell nanoplate boosts oxygen reduction catalysis". In: *Science* 354.6318, pp. 1410–1414. ISSN: 0036-8075. DOI: 10.1126/science.aah6133. eprint: https://science.sciencemag.org/content/354/6318/1410.full.pdf. URL: https://science.sciencemag.org/content/354/6318/1410.

C., Buurmans I. L. and Weckhuysen B. M (2012). "Heterogeneities of individual catalyst particles in space and time as monitored by spectroscopy". In: *Nat. Chem* 4, pp. 873–886.

C., Yao H. and Yao Y. F. Y. (1984). "Ceria in automotive exhaust catalysts". In: *J. Catal.* 86, pp. 254–265.

Carnegie Mellon University, statistic. "Motor cycle dataset". In: (). URL: http://www.stat.cmu.edu/~larry/all-of-statistics/=data/motor.dat?fbclid=IwAR0P-yF4HLqJoLO0ZEK-hNmAEEGcDxAQUg-kTYjWoCx1F3mqBc5QL8XpqpM.

Chancourtois, Beguyer de. "Table of the natural classification of elements, called the "telluric helix"". In: *Comptes Rendus de l'Académie des Sciences (in French)* 55.1862 (), 600–601.

Chen, Shuo et al. (2009). "Origin of Oxygen Reduction Reaction Activity on "Pt3Co"Nanoparticles: Atomically Resolved Chemical Compositions and Structures". In: *The Journal of Physical Chemistry C* 113, pp. 1109–1125. ISSN: 3. URL: https://doi.org/10.1021/jp807143e.

Christophe Biernacki GillesCeleux, GérardGovaertc (2003). "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models". In: *Computational Statistics and Data Analysis* 41.3. Recent Developments in Mixture Model, pp. 561 –575. ISSN: 0167-9473.

Colon-Mercado H. R.; Popov, B. N (2006). In: *J. Power Sources* 155, p. 253.

Corduneanu, Adrian and Christopher M. Bishop (2001). "Variational Bayesian model selection for mixture distributions". In: *Artificial Intelligence and Statistics*.

Cui, Yitao et al. "In Situ Hard X-ray Photoelectron Study of O2 and H2O Adsorption on Pt Nanoparticles". In: *The Journal of Physical Chemistry C* 120 (), pp. 10936–10940. ISSN: 20. URL: https://doi.org/10.1021/acs.jpcc.6b02402.

D., Mendeleev (1869). "Relationship of elements' properties to their atomic weights". In: *Journal of the Russian Chemical Society, (in Russian)* 1, 60–77. URL: https://babel.hathitrust.org/cgi/pt?id=mdp.39015065536586;view=1up;seq=70.

Dahlgaard, Soren, Mathias Baek Tejs Knudsen, and Mikkel Thorup (2017). "Practical Hash Functions for Similarity Estimation and Dimensionality Reduction". In:

Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6618–6628. URL: http://dl.acm.org/citation.cfm?id=3295222.3295407.

Dai, Sheng et al. (2017a). "In situ atomic-scale observation of oxygen-driven core-shell formation in Pt3Co nanoparticles (article)". In: *Nature Communications* 8.1. URL: https://doi.org/10.1038/s41467-017-00161-y.

Dai, Sheng et al. (2017b). "Revealing Surface Elemental Composition and Dynamic Processes Involved in Facet-Dependent Oxidation of Pt3Co Nanoparticles via in Situ Transmission Electron Microscopy". In: *Nano Letters* 17. URL: https://doi.org/10.1021/acs.nanolett.7b01325.

Dam, Hieu Chi et al. (2018). "Important Descriptors and Descriptor Groups of Curie Temperatures of Rare-earth Transition-metal Binary Alloys". In: *Journal of the Physical Society of Japan* 87.11, p. 113801. DOI: 10.7566/JPSJ.87.113801. eprint: https://doi.org/10.7566/JPSJ.87.113801. URL: https://doi.org/10.7566/JPSJ.87.113801.

David, Rogers and Hahn Mathew (2010). "Extended-Connectivity Fingerprints". In: *Journal of Chemical Information and Modeling* 50.5. PMID: 20426451, pp. 742–754. DOI: 10.1021/ci100050t. eprint: https://doi.org/10.1021/ci100050t. URL: https://doi.org/10.1021/ci100050t.

Davis, Jason V. et al. (2007). "Information- theoretic metric learning". In: *In Proc. of ICML*, 209–216.

Debe, Mark K. (2012). In: *Nature* 486.43, 43 EP. URL: https://doi.org/10.1038/nature11115.

Deudon, Michel (2018). "Learning semantic similarity in a continuous space". In: *Advances in Neural Information Processing Systems 31*, pp. 986–997. URL: http://papers.nips.cc/paper/7377-learning-semantic-similarity-in-a-continuous-space.pdf.

Dietterich, Thomas G. (2000). "Ensemble methods in machine Learning". In: *Proceedings of the First International Workshop on Multiple Classifier Systems*.

Domingos, Pedro, ed. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books.

Doshi-Velez, Finale and Been Kim (2017). "Towards A Rigorous Science of Interpretable Machine Learning". In: *arXiv*. URL: https://arxiv.org/abs/1702.08608.

Duangsoithong, Rakkrit and Terry Windeatt (2009). "Relevance and redundancy analysis for ensemble classifiers". In: *Machine learning and data mining in pattern recognition*. Ed. by Petra Perner. Springer Berlin Heidelberg, pp. 206–220.

