Diversity Preservation in Similarity-based Inference

Dang Tran Thai

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Diversity Preservation in Similarity-based Inference

Dang Tran Thai

Supervisor: Associate Professor Hieu Chi Dam

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Knowledge Science
September 2019

Abstract

Similarity-based inference has been widely used for recognition. The principle behind the similarity-based inference is that similar objects will share common properties. In machine learning, similarity-based inference is employed through various methods such clustering, k-nearest neighbors, etc. In addition, similarity-based inference is useful for controlling confounding factors in statistical causality inference.

There are several issues in using similarity-based inference in practice. The principles of the inference are applicable if the representation of objects and similarity measure used for this representation are ideal. In case these factors are not ideal, there has the inconsistency of the similarity measured based on the objects' representation with the similarity of objects' target values. In addition, in analogy-based causality inference, similar causes play the role of reference factors for assessing the relation between the cause of interest with effects. Hence, the main issue here is how to choose good similar causes for accurately recognizing confounding factors.

This work aims to solve the issues mentioned above through verifying the proposed hypothesis that conservation of diversity in selecting models and data samples can help to effective solve these issues. As such, we enrich the knowledge about the diversity preservation in machine learning.

We demonstrate issues in similarity-based inference through specific studies. The first one regards to measure the similarity between materials for effectively predicting materials' formation energies. The second one regards to control polypharmacy-induced confounding in assessing the cause of drug adverse reaction. Through these studies, we can evaluate the likelihood of our proposed hypothesis. In both studies, we focus on model interpretation and explanation based on model performance.

In the first study, we address the problem that most materials' descriptors in vector space are not ideal for representing materials for predicting formation energy, which induces the roughness of the energy surface. Hence, the similarity of materials measured based on their presentation is not consistent with the similarity of their energies. In this situation, finding an appropriate similarity measure for these descriptors may help to improve the performance of similarity-based learning models in approximating the energy surface. We hypothesize that to effectively approximate the energy function, similarity measures need to preserve the distinction of two objects in comparison with the third one. We propose a protocol for verifying this hypothesis that incorporates various methods for investigating the roughness of energy surface and similarity measures. In addition, we also proposed a method for estimating the loss of distinction of two objects in comparison with the third one when using similarity measures. The experimental results show the high likelihood of our proposed hypothesis. Furthermore, we establish general principles for effectively using similarity measures for mining materials data, which do not depend on any specific learning method.

In the second study, we concentrate on an important problem in post-marketing pharmaceutical surveillance that is drug-adverse reaction causality assessment. The main issue here is to deal with confounding factors induced by polypharmacy in the treatment. In this study, we employ reference sets constructed based on the analogy criterion – one of nine Bradford Hill criteria to control confounding factors. This criterion states that similar drugs may cause similar adverse events. We propose a novel model, called the

analogy-based active voting, for effectively assessing causal relations between drugs and adverse events. This model mimics the analogy criterion by a voting process of similar drugs. In this context, each drug is represented by a set of its associated adverse events extracted from electronic medical records. The diversity of these sets induce the conflict in voting of similar drugs, which plays an importance role for eliminating non-causal drug-adverse reaction pairs. This case study demonstrates the importance of diversifying reference in analogy-based causality inference.

Keywords: Similarity-based inference, diversity preservation, similarity measure, confounding, analogy-based causality inference

Acknowledgements

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. First all all, I would like to express my gratitude to my supervisor, Professor Dam Hieu Chi, Professor Ho Tu Bao for their supervision, advice, assistance during my master course and doctoral course. As my supervisor, they taught me not only knowledge about machine learning and data mining but also developing new idea, solving problems, and critical thinking. They also provided me kind encouragements and supports not only for my study but also for my life in Japan.

I wish to thank to my committee members: Professor Huynh Van Nam, Professor Hideomi Gokon, Professor Kenji Satou, and Professor Ho Tu Bao, for reading my thesis and providing me valuable feedback.

I would like to thank Professor Phung Quoc Dinh of Monash University for his guidance and support during my off-campus research in Australia.

I would like to acknowledge the 5D Scholarship Program of Japan Advanced Institute of Science and Technology (JAIST) for the financial supports during my study.

I would like to thank all members in Ho & Dam Laboratory of JAIST for their support during my research.

Finally, I would like to give a special thank to my family for their encouragements and supports during my study at JAIST.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context-dependent diversity

1.1.1 Example

Before define the concept of context-based diversity, we shows a simple example, which provides a first glance at this concept. Given a collection of three countries C ={United Kingdom, France, Germany} which are described by the continent they locate, the main language used in these countries, and the faction they belong to in the World War II (WWII), as shown in Table 1.1.

Relying on this table, if countries are compared with each other based on their continent, they are identical because all of them belong to the Europe. Meanwhile, when based on the faction in WWII, United Kingdom and France are different from Germany, so the collection C is more diverse. When based on the language, countries in each possible pair are different, thus, the collection in this context is most diverse. Through this example, we see that the diversity of a collection can vary when this collection is considered in different contexts.

1.1.2 Context-dependent similarity and difference evaluation

The evaluation of similarity and evaluation of difference are complementary (the difference can be considered a linear function of similarity with slope of -1), and depend on a specific context [93]. For example, the United Kingdom and France are similar in terms of continent, while are different in terms of the main language.

The evaluation of similarity or difference is commonly carried out with two objects by measuring how alike these objects are. In terms of mathematics, similarity measures are

Table 1.1: Example of context-based diversity

Country	Continent	Language	Faction in WWII
United Kingdom	Europe	English	Alliance
France	Europe	French	Alliance
Germany	Europe	German	Axis powers

real-valued functions that take representations of two objects as the function input, and then output a scalar. For example, distances between two points in the vector space such as the Euclidean, Manhattan, and cosine are used for measuring the similarity of these points. In fact, these distances are functions of two variables.

1.1.3 Context-dependent diversity evaluation

Differing from the similarity (or difference) evaluation, the evaluation of diversity is often conducted for a collection of n objects with $n \geq 3$. This evaluation is based on the set of pairwise difference of objects in this collection. In other words, the diversity evaluation is an aggregation of objects' pairwise differences. For example, in context of the faction in WWII, the number of country pairs whose elements are different is 2, while the number of these pairs in context of language is 3. Hence, in context of language, the collection is more diverse than that in context of the faction in WWII.

As mentioned above, the evaluation of similarity and difference depends on a specific context. Therefore, the evaluation of diversity in a collection also depends on a given context.

1.2 Diversity preservation in machine learning

Essentially, machine learning models aim to represent a collection of real-world objects and the relation among them in specifict contexts based on the "no free lunch" theorem [100]. In fact, each model here corresponds to a context of these objects. Obviously, objects in the real-world are essentially diverse. In several situations, capturing the diversity of these objects can help machine learning models attain high performance. To reflect the diversity of real-world objects, diversification of data, model parameters, and model ensemble is necessary [36]. Therefore, the term of "diversity preservation" in terms of machine learning refers to: the selection of data samples that maximize information contained for training process; the selection of models whose parameters can reflect much information in the data; and the selection of model ensembles whose based models are diverse.

1.2.1 Importance of diversity preservation

Machine learning techniques have been widely applied to solve real-world problems that is expected to make incredible improvements for people lives, and to accelerate scientific discovery. There are numerous factors that can affects the performance of machine learning systems, in which the diversity of training data and learning process plays an important role. Indeed, the diversity property in data and learning models can make the fairness in assessing learning models, enlarge the searching space of hypotheses in these models, and enhance the effective information exploitation from data. In addition, the diversity of recommended items can help to minimize the risk induced by the user dissatisfaction in information retrieval and recommender systems.