Dubaua, L. et al. (2010). "Durability of Pt3Co/C Cathodes in a 16 Cell PEMFC Stack: Macro/Microstructural Changes and Degradation Mechanisms". In: *Journal of The Electrochemical Society* 157, B1887–B1895. ISSN: 12. DOI: doi:10.1149/1.3485104.

Döbereiner, J. W. (1829). "An attempt to group elementary substances according to their analogies". In: *Annalen der Physik und Chemie, 2nd series (in German)* 15, 301–307. URL: http://web.lemoyne.edu/~giunta/dobereiner.html.

Escaño, Mary Clare Sison and Hideaki Kasai (2014). "First-principles study on surface structure, thickness and composition dependence of the stability of Pt-skin/Pt3Co oxygen-reduction-reaction catalysts". In: *Journal of Power Sources* 247, pp. 562 – 571. ISSN: 0378-7753. DOI: https://doi.org/10.1016/j.jpowsour.2013.09.013. URL: http://www.sciencedirect.com/science/article/pii/S0378775313015048.

Eto, Riki et al. (2014). "Fully-Automatic Bayesian Piecewise Sparse Linear Models". In: *AISTATS*.

Everitt, Brian S. et al., eds. (2011). *Cluster analysis, Chapter 4, Hierarchical clustering*. Wiley Series in Probability and Statistics.

F., Meirer and Weckhuysen B. M (2018). "Spatial and temporal exploration of heterogeneous catalysts with synchrotron radiation". In: *Nat. Rev. Mater.* 3, pp. 324–340.

F, Wang H. et al. (2009). "Maximizing the localized relaxation: the origin of the outstanding oxygen storage capacity of k- Ce2Zr2O8." In: *Angew. Chem. Int. Ed* 48, pp. 8289–8292.

Fraley, Chris (2006). "Algorithms for model-based Gaussian hierarchical clustering". In: *SIAM J. Sci. Comput.* 20.1.

Fukunaga, K. and D.R. Olsen (Feb. 1971). "An algorithm for finding intrinsic dimensionality of data". In: *IEEE Transactions on Computers* C-20. ISSN: 0018-9340. DOI: https://doi.org/10.1109/T-C.1971.223208.

G., Ertl et al., eds. (2008). *J. Handbook of Heterogeneous Catalysis, 2nd edn*. Wiley-VCH, Weinheim.

G., Schmidt, Mattern R., and Schueler F. (1981). "Biomechanical investigationto determinephysical and traumatologicaldifferentiationcriteriafor the maximumload capacity of head and vertebralcolumn with and without protectivehelmet under the effectsof impact". In: *EEC Research Programon Biomechanicsof Impacts* Final report, Phase III, Project G5.

G., Schwarz (1978). "Estimating the dimension of a model". In: *Ann. Statist.* 6.1.

Gan G Ma C, Wu J (2007). "Data Clustering theory, Algorithms, and Applications". In: *ASASIAM Series on Statistics and Applied. Society for Industrial and Applied Mathematics*.

Gavetti G, Rivkin JW. (2005). "How strategists really think. Tapping the power of analogy." In: *Harvard Business Review* 83.4, pp. 54–63. URL: https://www.ncbi.nlm.nih.gov/pubmed/15807039.

Gentner, Dedre (2002). "Analogy in Scientific Discovery: The Case of Johannes Kepler". In: *Model-Based Reasoning: Science, Technology, Values*. Boston, MA: Springer US, pp. 21–39. ISBN: 978-1-4615-0605-8. DOI: 10.1007/978-1-4615-0605-8_2. URL: https://doi.org/10.1007/978-1-4615-0605-8_2.

Ghahramani, Z, O.; von Luxburg, and G U.; Ratsch (2004). "Unsupervised Learning." In: *Advanced Lectures on Machine Learning, vol. 3176 of Lecture Notes in Computer Science* vol. 3176 of Lecture Notes in Computer Science, pp. 72–112.

Ghiringhelli, Luca M. et al. (2015). "Big data of materials science - critical role of the descriptor". In: *Physical Review Letters* 114. ISSN: 105503.

Goldsmith, Bryan R et al. (2017). "Uncovering structure-property relationships of materials by subgroup discovery". In: *New Journal of Physics* 19.1, p. 013031. DOI: 10.1088/1367-2630/aa57c2. URL: https://doi.org/10.1088%2F1367-2630%2Faa57c2.

Greeley, J. et al. (2009). "Alloys of platinum and early transition metals as oxygen reduction electrocatalysts". In: *Nature Chemistry* 1, p. 552. URL: https://doi.org/10.1038/nchem.367.

Hanawa, Hirotaka et al. (2012). "In Situ ATR-FTIR Analysis of the Structure of Nafion–Pt/C and Nafion–Pt3Co/C Interfaces in Fuel Cell". In: *The Journal of Physical Chemistry C* 116.40, pp. 21401–21406. DOI: 10.1021/jp306955q. eprint: https://doi.org/10.1021/jp306955q. URL: https://doi.org/10.1021/jp306955q.

Hansen, Katja et al. (2013). "Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies". In: *Journal of Chemical Theory and Computation* 9.8. PMID: 26584096, pp. 3404–3419. DOI: 10.1021/ct400195d.

eprint: https://doi.org/10.1021/ct400195d. URL: https://doi.org/10.1021/ct400195d.

Hastie Trevor Tibshirani Robert, Friedman Jerome (2009). *The Elements of Statistical Learning*. Springer.

Hayashi, Kohei and Ryohei Fujimaki (2013). "Factorized Asymptotic Bayesian Inference for Latent Feature Models". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., pp. 1214–1222. URL: http://papers.nips.cc/paper/5171-factorized-asymptotic-bayesian-inference-for-latent-feature-models.pdf.

Hirayama, Y. et al. (2017). "Intrinsic hard magnetic properties of Sm(Fe1-xCox)12 compound with the ThMn12 structure". In: *Scripta Materialia* 138, pp. 62 –65. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2017.05.029. URL: http://www.sciencedirect.com/science/article/pii/S1359646217302737.

Hirose, Makoto et al. (2018). "Visualization of Heterogeneous Oxygen Storage Behavior in Platinum-Supported Cerium-Zirconium Oxide Three-Way Catalyst Particles by Hard X-ray Spectro-Ptychography". In: *Angewandte Chemie International Edition* 57.6, pp. 1474–1479. DOI: 10.1002/anie.201710798. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201710798. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201710798.

Hirose, Makoto et al. (2019). "Oxygen-diffusion-driven oxidation behavior and tracking areas visualized by X-ray spectro-ptychography with unsupervised learning". In: *Communications Chemistry* 2.1. ISSN: 2399-3669. DOI: 10.1038/s42004-019-0147-y. URL: https://doi.org/10.1038/s42004-019-0147-y.

Hirosuke, Matsui et al. "Imaging of oxygen diffusion in individual Platinum/Ce2Zr2Ox catalyst particles during oxygen storage and release". In: *Angewandte Chemie International Edition* 55.39 (), pp. 12022–12025. DOI: 10.1002/anie.201606046. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201606046.

Hook J.R.; Hall, H.E., ed. (2010). *Solid State Physics*. Manchester Physics Series (2nd ed.), ohn Wiley and Sons.

Hwang, Seung Jun et al. (2013). "Supported Core@Shell Electrocatalysts for Fuel Cells: Close Encounter with Reality". In: *Scientific Reports* 3, 1309 EP –. URL: https://doi.org/10.1038/srep01309.

Ihli, J. et al. (2017). "A three-dimensional view of structural changes caused by deactivation of fluid catalytic cracking catalysts". In: *Nature Communications* 8.1, p. 809. URL: https://doi.org/10.1038/s41467-017-00789-w.

Ishiguro, Nozomu et al. (2016). "Kinetics and Mechanism of Redox Processes of Pt/C and Pt3Co/C Cathode Electrocatalysts in a Polymer Electrolyte Fuel Cell during an Accelerated Durability Test". In: *The Journal of Physical Chemistry C* 120.35, pp. 19642–19651. URL: https://doi.org/10.1021/acs.jpcc.6b04437.

J., Behler (2015). "Constructing High-Dimensional Neural Network Potentials: A Tutorial Review". In: *Int. J. Quantum Chem* 115, 1032–1050. DOI: 10.1002/qua.24890. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24890.

J., Einbeck, Evers L., and Bailer-Jones C. (2008). "Representing complex data using localized principal components with application to astronomical data". In: *Principal Manifolds for Data Visualization and Dimension Reduction. Lecture Notes in Computational Science and Enginee* 58.

J., Kasper, Fornasiero P., and M Graziani (1999). "Use of CeO2-based oxides in the three-way catalysis". In: *Catal. Today* 50, pp. 285–298.

J., Miao et al. (2015). "eyond crystallography: diffractive imaging using coherent x-ray light sources". In: *Science* 348, pp. 530–535.

Jacobs, Robert A. et al. (1991). "Adaptive Mixtures of Local Experts". In: *Neural Comput.* 3.1, pp. 79–87. ISSN: 0899-7667. DOI: 10.1162/neco.1991.3.1.79. URL: http://dx.doi.org/10.1162/neco.1991.3.1.79.

Jain, Anubhav et al. (2013). "The materials project: a materials genome approach to accelerating materials innovation". In: *APL Materials* 1.1, p. 011002. ISSN: 2166532X.

Jia, Qingying et al. (2015). "Improved Oxygen Reduction Activity and Durability of Dealloyed PtCox Catalysts for Proton Exchange Membrane Fuel Cells: Strain, Ligand, and Particle Size Effects". In: 5, pp. 176–186. ISSN: 1. URL: https://doi.org/10.1021/cs501537n.

Jiang, Kezhu et al. (2016). "Ordered PdCu-Based Nanoparticles as Bifunctional Oxygen-Reduction and Ethanol-Oxidation Electrocatalysts". In: *Angewandte Chemie International Edition* 55.31, pp. 9030–9035. DOI: 10.1002/anie.201603022. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201603022. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201603022.

Jiang, Kezhu et al. (2017). "Efficient oxygen reduction catalysis by subnanometer Pt alloy nanowires". In: *Science Advances* 3.2. DOI: 10.1126/sciadv.1601705. eprint: https://advances.sciencemag.org/content/3/2/e1601705.full.pdf. URL: https://advances.sciencemag.org/content/3/2/e1601705.

Karim, Waiz et al. (2017). "Catalyst support effects on hydrogen spillover". In: *Nature* 541.68, 68 EP –. URL: https://doi.org/10.1038/nature20782.

Katja, Hansen et al. (2011). "Visual Interpretation of Kernel-Based Prediction Models". In: *Molecular Informatics* 30.9, pp. 817–826. DOI: 10.1002/minf.201100059. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201100059. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201100059.