In machine learning, the "no free lunch" theorem [100] states that there is no search and optimization algorithm is expected to perform better than any other algorithms. In other words, there has no a model which is the best for solving all problems. Hence,

the selection of appropriate models is an inevitable step in most of learning methods. To perform the suitable selection, evaluating and comparing models play an important role. One of important criteria is to assess how a model is generalized for achieving better prediction performance. The evaluation on the generalization of models requires the use of different testing or validation data samples. Thus, diversifying the testing sets helps with the fairness in model assessment. For example, cross-validation, which is a well-known method for model assessment, generates one several pairs of training and validation sets for estimating the risk of the model in prediction [5].

In principle, a model is tuned for fitting with an available dataset, hence, to avoid the bias of such a model and overfitting as well as enhance the generalization of such a model for effectively predicting new instances, the learning process should be carried out with difference data samples. A well-known approach, which attempts to generate multiple learners and then incorporates these learners for improve prediction, is the ensemble learning. Several methods in the ensemble learning, e.g., bagging and adaBoost, have been widely used in many applications. These methods target to enlarge the searching space of hypotheses from the amount of available training dataset, and then aggregate these hypotheses for making the improvement in prediction [27, 106]. In ensemble learning methods, diversifying classifiers is needful to be considered when building classifiers ensembles for real-life pattern recognition, which was proven in [57].

The diversity property helps machine learning techniques be able to adapt with real-world problems, furthermore, enhance their ability for solving these problems. In the area of recommendation system, the diversity of output has been taken into account in information retrieval and recommender systems for a long time. The reason is that in searching engines and recommender systems, diversifying results helps to minimize the risk of dissatisfaction of users [2]. In other words, diverse items that are recommended for users will help the users get more options for selection, and improve their satisfaction. The importance of result diversification has been discussed in early work on information retrieval. This problem is stated that the relevance of retrieval documents depends on not only the individual relevance of each user, but also on how they are related to other users [19]. Ideally, we expect that recommended documents should be relevant to the common interest of most of users in the population [22]. Because of the importance of result diversification, numerous studies so far have attempted to develop diversity-based ranking methods for improving the quality of retrieved items. Besides the result diversification in recommendation systems, in [45], the author proved that the diversity and size of data are important factors for better performance of machine learning methods. This proof was based on a comparison of performance in playing between the AlphaGo and AlphaGo Zero systems.

1.2.2 The need of measuring the diversity

As mentioned above, diversity preservation is known as the selection of data samples and learning models that maximizes the ability of reflecting the diversity of real-world objects. Hence, defining and measuring the diversity in specific situations are needful because they are used for establishing criteria for selecting data samples and models.

In general, quantifying the diversity plays an important role for effectively constructing learning models that minimize the risk when performing with real-world data. The risk

can be caused by the limited labeled data for training as the motivation of active learning, additionally can be caused when the events happening the future are out of intended outcome of the model as the issue in recommendation system.

In recommendation systems, although the recommending models can be fitted with the historical information of customers' interest, these models can be poor to suggest items for the customers in the future because the interest of customers may change over time, even the interest can be significantly different from that collected in the historical data. Hence, diversifying retrieval results is important for search engines and recommender systems as mentioned above. In these systems, the measure for ranking items needs to make the trade-off between the relevance level of items to a specific customer and the novelty level of these items to this customer.

In machine learning, data diversification is important that aims to provide informative samples for training machine learning models. Considering the diversity of instances in these samples aims to maximize the amount of information contained in these ones. Thus, measuring the diversity of samples, a.k.a. measuring the informativeness of samples, is needful to provide the basis for selecting informative training samples. In active learning, we aim to enrich existing labeled data for effectively training, while reduce the cost of labeling and time consuming. Thus, selecting informative samples for labeling plays an important role, which are based on evaluations on the diversity of samples [104, 91].

1.3 Problem and research objectives

In this study, we focus on the problem of diversity preservation in similarity-based inference. This problem is specified through two studies: (i) preserving the distinction of pairwise comparison in triplet of objects in measuring similarity for approximating roughness target function; (ii) diversifying the reference in analogy-based causality inference.

1.3.1 Similarity-based inference

Measuring the similarity is a fundamental process in analogy-based recognition. The principle behind the similarity-based inference is that similar objects will share common properties. In machine learning, similarity-based inference has been widely used through several well-known methods such as clustering, k-nearest neighbors methods, etc. By this methods, unknown target values of new instances are inferred by comparing these instances with existing ones based on similarity measures. In addition, the similarity-based inference can be used in causality inference that similar presume causes may result in similar effects.

There are several issues in similarity-based inference. In practice, the principle of similarity-based inference is applicable if the representation of objects and similarity measures used for this representation are ideal. However, finding an ideal representation is so difficult and takes lots of time. In case the representation and similarity measure are not ideal, there will be an inconsistency of measuring objects similarity based on their representation with the similarity of their target values. In fact, this induces the roughness of target function. In analogy-based causality inference, the use of this principle can help for controlling confounding, in which similar causes play the role of reference for assessing

the relation between the cause of interest with effects. The main issue here is that how to choose good similar causes for accurately recognizing confounding factors.

1.3.2 Objectives

As mentioned in previous sections, we address two issues in similarity-based inference: (i) the inconsistency of measuring objects similarity based on their representation with the similarity of their target values because of using non-ideal representation; (ii) how to design and select good similar causes for effectively controlling confounding in analogy-based causality inference. Our work aims to solve these issues. We hypothesize that conservation of the diversity in selecting models and data samples can help to effectively solve these issues.

Regarding (i), we hypothesize that the use of similarity measures that preserve the distinction of pairwise comparison in a triplet of objects can help similarity-based learning models improve the performance in approximating rough target functions induced by the use of non-ideal representation. Regarding (ii), we hypothesize that diversifying the reference (similar causes) can help to improve the performance of similarity-based causality inference. In this work, we aim to verify and estimate the likelihood of proposed hypotheses.

Through solving these issues, we aim to enrich the knowledge about diversity preservation in machine learning by providing additional views in such a situation. To verify our proposed hypotheses, we address two main objectives as follows:

- Defining and measuring the concept of diversity in specific context when solving each issue. This helps for assessing similarity measures and collections of similar causes used for controlling confounding factors.

- Interpreting why preserving the diversity when selecting similarity measures and samples of reference factors (similar causes) can help to solve these issues.

1.4 Contributions

We demonstrate issues in similarity-based inference as mentioned above by specific studies. For the first issue, we carry out a study on measuring the similarity between materials for effectively predicting materials' formation energies. For the second one, we carry a study on controlling polypharmacy-induced confounding in assessing the cause of drug adverse reaction. Through investigations in these studies, we can verify and estimate the likelihood of proposed hypotheses on the role of diversity preservation.

1.4.1 Measuring the similarity between materials for effectively predicting materials' formation energies

The main problem is this study is that most material descriptors (representations) of interest in the vector space are not ideal for representing materials for predicting formation energies, which induces the roughness of the energy surface. Hence, measuring the similarity of materials based on their presentation in vector space is not compatible with the

similarity of their formation energy. In other words, neighbors of materials in a vicinity, determined by similarity measures in the vector space, may have the energies that are extremely different from the energy of these materials.