Kevin, Williams et al. (2015). "Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases". In: *Journal of The Royal Society Interface* 12.104. URL: https://doi.org/10.1098/rsif.2014.1289.

King, R. D. et al. (2009). "The Robot Scientist Adam". In: *Computer* 42.8, pp. 46–54. ISSN: 0018-9162. DOI: 10.1109/MC.2009.270.

King, Ross D. et al. (2004). "Functional genomic hypothesis generation and experimentation by a robot scientist". In: *Nature* 427.6971, pp. 247–252. URL: https://doi.org/10.1038/nature02236.

Koh, S. et al. (2007). "Activity-Stability Relationships of Ordered And Disordered Alloy Phases of Pt3Co Electrocatalysts for the Oxygen Reduction Reaction (ORR)". In: *Electrochimica Acta* 52. DOI: 10.1016/j.electacta.2006.08.039. URL: https://www.osti.gov/servlets/purl/901833.

Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*. Vol. 2. Morgan Kaufmann Publishers, pp. 1137–1143.

Kohavi, Ron and George H. John (1997). "Wrappers for feature subset selection". In: *Artificial Intelligence* 97.1. Relevance, pp. 273 –324. ISSN: 0004-3702. URL: http://www.sciencedirect.com/science/article/pii/S000437029700043X.

Kruskal, J. B. (1964). "Nonmetric multidimensional scaling: A numerical method". In: *Psychometrika* 29.2, pp. 115–129. ISSN: 1860-0980. DOI: 10.1007/BF02289694. URL: https://doi.org/10.1007/BF02289694.

Kvalseth, Tarald O. (1985). "Cautionary note about R^2". In: *The American Statistician* 39.4.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015, 05). "Deep learning (article)". In: *Nature* 521, 436 EP. DOI: https://doi.org/10.1038/nature14539. URL: https://doi.org/10.1038/nature14539.

Lei, Yu et al. (Apr. 2011). "Effect of Particle Size and Adsorbates on the L3, L2 and L1 X-ray Absorption Near Edge Structure of Supported Pt Nanoparticles". In: *Topics in Catalysis* 54, pp. 334–348. DOI: 10.1007/s11244-011-9662-5.

Letham, Benjamin et al. (Sept. 2015). "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model". In: *Ann. Appl. Stat.* 9.3, pp. 1350–1371. DOI: 10.1214/15-AOAS848. URL: https://doi.org/10.1214/15-AOAS848.

Li, Mufan et al. (2016). "Ultrafine jagged platinum nanowires enable ultrahigh mass activity for the oxygen reduction reaction". In: *Science* 354.6318, pp. 1414–1419. ISSN: 0036-8075. DOI: 10.1126/science.aaf9050. eprint: https://science.sciencemag.org/content/354/6318/1414.full.pdf. URL: https://science.sciencemag.org/content/354/6318/1414.

Lilienfeld O. Anatole, von et al. (2015). "Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties". In: *International Journal of Quantum Chemistry* 115.16, pp. 1084–1093. DOI: 10.1002/qua.24912. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24912. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.24912.

Liu, Huan and Lei Yu (2005). "Toward integrating feature selection algorithms for classification and clustering". In: *IEEE Transactions on Knowledge and Data Engineering* 17. ISSN: 1041-4347.

Lo, Yu-Chen et al. (2018). "Machine learning in chemoinformatics and drug discovery". In: *Drug Discovery Today* 23.8, pp. 1538 –1546. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2018.05.010. URL: http://www.sciencedirect.com/science/article/pii/S1359644617304695.

Luo, Huimin et al. (May 2016). "Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm". In: *Bioinformatics* 32.17, pp. 2664–2671. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw228. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/32/17/2664/17345874/btw228.pdf. URL: https://doi.org/10.1093/bioinformatics/btw228.

Lázaro-Gredilla, Miguel, Steven Van Vaerenbergh, and Neil D. Lawrence (2012). "Overlapping Mixtures of Gaussian Processes for the data association problem". In: *Pattern Recognition* 45.4, pp. 1386 –1395. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2011.10.004. URL: http://www.sciencedirect.com/science/article/pii/S0031320311004109.

López-Haro, M. et al. "Advanced electron microscopy investigation of Ceria–Zirconia-based catalysts". In: *ChemCatChem* 3.6 (), pp. 1015–1027. DOI: 10.1002/cctc.201000306. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cctc.201000306.

M., Hirose et al. (2017). "Use of Kramers-Kronig relation in phase retrieval calculation in X-ray spectro- ptychography". In: *Opt. Express* 25, pp. 8593–8603.

M, Hirose et al. (2018). "Visualization of heterogeneous oxygen storage behavior in platinum-supported cerium-zirconium oxide three-way catalyst particles by hard X-ray spectro-ptychography". In: *Angew. Chem. Int. Ed* 130, pp. 1490–1495.

M., Maiden A. and Rodenburg J. M (2009). "An improved ptychographical phase retrieval algorithm for diffractive imaging". In: *Ultramicroscopy* 109, pp. 1256–1262.

Marr, Bernard (2018). "27 Incredible Examples Of AI And Machine Learning In Practice". In: *Forbes*. URL: https://www.forbes.com/sites/bernardmarr/2018/04/30/27-incredible-examples-of-ai-and-machine-learning-in-practice/#66b02f5d7502.