We hypothesize that in this situation, finding an appropriate similarity measure for these descriptors may help to improve the performance of similarity-based learning models in approximating the energy surface. We evaluate the appropriateness of similarity measures in fitting the rough energy surface though the use of these measures for local approximation. Indeed the number of neighbors of each instance affects the approximation accuracy at this instance, which depends on the similarity measure used. Hence, relying on this, we hypothesize that to effectively approximate the energy function, similarity measures need to preserve the distinction of two objects in comparison with the third one.

Relying on the dependency among data presentation, similarity measure, and learning model, we propose a protocol for verify the proposed hypothesis, which includes two main steps: (1) examining the roughness of target function; and (2) evaluating appropriateness of similarity measures in fitting rough target function. The roughness of target function is quantitatively evaluated by estimating the roughness level based on function derivative, examining whether the target variable distribution is close to uniform, and examining whether the target function can be approximated by a linear function. For investigating similarity measures, we interpret empirical performance of $k-$nearest neighbors and kernel ridge regression which use these measures.

Inspired by the fixed-radius nearest neighbors regression, we propose a method for measuring the loss of distinction of two objects in comparison with the third one when using similarity measures. Let $DMIN$ be the distance from each data point to its closest neighbors, we enlarge the neighboring region of this data point for determining other neighbors with a radius of $DMIN \times (1 + \varepsilon)$, ε is a predefined scalar. This loss is defined as the average of the number of neighbors of each instance in a given dataset. This is used for evaluate similarity measures.

The experimental results show the high likelihood of our proposed hypothesis. That can help to explain why the Manhattan distance and Bray-Curtis dissimilarity provide better prediction performance with most of descriptors and material datasets. Furthermore, we establish general principles for effectively using similarity measures for mining material data, which do not depend on any specific learning method.

1.4.2 Controlling for confounding in assessing the cause of drug adverse reactions

In this study, we concentrate on an essential problem in post-marketing pharmaceutical surveillance (i.e., pragmatic clinical trials) – assessing the causality between drugs and adverse drugs reactions (ADRs) observed during the treating process. The main difficulty is that the presence of co-morbidity in patients requires polypharmacy for treating, hence, we have to face with the problem of confounding factors in the statistical causality inference. The polypharmacy-induced confounding makes most of existing methods poor in detecting actually causal drug-ADR pairs. Therefore, reducing bad impacts of confounding factors on the causality inference process motivates our work.

In our work, confounding factors are defined as non-causal drug-ADR pairs which frequently and coincidentally co-occur in treatment. To control confounding factors, we need additional references which are constructed based on the analogy criterion – one of nine Bradford Hill criteria. The analogy criterion states that similar drugs may causes similar ADRs, hence, a drug is believed more to cause an ADR if we found other drugs that are similar to the drug of interest and also have associations with the ADR. Therefore, in this context, we use similar drugs with their associated and ADRs (extracted from clinical narratives) as the reference for recognizing confounding factors.

We propose a novel semi-supervised model for inferring drug-ADR causality based on the analogy criterion, called the analogy-based active voting (AAV). This model represents this criterion as a voting process of similar drugs, in which similar drugs vote for a drug-ADR association to be causal if they also have the association with the ADR. The set of similar drugs is called the committee. We present each drug by two features: the mechanism of actions and targets; and the list of its associated ADRs extracted from electronic medical records. The first feature is used to identify similar drugs for establishing committees. The second one is used for voting of these drugs.

For effectively controlling confounding factors, we hypothesize that it is needful to diversify committee according to the feature of its drugs. It aims to create a strict inspection for distinguishing causal drug-ADR pairs from non-causal ones, and then can help to improve the causality inference performance. Similar to active learning, we need to select similar drugs that maximize the committee in terms of the second feature (list of associated ADRs). Because the second feature of each drug is bag of associated ADRs denoted by F_{x_i} where x_i indicates a drug, we measure the diversity of committee by considering the intersection of F_{x_i}, and using the Hamming distance for modeling the conflict in voting of each drug in committee with the rest. By using the Hamming distance, sets F_{x_i} is represented by one-hot vectors. The experimental results show that the use of diverse committee results in higher accuracy in detecting causal drug-ADR pairs. In other words, this shows the high likelihood of our proposed hypothesis.

1.4.3 Dissertation structure

The dissertation includes five chapters, in which Chapters 3 and 4 present the main content of our work. In Chapter 2, we make an overview of diversity preservation in machine learning via existing studies. In Chapter 3, we demonstrate the problem of inconsistency of similarity estimated based on object representation with the similarity of target values via the study on measuring materials similarity for predicting their formation energies. In Chapter 4, we discuss about the reference diversification in analogy-based causality inference through the study on controlling polypharmacy-induced confounding in assessing the cause of adverse drug reactions. Chapter 5 shows conclusions and future work.

Chapter 2

Diversity Preservation in Machine Learning

2.1 Introduction

As mentioned in previous chapter, diversity preservation in machine learning refers to the selection of data samples the maximize information contained for training process; the selection of models whose parameters can capture the information in training data as much as possible; the selection of models ensembles in which the output of each based model is different from that of the others. In this chapter, we make an overview of data diversification, model diversification, inference diversification in machine learning, and methods for measuring the diversity. This inspires our work on investigating the diversity preservation in similarity-based inference.

2.2 Diversifying data in machine learning

2.2.1 Increasing the number of dimensions in representation

Data representation in terms of computer science refers to methods to structure data for storing, processing, and transmitting by the computer. Data representation is almost the first step in machine learning and data mining. There are various forms for representing data, in which the vector form are widely used. In this representation, each object is described by a number of attributes, each attribute corresponds to a dimension of vector. Each attribute is considered an aspect for comparing data instances. For example, in healthcare data, a patient can be described by a vast amount of variables (or attributes), e.g., blood pressure, weight, cholesterol level, etc. Typically, data can be represented as a table or matrix whose columns represent dimensions. High-dimensional data simply means that the number of dimensions are staggeringly high.

As the common sense, the concept of diversity refers to different things, hence, we can define the concept of data diversity as the distinctiveness of data instances. However, it is trivial to say that because instances are manifestly distinct, and duplicated instances are usually removed. What we would like to mention here is how the number of dimensions is associated with the distinct level of instances. For example, let $A, B \in \mathbb{R}^3$ be two

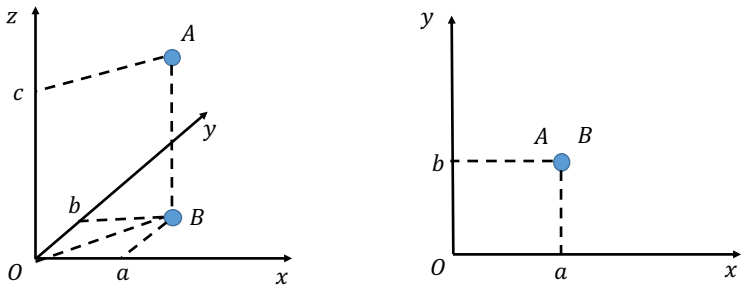

Figure 2.1: Visualization of two distinct points A and B in 3D space, and project them to 2D space.

distinct 3-dimensional vectors, which are plotted in 3D space as shown in Figure 2.1. If we project A and B to the plane Oxy, these points stack up. In fact, higher the number of dimensions is, more distinct instances are.

We demonstrate how the distinctiveness of data instances is reflected in high-dimensional space through investigating the pairwise distance between such instances. In 2-dimensional space, the distance between two points is $\sqrt{\Delta x^2 + \Delta y^2}$. When the third dimension is added, this extends to $\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}$, which is probably larger. Hence, the pairwise distances grow in high-dimensional space. To confirm that, we randomly generate a data sample with 10000 dimensions, and then we project such a sample to hyperplanes of 2, 5, 10, 100, and 1000 dimensions by using principal component analysis (PCA). The distributions of pairwise distances between data points in the original space and projected spaces are shown in Figure 2.2. The figure shows that pairwise distances increase when increasing the number of dimensions. In addition, histograms also show that in the low-dimensional spaces (the number of dimensions is 2, 5, 10, 100) the range of pairwise distances is larger than that in the high-dimensional spaces (the number of dimensions is 1000, 10000). In context of k-nearest neighbors method, the narrowed range of pairwise distances between points in high-dimension space means that all points in the dataset are almost equidistant to the query point (the point needs to be predicted its label or target value). That means the distinction between close points and distant points to the query point is small, which is called "contrast loss" [63]. The contrast loss can hinder the clustering or other machine learning techniques for generalizing data into patterns. No generalization reflects the distinctiveness of data instances.

2.2.2 Active learning

Active learning is a semi-supervised method that helps to reduce the cost of labeling data by selecting informative samples for human to assign their labels. Samples are selected based on examining that they can enrich information contained in the current data that is known as enhancing the diversity of data. To this end, several criteria are used for selecting samples. The first one is uncertainty sampling that select samples which the current model θ is the most uncertain about. There are various ways to measure the uncertainty in predicting labels of new samples. If the model is characterized as a hyperplane, we can estimate the distance from new instance to this hyperplane. In addition, we can measure

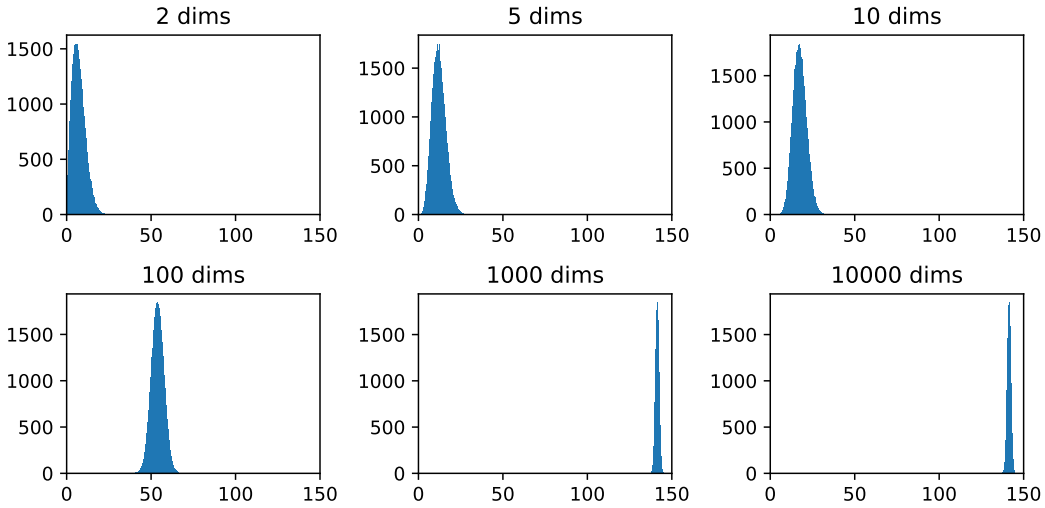

Figure 2.2: Histogram plots show the distribution of all pairwise distances between randomly distributed points in d-dimensional space.

the uncertainty based on labels probabilities as follows:

- Least confident:
$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\, 1 - P_\theta(\hat{y}|x) \tag{2.1}$$
where $\hat{y}$ is the most probable label for x under the current model θ.

- Smallest margin:
$$x^*_{SM} = \underset{x}{\operatorname{argmin}}\, P_\theta(y_1|x) - P_\theta(y_2|x) \tag{2.2}$$
where y_1, y_2 are the two most probable labels for x under the current model

- Label entropy: choose example whose label entropy is maximum:
$$x^*_{LE} = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) log P_\theta(y_i|x) \tag{2.3}$$
where y_i ranges over all possible labels.

Besides, samples can be selected based on query by committee (QBC). QBC uses a committee of models $C = \{\theta^1, ..., \theta^n\}$. All models are trained by using the current labeled data, and then vote their predictions on the unlabeled data. Examples with maximum disagreement are chosen for labeling. The disagreement is measured by the vote entropy:

$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} log \frac{V(y_i)}{C} \tag{2.4}$$

where y_i ranges over all possible labels, $V(y_i)$: number of votes received to label y_i.

2.3 Diversifying models in machine learning

In addition to diversify the data for learning by adding informative samples, we can also diversify models to enhance the representation ability of these models. Indeed, diversifying model includes diversifying parameters of an single model and diversifying an ensemble of models.

Diversifying parameters can enhance the representation ability of models because this helps to capture the information from the data as much as possible. Additionally, it makes the model complex and flexible. Obviously, increasing the flexibility of model is equivalent to enhancing the representation ability of this model. To enforce the diversity of parameters, we can utilize Bayesian method, posterior regularization method [33], diversity regularization by using distance-based measurement [17], angular-based measurement [108], etc.

For effectively solving real-world problems, ensembles of models are often used. Diversifying based models in these ensembles can help to improve the performance. Indeed, this induces various representations of data. A well-known method that are commonly used for encouraging the diversity of ensembles is sample-based methods that attempt to generate different models by randomly dividing the training data into subsets.

2.4 Diversifying inference in machine learning

Diversifying data and models can help to improve the performance of machine learning models. Besides, there has several methods that focus on obtaining multiple choices in inference of machine learning models. By using machine learning models, the predicted labels of data instances often converge to sub-optimal results because of the limitation of data and representation ability of the models. Hence, it motivates the use of multiple choices in inference of machine learning models. Diversifying choices can minimize the risk in prediction of models. There are several methods for diversifying the choices in inference such as diversity-promoting multiple choice learning (D-MCL), submodular, M-Modes, and M-NMS [36].

2.5 Measuring the diversity

Quantifying the diversity in machine learning is important, however, it is not straightforward, and depends on a specific domain. Measuring the diversity has attracted the interest of researchers for a long time such that many models have been proposed for attempting to measure such a property. Nevertheless, so far we have no a general measure method for this property.

As mentioned above, diversifying result in recommendation systems is important, which attracts many studies on this. Most studies attempted to propose quantitative models for ranking items that makes a tradeoff between the relevance (i.e., utility) and the diversity. Agrawal *et al.* [2] proposed an algorithm with an explicit objective of tradeoff between the relevance and the diversity. Additionally, they generalized several classical information retrieval metrics to explicitly account for the value of diversification. Carbonell *et al.* proposed a ranking method which combines query-relevance with

information-novelty for text retrieval and summarization [22]. The method is called the maximal marginal relevance. In [18], an algorithm, called optSelect, was proposed with a novel utility measure. This allows the diversification task to be accomplished effectively.

The diversity can be quantified by pairwise distance measures [105, 110]. Given elements in a collection, the diversity in such a collection is defined as an aggregate function (e.g., sum) of pairwise distances between elements. In addition, coverage-based measures have been used for estimating the diversity [76, 102]. This measure relies on the existence of a predefined number of aspects, that is, topics, interpretations, or opinions. In [57], the authors pointed out the connection between the diversity in classifier ensembles and accuracy. Furthermore, they presented ten statistics that can measure the diversity of binary classifier outputs, in which: averaged pairwise measures include the Q statistic, the correlation, the disagreement and the double fault; and non-pairwise measures include the entropy of votes, the difficulty index, the Kohavi-Wolpert variance, the interrater agreement, the generalized diversity, and the coincident failure diversity.