Matsui, Hirosuke et al. "Operando 3D Visualization of Migration and Degradation of a Platinum Cathode Catalyst in a Polymer Electrolyte Fuel Cell". In: *Angewandte Chemie International Edition* 56.32 (), pp. 9371–9375. DOI: 10.1002/anie.201703940. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201703940. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201703940.

Matthias, Rupp (2015). "Machine learning for quantum mechanics in a nutshell". In: *International Journal of Quantum Chemistry* 115.16, pp. 1058–1073.

Mead, A. (1992). "Review of the Development of Multidimensional Scaling Methods". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 41.1, pp. 27–39. URL: http://www.jstor.org/stable/2348634.

Meeds, Edward and Simon Osindero (2006). "An Alternative Infinite Mixture Of Gaussian Process Experts". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, B. Schölkopf, and J. C. Platt. MIT Press, pp. 883–890. URL: http://papers.nips.cc/paper/2768-an-alternative-infinite-mixture-of-gaussian-process-experts.pdf.

Meirer, Florian et al. (2015). "Mapping metals incorporation of a whole single catalyst particle using element specific X-ray nanotomography". In: *J. Am. Chem. Soc* 137, pp. 102–105. URL: https://doi.org/10.1021/ja511503d.

Meyer, Julius Lothar, ed. (1864). *Die modernen Theorien der Chemie*. URL: https://reader.digitale-sammlungen.de/de/fs1/object/goToPage/bsb10073411.html?pageNo=147.

Moseley, Henry Gwyn Jeffreys (1914). "The high-frequency spectra of the elements". In: *Philosophical Magazine, 6th series* 27, 703–713.

Murphy, Kevin P., ed. (2012a). *Machine learning: a probabilistic perpective*. MIT Press.

— ed. (2012b). *Machine learning: a probabilistic perpective, Chapter 11.2.1*. MIT Press, p. 339.

Murtagh, Fionn and Pedro Contreras (2011). "Methods of Hierarchical Clustering". In: *CoRR* abs/1105.0121. arXiv: 1105.0121. URL: http://arxiv.org/abs/1105.0121.

Nadaraya, E. (1964). "On Estimating Regression". In: *Theory of Probability & Its Applications* 9.1, pp. 141–142. DOI: 10.1137/1109020. eprint: https://doi.org/10.1137/1109020. URL: https://doi.org/10.1137/1109020.

Newlands, John A. R (8 August 1865). "On the law of octaves". In: *The Chemical News* 1, 12: 83. URL: https://babel.hathitrust.org/cgi/pt?id=nyp.33433062749274;view=1up;seq=97.

Nguyen, Duong-Nguyen et al. (2018). "Committee machine that votes for similarity between materials". In: *IUCrJ* 5.6, pp. 830–840. DOI: 10.1107/S2052252518013519. URL: https://doi.org/10.1107/S2052252518013519.

Nguyen, Duong-Nguyen et al. (2019). "Ensemble learning reveals dissimilarity between rare-earth transition binary alloys with respect to the Curie temperature". In: *Journal of Physics: Materials*. URL: http://iopscience.iop.org/10.1088/2515-7639/ab1738.

Nikkuni, Flávio R. et al. (2015). "Accelerated degradation of Pt3Co/C and Pt/C electrocatalysts studied by identical-location transmission electron microscopy in polymer electrolyte environment". In: *Applied Catalysis B: Environmental* 176-177, pp. 486 –499. ISSN: 0926-3373. DOI: https://doi.org/10.1016/j.apcatb.2015.04.035. URL: http://www.sciencedirect.com/science/article/pii/S0926337315002234.

Nørskov, J. K. et al. (2004). "Origin of the Overpotential for Oxygen Reduction at a Fuel-Cell Cathode". In: *The Journal of Physical Chemistry B* 108.46, pp. 17886–17892. URL: https://doi.org/10.1021/jp047349j.

Ouyang, Runhai et al. (2018). "SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates". In: *Phys. Rev. Materials* 2 (8), p. 083802. DOI: 10.1103/PhysRevMaterials.2.083802. URL: https://link.aps.org/doi/10.1103/PhysRevMaterials.2.083802.

Ozawa, Saki et al. (2018a). "Operando Time-Resolved X-ray Absorption Fine Structure Study for Pt Oxidation Kinetics on Pt/C and Pt3Co/C Cathode Catalysts by Polymer Electrolyte Fuel Cell Voltage Operation Synchronized with Rapid O2 Exposure". In: *The Journal of Physical Chemistry C*, pp. 14511–14517. URL: https://doi.org/10.1021/acs.jpcc.8b02541.

— (2018b). "Operando Time-Resolved X-ray Absorption Fine Structure Study for Pt Oxidation Kinetics on Pt/C and Pt3Co/C Cathode Catalysts by Polymer Electrolyte Fuel Cell Voltage Operation Synchronized with Rapid O2 Exposure". In: *The Journal of Physical Chemistry C* 122. ISSN: 26. URL: https://doi.org/10.1021/acs.jpcc.8b02541.

P., Dholabhai P., Perriot R., and Uberuaga B. P (2016). "Atomic-scale structure and stability of the low-index surfaces of pyrochlore oxides". In: *J. Phys. Chem. C* 120, pp. 10485–10499.

Picard, Richard R. and R. Dennis Cook (1984). "Cross validation of regression models". In: *Journal of the American Statistical Association* 79.387. URL: http://www.jstor.org/stable/2288403.

Pilania, Ghanshyam et al. (2013). "Accelerating materials property predictions using machine learning". In: *Scientific Reports* 3.