Chapter 3

Measuring Similarity: The Need of Preserving Objects Distinction in Reference-based Comparison

3.1 Introduction

In this chapter, we make an intensive discussion about an important problem in similarity-based learning (i.e., instance-based learning). The problem is that the use of inappropriate combination of representations and similarity measures can make objects with significant differences in their target values lost the distinction. This induces abrupt changes in the target surface (rough surface), which makes similarity-based learning methods become ineffective. To solve this problem, we focus on finding an appropriate similarity measure for a given previously designed representation. We found that it is needful to make a trade-off between the preservation of objects distinction when comparing them using a referenced object and the loss of this distinction. We quantify the loss of this distinction when using a similarity measure. To validate our statement, we employ a protocol that aims to point out the relation among: the roughness of target surface with a given representation; the loss of objects distinction as mentioned above; and the high predictive accuracy.

The problem mentioned above is demonstrated in our study on approximating the materials' formation energy surface that is evaluated to be rough towards most existing material representations (e.g., orbital field matrix, Coulomb matrix, and smooth overlap of atomic positions). We investigate several well-know dissimilarity measures (e.g., $p-$norm, chebyshev, cosine distances, Bray-Curtis and Canberra dissimilarities) where these measures are used in similarity-based learning models such as $k-$nearest neighbors regression and kernel ridge regression for predicting formation energies. The empirical experiments with several well-known material datasets and representations show the high potential of our finding as mentioned above. In addition, relying on this, we propose a policy for effectively designing similarity measures for material data.

Distance metric learning (DML) is a class of similarity measure learning methods that is useful for dealing with the problem of inconsistency of representation-based similarity measurement with the similarity of target values. Indeed, DML aims to learn an appropriate Mahalanobis distance between object representations to maximize the consistency

with the similarity of their target values. By examining the model complexity of kernel ridge regression using the learned Manhalanobis distance, we found that the appropriate distance for represented materials makes the model complexity increase. This is consistent with and supports our investigations of selecting available similarity measures.

3.2 Overview of similarity measurement in materials data mining

A small change in the chemical composition or structure of materials can lead to a significant change in the properties of materials. For example, differences in the chirality of a honeycomb network of carbon atoms can lead to a distinctive difference in physical properties of nanotubes. In fact, the distinctiveness of materials, which makes the diversity of materials in the nature, is the main characteristic of the material data. Therefore, this characteristic needs to be represented in a metric that allows for a comparison of materials in a reliable, efficient, and useful way.

The main target of machine learning systems when mining material data is to determine a likely function $f(x)$, which indicates the relation between the materials' attributes and their physical/chemical properties. Typically, these systems includes two main components: (i) data representations which are also called descriptors; and (ii) operators including similarity measures between materials and learning methods (which map materials' attributes to physical properties). For efficient mining, these components are designed with the aim of reflecting domain knowledge and the nature of material data.

To render computational methods tractable for materials in datasets, the geometrical, topological, or electronic characteristics of the materials need to be represented in form of numerical variables. Descriptors commonly encode the information of a material A by a vector $\vec{x}_A = (x_A^1, x_A^2, ..., x_A^m)$ whose number of dimensions, and values in each dimension depend on the information selected to describe the materials with a specific purpose for mining tasks. To represent material structures, several descriptors have been proposed. Behler *et al.* utilized atom-distribution-based symmetry functions to represent the local chemical environment of atoms [12]. Rupp *et al.* proposed the Coulomb matrix (CM), which represents materials via the Coulomb repulsion between all possible nuclei in the material [88]. Bartok *et al.* proposed the smooth overlap of atomic positions (SOAP) that is effective to represent molecules [10, 26]. In addition, Isayev *et al.* used the band structure and density of states (DOS) fingerprint vectors as descriptors of materials to visualize material space [47]. Zhu *et al.* introduced another fingerprint representation for crystals and used this to define the configurational distance between crystalline structures [109]. Pham *et al.* proposed a descriptor for encoding atomic orbital information, called the orbital field matrix (OFM) [59, 80].

Similarity measures aim to quantify how alike two materials are, which are mathematically implemented as scalar valued functions that take two vectors representing materials A and B as input: $S(A, B) = S(\vec{x}_A, \vec{x}_B)$. The use of these measures is subjective because they depend on a specific domain or application. Similarity measure is an important operator in many learning models. Conventionally, materials science studies begin by grouping similar materials in order to explore the patterns and rules in these

materials. Consequently, measuring material similarity is considered a key technique in material informatics [13]. The advantages and disadvantages of many similarity measures were addressed in [70] and the argument that similar structures lead to similar properties was offered in [9, 99]. However, the validity of this argument was reconsidered by Maggiora *et al.*, who showed that small chemical modifications can lead to significant changes in biological activity [69]. Because the nature of materials is fundamentally diverse, Riniker *et al.* addressed the problem of partially losing the transparency among fingerprint types by using fuzzier similarity methods [86]. In addition, Maldonado *et al.* optimized measures of molecular similarity and diversity based on selecting and classifying descriptors [72]. Moreover, several methods have been proposed for comparing crystalline materials [58, 109].

Although similarity measures are disseminated in many studies in machine learning, to the best of our knowledge, most previous work rarely makes the discussion about properties of these measures, and why they perform well in specific contexts. It makes the explanation and interpretation of these measures poor, so this motivates our work to overcome such a limitation.

3.3 Roughness of target function subject to representation

3.3.1 Similarity-based inference

Similarity assessment is a fundamental operator in recognition. In machine learning, similarity-based inference has been widely used in various learning methods such as clustering, $k-$nearest neighbors methods, kernel methods. The principle of the similarity-based inference is that similar instances result in similar target values. Hence by this inference, unknown target values of new instances are inferred by finding neighbors (similar instances of the ones of interest) using various similarity measures. In practice, for similarity-based learning methods to effectively perform, representation and similarity measure need to reflect the nature of data. However, finding appropriate representation and similarity measure is not straightforward that requires intensively digging into the nature of data for understanding.

3.3.2 Problem Statement

Given a collection of objects $O = \{o_1, ..., o_n\}$, and $T = \{T_1, ..., T_n\}$ is the set of target property corresponding to each object in O. Suppose that there has a function $f(o)$ with $o \in O$ that maps each object to its target value: $f : O \to T$.