P.Myers, H., ed. (1997). *Introductory Solid State Physics (Second Edition)*. CRC Press.

R, Maitra (2009). "Initializing partition-optimization algorithms." In: *IEEE/ACM Trans Comput Biol Bioinform.* 6.1. ISSN: 144-57.

R. Hofstadter, Douglas (Jan. 2006). "Analogy as the Core of Cognition". In: *Current Opinion in Neurobiology* 12, pp. 1–21.

Rasmussen, Carl E. and Zoubin Ghahramani (2002). "Infinite Mixtures of Gaussian Process Experts". In: *Advances in Neural Information Processing Systems 14*. Ed. by T. G. Dietterich, S. Becker, and Z. Ghahramani, pp. 881–888. URL: http://papers.nips.cc/paper/2055-infinite-mixtures-of-gaussian-process-experts.pdf.

Rasouli, S. et al. (2017). "Surface area loss mechanisms of Pt3Co nanocatalysts in proton exchange membrane fuel cells". In: *Journal of Power Sources* 343, pp. 571 – 579. ISSN: 0378-7753. DOI: https://doi.org/10.1016/j.jpowsour.2017.01.058. URL: http://www.sciencedirect.com/science/article/pii/S0378775317300587.

Ravenhorst, Ilse K. van et al. "Capturing the Genesis of an Active Fischer–Tropsch Synthesis Catalyst with Operando X-ray Nanospectroscopy". In: *Angewandte Chemie International Edition* 57.37 (), pp. 11957–11962. DOI: 10.1002/anie.201806354. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201806354.

Rodenburg, John Marius et al. (2007). "Hard-x-ray lensless imaging of extended objects." In: *Physical review letters* 98 3, p. 034801.

Rokach, Lior and O Maimon, eds. (2008). World Scientific.

Ross, James C. and Jennifer G. Dy (2013). "Nonparametric Mixture of Gaussian Processes with Constraints". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13, pp. III–1346–III–1354. URL: http://dl.acm.org/citation.cfm?id=3042817.3043087.

Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5, pp. 206–215. URL: https://doi.org/10.1038/s42256-019-0048-x.

Rupp, Matthias and Gisbert Schneider (2010). "Graph Kernels for Molecular Similarity". In: *Molecular Informatics* 29.4, pp. 266–273. DOI: 10.1002/minf.200900080. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.200900080. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.200900080.

Rupp, Matthias et al. (2012). "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning". In: *Phys. Rev. Lett.* 108 (5), p. 058301. DOI: 10.1103/PhysRevLett.108.058301. URL: https://link.aps.org/doi/10.1103/PhysRevLett.108.058301.

S, Matsumoto (2004). "Recent advances in automobile exhaust catalysts". In: *Catal. Today* 90, pp. 183–190.

S., Mitchell, Michels N. L., and Perez-Ramirez J. (2013). "From powder to technical body: the undervalued science of catalyst scale up". In: *Chem. Soc. Rev.* 42, pp. 6094–6112.

SASAKI, Tsuyoshi et al. (2003). "Oxygen absorption behavior of Ce2Zr2O7+x and formation of Ce2Zr2O7.5". In: *J. Ceram. Soc. Jpn* 111, pp. 382–385.

Sasaki, Tsuyoshi et al. (2004). "Crystal Structure of Ce2Zr2O7 and .BETA.-Ce2Zr2O7.5". In: *Journal of The Ceramic Society of Japan - J CERAMIC SOC JPN* 112, pp. 440–444. DOI: 10.2109/jcersj.112.440.

Schweitzer, Neil et al. (2010). "Establishing Relationships Between the Geometric Structure and Chemical Reactivity of Alloy Catalysts Based on Their Measured Electronic Structure". In: *Topics in Catalysis* 53.5, pp. 348–356. ISSN: 1572-9028. DOI: 10.1007/s11244-010-9448-1. URL: https://doi.org/10.1007/s11244-010-9448-1.

Seh, Zhi Wei et al. (2017). "Combining theory and experiment in electrocatalysis: Insights into materials design". In: 355.6321. DOI: 10.1126/science.aad4998.

Seniha, Yuksel, Wilson Joseph, and Gader Paul. (2012). "Twenty Years of Mixture of Experts." In: *Neural Networks and Learning Systems, IEEE Transactions* 23, pp. 1177–1193. ISSN: 10.1109/TNNLS.2012.2200299.

Settles, Burr (2010). "Active learning literature survey". In: *Computer Sciences Technical Report 1648 University of Wisconsin Madison*.

Seung, H. Sebastian et al. (1992). "Query by committee". In: *Proceedings of the fifth annual workshop on Computational learning theory. ACM*.

Shapiro, David A. et al. (2014). "Chemical composition mapping with nanometre resolution by soft X-ray microscopy". In: *Nat. Photonics* 8, pp. 765–769. URL: https://doi.org/10.1038/nphoton.2014.207.

Silverman, B. W. (1985). "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 47.1, pp. 1–52. URL: https://www.ece.uvic.ca/~bctill/papers/mocap/Silverman_1985.pdf.

Smit, Emiel de et al. (2008). "Nanoscale chemical imaging of a working catalyst by scanning transmission X-ray microscopy". In: *Nature* 456, pp. 222–225.

Snyder, John C. et al. (2012). "Finding Density Functionals with Machine Learning". In: *Phys. Rev. Lett.* 108 (25), p. 253002. DOI: 10.1103/PhysRevLett.108.253002. URL: https://link.aps.org/doi/10.1103/PhysRevLett.108.253002.