To approximate the function f, we first encode objects in O in various forms, in which the vector form is widely adopted. Let r be a representation function that converts each object to its corresponding vector, $r : O \to V$ where V is the set of vectors. To preserve the identification of objects, two distinct objects o_i, o_j need to have different representations, $v_i \neq v_j$. The target properties in T are approximated by a model $\hat{f}$ on vectors: $T_i \approx \hat{f}(v_i)$ where $v_i \in V$. To find the function $\hat{f}$, if T_i and T_j are different, v_i and v_j must be different:

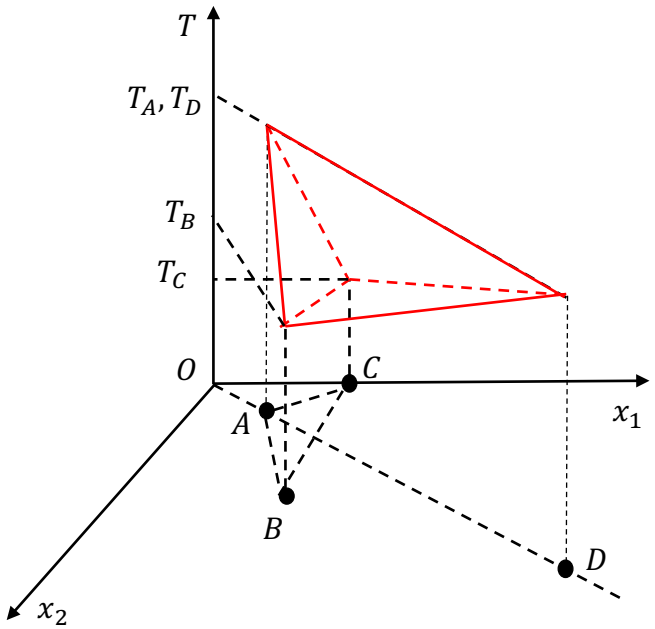

Figure 3.1: The illustration of abrupt changes in the target surface that is induced by the use of inappropriate combination of representation and similarity measure.

$$T_i \neq T_j \Rightarrow v_i \neq v_j \tag{3.1}$$

If we find two objects o_i and o_j that $v_i = v_j$ and $T_i \neq T_j$, it is hard to find an appropriate function $\hat{f}$.

Inappropriately using the combination of representation and similarity measure

As mentioned above, a small change in material structure can lead to significant change in material physical property. Suppose that the change between two objects o_i and o_j is defined by the Euclidean distance between v_i and v_j, denoted by $d_{euc}(v_i, v_j)$. Let consider an example shown in Figure 3.1, when using inappropriate representation, $d_{euc}(A, B)$ and $d_{euc}(A, C)$ are small while $|T_A - T_B|$ and $T_A - T_C$ are significantly large. This makes estimating the T_A based on T_B and T_C is imprecise. In general, inappropriately using representation method can make the surface indicating the target property become rough, and then makes similarity-based learning methods ineffective.

Suppose, with this representation, we use the cosine distance instead of Euclidean distance. Thus, we have $d_{cos}(A, D) = 0$ and $d_{cos}(A, D) < d_{cos}(A, B), d_{cos}(A, C)$. In addition, $T_A = T_D$, so we can properly infer T_A based on T_D. That means the cosine distance is more appropriate than the Euclidean distance for this representation.

We address the problem that similarity-based learning methods become ineffective because of the inappropriate use of combination of representations and similarity measures that poorly distinguish materials with significant differences in physical property values. There are two potential approaches for solving this problem: (i) finding an appropriate representation; (ii) finding an appropriate similarity measure for a given representation.

To find an appropriate representation of objects, it requires the prior knowledge about the nature of data. In fact, this knowledge is almost hidden, even though is out of our knowledge, hence, finding an appropriate representation is extremely difficult. In case of lacking prior knowledge, finding an appropriate similarity measure for a given representation is a potential solution. In this study, we concentrate on this approach.

Objectives

For selecting an appropriate similarity measure, simply, we can use the performance of existing similarity measures when they are used in instance-based learning methods. In fact, this criterion is just applicable for selecting available similarity measures, and is meaningless for interpreting and explaining the nature of data. Hence, we demand a more informative criterion that helps to gain insight into the nature of data. In addition, this criterion can be useful for designing new similarity measures.

3.4 Hypothesis on the influence of preserving the instances distinction in their reference-based similarity evaluation on the performance of similarity-based learning methods

As mentioned above, we attempt to explore an informative criterion for similarity measure selection, which is useful for gaining insight into the nature of data. In this section, we aim to clarify how similarity measures preserve the distinction of two instances when comparing them using a specific instance as a reference. In addition, we hypothesize the dependency of similarity-based model performance on this characteristic.

3.4.1 A comparison between the Manhattan distance and Euclidean distance in terms of their ability of preserving instance distinction in the reference-based similarity evaluation

We demonstrate the preservation of the distinction between two particular objects when these objects are compared using another object as a reference by considering the Manhattan (1-norm) distance and Euclidean (2-norm) distance. As illustrated in Figure 3.2, let $S = \{X | d_{euc}(X, O) \leq r\}$ be the set of points X whose Euclidean distances to O are smaller or equal than r; and $S' = \{X | d_{man}(X, O) \leq r\}$ be the set of points X whose Manhattan distances to O are smaller or equal than r. We compare two points A and B only based on their distance to O, thus, O is known as a referenced point for this comparison. By using the Euclidean distance, we see that A and B lost the distinction because of $d_{euc}(A, O) = d_{euc}(B, O)$. Meanwhile, by using the Manhattan distance, we still preserve the distinction of A and B because of $d_{man}(A, O) = d_{man}(B, O)$. Hence, we conclude that the Manhattan distance preserves the distinction between A and B when comparing them based on their distance to O, but the Euclidean does not.

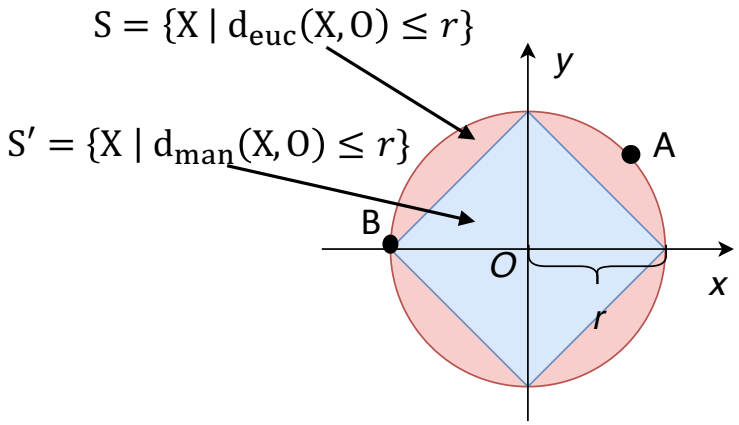

Figure 3.2: The comparison between the Manhattan distance and Euclidean distance in terms of preserving the instances distinction when comparing these instances using a referenced instance. Area of the blue square indicates the set of instances that the Manhattan distance between these instances and O is smaller or equal than r. Area of the red circle indicates the set of instances that the Euclidean distance between them and O is smaller or equal than r.

3.4.2 The influence of preserving the instance distinction in their reference-based similarity evaluation on the performance of similarity-based learning models

Suppose that we need to estimate the target value at the instance O that is denoted by y_O. If r is small enough, S and S' are the sets of O's neighbors determined by the Euclidean and Manhattan distances, respectively. With neighbors in S (or in S'), the value of y_O can be estimated as the following:

$$\begin{aligned} \hat{y}_O &= \frac{1}{|S|} \times \sum_{X \in S} y_X \\ &= \frac{1}{|S|} \times \left(\sum_{X_+ \in S_+} y_{X_+} + \sum_{X_- \in S_-} y_{X_-} \right) \end{aligned} \tag{3.2}$$

where $S_+ \subset S$ is the set of neighbors in S whose target values are greater or equal than y_O, $y_{X_+} = y_O + \sigma_{X_+}$ with $\sigma_{X_+} \geq 0$; $S_- \subset S$ is the set of neighbors whose target values are smaller than y_O, $y_{X_-} = y_O - \sigma_{X_-}$ with $\sigma_{X_-} > 0$. Let $\sigma_{S_+} = \sum_{X_+ \in S_+} \sigma_{X_+}$, and $\sigma_{S_-} = -\sum_{X_- \in S_-} \sigma_{X_-}$. Actually, to precisely estimate y_O, the ground truth is that the sum of σ_{S_+} and σ_{S_-} approaches to 0, $\sigma_{S_+} + \sigma_{S_-} \approx 0$. Hence, if O is an extremum, this sum is significantly different from 0 because either S_+ or S_- is empty. Such situation is undesirable in which we cannot estimate y_O correctly.

As mentioned above, the use of inappropriate descriptor can induce the roughness of the target surface. In other words, there has many abrupt changes of target values in the neighboring region of O. In fact, these changes can produce the previously mentioned undesirable situation in which the sum of σ_{S_+} and σ_{S_-} is significantly different from 0, consequently, the estimation of y_O is inaccurate. Therefore, we should determine neigh-

boring regions of O in which the changes of target values are as small as possible. If O is an extreme point, we expect that the number of neighbors used for estimating y_O is also as small as possible.

Given the same value r, area of the neighboring region of O determined by the Euclidean distance is larger than that determined by the Manhattan distance, as shown in Figure 3.2. With the larger neighboring region, instances that have the small distance to O but significantly different target values to O and other neighbors of O have a high chance to be included in this region. By using the Manhattan distance that produces a smaller neighboring region, this chance can be lowered. Hence, we hypothesize that to improve the performance of similarity-based models for fitting rough target surface, it is needful to select similarity measures that preserve the distinction of two particular instances when they are compared using a specific instance as a reference. However, the exceedingly preserving this distinction can lead to the overfitting problem because the number of neighbors used for inferring the target value at the instance of interest is extremely small. This induces the low prediction performance. Hence, the function indicating the dependency of the predictive performance on this distinction has an unique extreme point. At this point, the prediction accuracy attains the highest value.

3.5 Protocol for validating the proposed hypothesis

We demonstrate the proposed hypothesis mentioned in previous section through the study on selecting appropriate similarity measures used in instance-based learning methods for predicting materials' formation energies. To validate this hypothesis, we need to point out the relation among: (i) the roughness of the target surface indicating the material formation energy given a representation of materials; (ii) the loss of instance distinction in reference-based similarity evaluation when using each similarity measure; and (iii) the high formation energy prediction accuracy. To this end, we employ a protocol, in which we attempt to quantify the loss mentioned in (ii), and to derive features indicating factors (i) and (iii). This protocol takes into account several well-known material representations, (dis)similarity measures, and similarity-based learning methods (e.g., $k-$nearest neighbors regression, kernel ridge regression). In this section, firstly we introduce several well-known material representations and dissimilarity measures that are used for evaluating the similarity of objects. Next, we present about components in the proposed protocol in detail.

3.5.1 Material representation (descriptor)

Material descriptors aim to represent or encode real materials into mathematical forms for computation, in which, vector form is widely utilized, as shown in Figure 3.3. In this study, we investigate three well-known material descriptors: *orbital field matrix*; *Coulomb matrix*; and *smooth overlap of atomic positions*. To the best of our knowledge, these descriptors often help to improve the performance of machine learning models for predicting materials' physical/chemical properties.

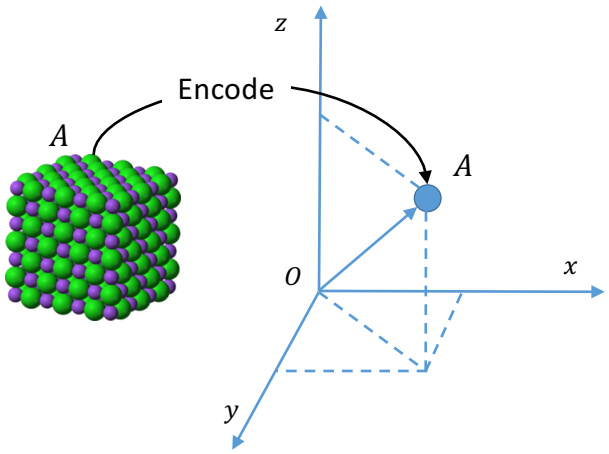

Figure 3.3: Representing materials in the vector space.

Orbital field matrix

The orbital field matrix (OFM) is a novel descriptor that was proposed recently [59, 80], which uses the valence atomic configuration to represent the structure of materials. In the OFM descriptor, a material is assumed to be composed of building blocks that are called local structures. Each local structure includes a central atom and its environmental (or neighboring) atoms. First, each atom is represented by a one-hot vector based on a dictionary of subshell orbitals: $D = \{s^1, s^2, p^1, ..., p^6, d^1, ..., d^{10}, f^1, ..., f^{14}\}$. We denote the vector of the central atom by $\vec{O}_{central}$, and the vector of the k^{th} neighboring atom by $\vec{O}_k$. Second, the vector representing the environment of each atom in a structure, $\vec{O}_{env}$, is computed as follows:

$$\vec{O}_{env} = \sum_{k}^{K} w_k \vec{O}_k, \tag{3.3}$$

where the weight, w_k, measures the contribution of the k neighboring atom, and K is the number of neighboring atoms. The local structure is represented by a matrix, X, where X_{ij} represents the number of an environment atomic orbital (orbital j) coordinated with a central atomic orbital (orbital i). Hence, the representation matrix of a local structure is

$$\begin{aligned} X &= \vec{O}^T_{central} \times \vec{O}_{env} \\ &= \vec{O}^T_{central} \times \Big(\sum_{k}^{K} \vec{O}_k \frac{\theta_k}{\theta_{max}} \Big) \end{aligned} \tag{3.4}$$

where $w_k = \frac{\theta_k}{\theta_{max}}$ is the weight representing the contribution from atom k^{th} to the coordination number of the central atom; θ_k is the solid angle determined by the face of the Voronoi polyhedral that separates the k^{th} atom and the central atom; and θ_{max} is the maximum of all solid angles determined by this Voronoi polyhedral.

The distance r_k between the central atom and the k^{th} neighboring atom is incorporated

in the representation of local structures as follows:

$$X = \vec{O}^T_{central} \times \Big(\sum_k^K \vec{O}_k \frac{\theta_k}{\theta_{max}} \zeta(r_k) \Big), \tag{3.5}$$

where $\zeta(r_k) = 1/r_k$ is the distance-dependent weight function. Finally, the descriptor for the entire material is a mean of descriptors for its local structures.

In an extension of the OFM, the information regarding the central atom is incorporated by simply concatenating $\vec{O}^T_{central}$ to the matrix X as a new column, as follows:

$$X = \vec{O}^T_{central} \times \Big(1.0, \sum_k^K \vec{O}_k \frac{\theta_k}{\theta_{max}} \zeta(r_k) \Big) \tag{3.6}$$

In this study, we use this extension to the OFM to predict crystals' formation energies.

Coulomb matrix

The Coulomb matrix (CM) [88, 74] is a descriptor that encodes the structure of a material using nuclear charges Z_i and the 3D coordinates $\boldsymbol{R}_i$ of each constituent atom in the material, as follows:

$$C_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \forall i = j \\ \frac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|} & \forall i \neq j \end{cases} \tag{3.7}$$

To deal with the atom-ordering problem in CM, the authors used (i) the eigenspectrum representation that first obtains eigenvalues of each Coulomb matrix, and then uses the sorted eigenvalues (i.e., spectrum) as the representation, and (ii) sorted Coulomb matrices that choose the permutation of atoms whose associated Coulomb matrix C satisfies $||C_i|| \geq ||C_{i+1}|| \quad \forall i$ where C_i is the i^{th} row of the Coulomb matrix. In practice, padding the Coulomb matrices by zero-valued entries is required in order to avoid the difference in matrix size induced by the difference in the number of atoms in each material.