Souza, Camila P. E. de and Nancy E. Heckman (2014). "Switching nonparametric regression models". In: *Journal of Nonparametric Statistics* 26.4, pp. 617–637. DOI: 10.1080/10485252.2014.941364. eprint: https://doi.org/10.1080/10485252.2014.941364. URL: https://doi.org/10.1080/10485252.2014.941364.

Stamenkovic, Vojislav et al. (2006). "Changing the Activity of Electrocatalysts for Oxygen Reduction by Tuning the Surface Electronic Structure". In: *Angewandte Chemie International Edition* 45.18, pp. 2897–2901. DOI: 10.1002/anie.200504386. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200504386. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200504386.

Stamenkovic, Vojislav R. et al. (2007). "Trends in electrocatalysis on extended and nanoscale Pt-bimetallic alloy surfaces". In: *Nature Materials* 6, p. 241. URL: https://doi.org/10.1038/nmat1840.

Stephens, Ifan, Jan Rossmeisl, and Ib Chorkendorff (2016). "Toward sustainable fuel cells". English. In: *Science* 354.6318, pp. 1378–1379. ISSN: 0036-8075. DOI: 10.1126/science.aal3303.

Stone, M. (1974). "Cross validatory choice and assessment of statistical predictions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2.

Sung-Hyuk, Cha (2007). "Comprehensive survey on distance/similarity measures between probability density functions". In: *Int J Math Model methods Appl Sci* 1, 300–307. DOI: 10.1.1.154.8446.

Suzuki, Hiroyuki (2017). "Metastable phase YFe12 fabricated by rapid quenching method". In: *AIP Advances* 7.5, p. 056208. DOI: 10.1063/1.4973799. eprint: https://doi.org/10.1063/1.4973799.

Tada, Mizuki et al. (2011). "μ-XAFS of a single particle of a practical NiOx/Ce2Zr2Oy catalyst". In: *Phys. Chem. Chem. Phys* 13, pp. 14910–14913.

Tada, Mizuki et al. (2012). "The Active Phase of Nickel/Ordered Ce2Zr2Ox Catalysts with a Discontinuity (x=7–8) in Methane Steam Reforming". In: *Angewandte Chemie International Edition* 51.37, pp. 9361–9365. DOI: 10.1002/anie.201205167. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201205167. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201205167.

Takagi, Yasumasa et al. (2017). "X-ray photoelectron spectroscopy under real ambient pressure conditions". In: *Applied Physics Express* 10.7. ISSN: 076603. URL: https://iopscience.iop.org/article/10.7567/APEX.10.076603/meta#back-to-top-target.

Takahashi, Keisuke et al. (2017). "Descriptors for predicting the lattice constant of body centered cubic crystal". In: *The Journal of Chemical Physics* 146.204104, p. 011002. DOI: 10.1063/1.4984047.

Tanimoto, T. (1957). "An Elementary Mathematical theory of Classification and Prediction". In: *Internal IBM Technical Report*.

Tenenbaum, J.B. and T.L. Griffiths (Sept. 2001). "Generalization, similarity, and Bayesian inference". In: *The Behavioral and brain sciences* 24, 629–40; discussion 652.

Tenenbaum, Joshua B. (1996). "Learning the Structure of Similarity". In: *Advances in Neural Information Processing Systems 8*. Ed. by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo. MIT Press, pp. 3–9. URL: http://papers.nips.cc/paper/1052-learning-the-structure-of-similarity.pdf.

— (2000). "Rules and Similarity in Concept Learning". In: *Advances in Neural Information Processing Systems 12*. Ed. by S. A. Solla, T. K. Leen, and K. Müller. MIT Press, pp. 59–65. URL: http://papers.nips.cc/paper/1666-rules-and-similarity-in-concept-learning.pdf.

Tibshirani, Robert. (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society, Series B (Methodological)* 58. ISSN: 1.

Tien-Lam, Pham et al. (2018). "Learning structure-property relationship in crystalline materials: A study of lanthanide–transition metal alloys". In: *The Journal of Chemical Physics* 148.20, p. 204106. DOI: 10.1063/1.5021089. eprint: https://doi.org/10.1063/1.5021089. URL: https://doi.org/10.1063/1.5021089.

Tresp, Volker (2001). "Committee machines". In: *Neural Computation* 12, p. 2000.

Tversky, Amos (1977). "Features of Similarity". In: *Psychological Review* 84.4, pp. 327–352. DOI: 10.1037/0033-295X.84.4.327.

Urban, Sven et al. (2017). "In situ study of the oxygen-induced transformation of pyrochlore Ce2Zr2O7+x to the k-Ce2Zr2O8 phase". In: *Chem. Mater.* 29, pp. 9218–9226. URL: https://doi.org/10.1021/acs.chemmater.7b03091.

Vidal, Rene, Yi Ma, and Shankar Sastry (Dec. 2015). "Generalized principal component analysis (GPCA)". In: *IEEE transactions on pattern analysis and machine intelligence* 27.12.

Villars, P. et al. (2004). "The Pauling File, Binaries Edition". In: *Journal of Alloys and Compounds* 367.

Visalakshi, S. and V. Radha (2014). "A literature review of feature selection techniques and applications: Review of feature selection in data mining". In: *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–6.

Vliet, Dennis F. van der et al. (2012). "Unique Electrochemical Adsorption Properties of Pt-Skin Surfaces". In: *Angewandte Chemie International Edition* 51.13, pp. 3139–3142. DOI: 10.1002/anie.201107668. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201107668. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201107668.

Wang, Bo and D. M. Titterington (2006). "Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model". In: *Bayesian Analysis* 1.3.

Wang, Deli et al. (2012). "Structurally ordered intermetallic platinum–cobalt core–shell nanoparticles with enhanced activity and stability as oxygen reduction electrocatalysts". In: *Nature Materials*. URL: https://doi.org/10.1038/nmat3458.

Wanjala, Bridgid N. et al. (2011). "Correlation between Atomic Coordination Structure and Enhanced Electrocatalytic Activity for Trimetallic Alloy Catalysts". In: *Journal of the American Chemical Society* 133. ISSN: 32. URL: https://doi.org/10.1021/ja2040464.

Watson, Geoffrey S (1964). "Smooth Regression Analysis." In: *JSTOR* 26.4, 359–372. URL: www.jstor.org/stable/25049340.

Wise, Anna M. et al. (2016a). "Nanoscale chemical imaging of an individual catalyst particle with soft X-ray ptychography". In: *ACS Catal* 6, pp. 2178–2181. URL: https://doi.org/10.1021/acscatal.6b00221.

— (2016b). "Nanoscale Chemical Imaging of an Individual Catalyst Particle with Soft X-ray Ptychography". In: *ACS Catalysis* 6.4, pp. 2178–2181. URL: https://doi.org/10.1021/acscatal.6b00221.

Wu, Juan et al. (2018a). "4D imaging of polymer electrolyte membrane fuel cell catalyst layers by soft X-ray spectro-tomography". In: *Journal of Power Sources* 381, pp. 72 –83. ISSN: 0378-7753. DOI: https://doi.org/10.1016/j.jpowsour.2018.01.074. URL: http://www.sciencedirect.com/science/article/pii/S0378775318300740.

Wu, Juan et al. (2018b). "Four-dimensional imaging of ZnO-coated alumina aerogels by scanning transmission X-ray microscopy and ptychographic tomography". In: *J. Phys. Chem. C* 122, pp. 25374–25385. URL: https://doi.org/10.1021/acs.jpcc.8b07363.

Wu, Juan et al. (2018c). "High-resolution imaging of polymer electrolyte membrane fuel cell cathode layers by soft X-ray spectro-ptychography". In: *J. Phys. Chem. C* 122, pp. 11709–11719. URL: https://doi.org/10.1021/acs.jpcc.8b02933.

— (2018d). "High-Resolution Imaging of Polymer Electrolyte Membrane Fuel Cell Cathode Layers by Soft X-ray Spectro-Ptychography". In: *The Journal of Physical Chemistry C* 122.22, pp. 11709–11719. URL: https://doi.org/10.1021/acs.jpcc.8b02933.

Xia, Wei et al. (2016). "Earth-Abundant Nanomaterials for Oxygen Reduction". In: *Angewandte Chemie International Edition* 55.8, pp. 2650–2676. DOI: 10.1002/anie.201504830. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201504830. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201504830.

Xu, Dongkuan and Yingjie Tian (2015). "A Comprehensive Survey of Clustering Algorithms". In: *Annals of Data Science* 2.2, pp. 165–193. ISSN: 2198-5812. DOI: 10.1007/s40745-015-0040-1. URL: https://doi.org/10.1007/s40745-015-0040-1.

Xu, Yibin, Masayoshi Yamazaki, and Pierre Villarsß (2011). "Inorganic Materials Database for Exploring the Nature of Material". In: *Jpn. J. Appl. Phys.* 50.11RH02.

Yamamoto, Takashi et al. "Origin and Dynamics of Oxygen Storage/Release in a Pt/Ordered CeO2–ZrO2 Catalyst Studied by Time-Resolved XAFS Analysis". In: *Angewandte Chemie International Edition* 46.48 (), pp. 9253–9256. DOI: 10.1002/anie.200703085. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200703085.

Yu, Lei and Huan Liu (2004). "Efficient Feature Selection via Analysis of Relevance and Redundancy". In: *J. Mach. Learn. Res.* 5, pp. 1205–1224.

Yu, Young-Sang et al. (2015). "Dependence on crystal size of the nanoscale chemical phase distribution and fracture in LixFePO4". In: *Nano Lett.* 15, pp. 4282–4288. URL: https://doi.org/10.1021/acs.nanolett.5b01314.

Yu, Young-Sang et al. (2018). "Three-dimensional localization of nanoscale battery reactions using soft X-ray tomography". In: *Nat. Commun* 9.921, https://doi.org/10.1038/s41467–018–03401–x.

Zhang, Cha and Yunqian Ma (2012). *Ensemble machine learning: methods and applications*. Springer Publishing Company, Incorporated. ISBN: 1441993258, 9781441993250.

Zhao, Xiaojing et al. (May 2013). "Evaluation of change in nanostructure through the heat treatment of carbon materials and their durability for the start/stop operation of polymer electrolyte fuel cells". English. In: *Electrochimica Acta* 97, pp. 33–41. ISSN: 0013-4686. DOI: 10.1016/j.electacta.2013.02.062.

Zhu, Xiaohui et al. (2016). "Measuring spectroscopy and magnetism of extracted and intracellular magnetosomes using soft X-ray ptychography". In: *Proc. Natl Acad. Sci. USA* 113, E8219–E8227. URL: https://doi.org/10.1073/pnas.1610260114.