Smooth overlap of atomic positions

Smooth overlap of atomic positions (SOAP) [10, 26] is a descriptor that encodes regions of atomic geometries by using local expansion of Gaussian smeared atomic density with orthonormal basis functions based on spherical harmonics. This is based on the similarity kernel between two environments of atoms which is defined as the overlap of the two local atomic neighbor densities.

The output of this descriptor is a vector $p(\boldsymbol{r})$, where the different elements are formed from the partial SOAP power spectrum defined as[1]:

$$p(\boldsymbol{r})^{ZZ'}_{nn'l} = \pi \sqrt{\frac{8}{2l+1}} \sum_m c^Z_{nlm}(\boldsymbol{r})^\dagger c^{Z'}_{n'lm}(\boldsymbol{r}), \tag{3.8}$$

where $\boldsymbol{r}$ is a atomic position in space; c^Z_{nlm} are the expansion coefficients of the Gaussian smoothed atomic density at position $\boldsymbol{r}$ that is expanded in the basis of spherical harmonics

[1] https://singroup.github.io/dscribe/tutorials/soap.html

and orthonormal radial basis functions; n and n' are indices for different radial basis functions up to n_{max}; l is the angular degree of spherical harmonics up to l_{max}; Z and Z' are atomic species. This form ensures stratification of the output by species and also provides information about cross-species interaction.

3.5.2 Similarity measures of interest

Commonly, the similarity between two objects is assessed by estimating the difference of two objects. There are various methods for quantifying the dissimilarity between two objects. In dissimilarity measures, ones are called the distance if they satisfy all conditions of a metric that include non-negativity, identity of indiscernibles, symmetry, and triangular inequality. Besides, there are dissimilarity measures which are asymmetric and do not obey the triangular inequality.

In this study, because materials are represented by numerical vectors, we focus on distances or dissimilarity measures used for numerical vectors. Let $u, v \in \mathbb{R}^m$ be two vectors, dissimilarity measures are essentially mathematical functions that take the vectors u and v as their input then produce a scalar as their output. In our work, we investigate several well-known distances and dissimilarity measures which are commonly used for numerical vectors, as follows:

- $p-$norm distance

$$d(u, v) = ||u - v||_p = (\sum_{i=1}^{m} |u_i - v_i|^p)^{\frac{1}{p}} \tag{3.9}$$

  with $p = 1, 2, 3$ in which the 1-norm and 2-norm are known as the Manhattan and Euclidean distances, respectively.

- Cosine distance

$$d(u, v) = 1 - \frac{u.v}{||u||_2||v||_2} \tag{3.10}$$

  Noting that $cos(u, v) = \frac{u.v}{||u||_2||v||_2}$ measures the similar between u and v.

- Bray-Curtis (B-C) dissimilarity: this is not a distance measure because it does not obey the triangular inequality

$$d(u, v) = \frac{\sum_{i=1}^{m} |u_i - v_i|}{\sum_{i=1}^{m} |u_i + v_i|} \tag{3.11}$$

- Canberra distance

$$d(u, v) = \sum_{i=1}^{m} \frac{|u_i - v_i|}{|u_i| + |v_i|} \tag{3.12}$$

- Chebyshev distance

$$d(u, v) = \max_{i=1,\ldots,m} |u_i - v_i| \tag{3.13}$$

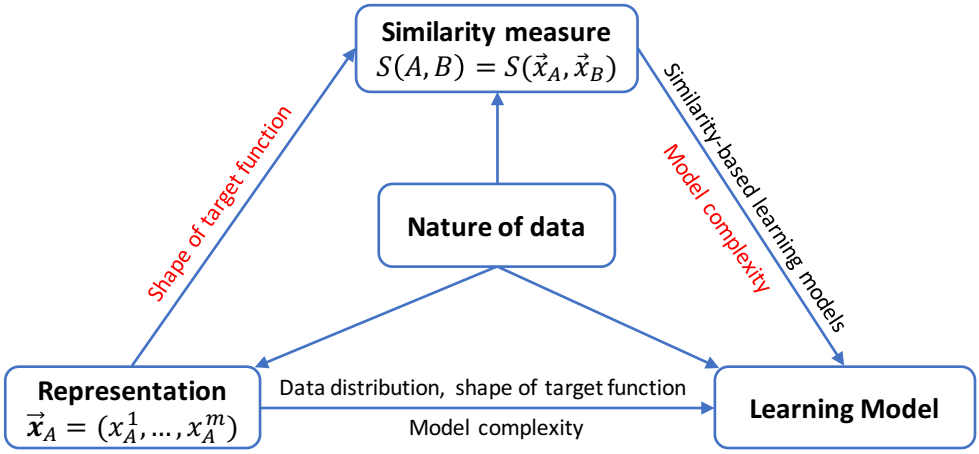

Figure 3.4: Directed graph showing the dependence of material descriptor, similarity measure, and learning model on each other, and the dependence of them on the nature of material data.

3.5.3 Dependency among data representation (descriptor), similarity measure, and learning method

Typically, for designing an appropriate machine learning system for solving a mining task, it is needful to consider the association among the data representation (i.e., descriptor), the similarity measure, and the learning model. In addition, these factors must reflect the nature of data. In this study, we demonstrate the association of the proposed criterion of dissimilarity measures with the representation and learning model. This helps to interpret and explain the performance of learning models with material datasets.

Regarding the association between data representation and learning model, in fact, different representations make different locations of data instances, in other words, result in different distributions of data instances. Hence, the shapes of the target function corresponding to each data distribution are also different. In fact, understanding about the shape of target function is useful to select learning models as well as their parameters. In addition, the complexity of learning models also depends on the data representation because it depends on the shape of target function. For example, if the data representation results in a rough surface of the target function, we need high-complexity models for fitting this surface.

Regarding the association between similarity evaluation and data representation, in fact, the similarity is evaluated by dissimilarity measures which are essentially real-valued functions taking two represented vectors as their input. Hence, obviously, selecting dissimilarity measures depends on the nature of representation. In addition, as mentioned above, data representations affect the distribution of data instance and the shape of the target surface. Indeed, the shape of target surface affects the selection of appropriate dissimilarity measures that we focus on clarifying in this study.

Regarding the association between the similarity evaluation and the learning model, so far many machine learning methods have been developed based on similarity assessment (inference) such as clustering, $k-$nearest neighbors, kernel methods, etc. These methods are often called the instance-based (or distance-based) learning methods. Hence, the

selection of appropriate dissimilarity measures plays an important role that affects the performance of these methods. In addition, in this study, we point out that dissimilarity measures selection affects the model complexity in terms of kernel method.

Take into account the dependency among the data representation, similarity evaluation, and learning model is important for model interpretation and explanation. In addition, this helps to select or design representations, measures for similarity evaluation, and learning model for attaining high prediction accuracy. In this study, we investigate dissimilarity measures based on analyzing their association with the nature of material data, material representations, and instance-based learning models, which is shown in Figure 3.4.

3.5.4 Protocol

To validate the proposed criterion of the nature of dissimilarity measures, we need to derive features that indicate the three factors as mentioned above. The features are shown in Figure 3.5 that include:

1. Sensitivity to the change of target values towards the change of instances
2. Likelihood of globally and linear approximating the target surface
3. Performance of $k-$nearest neighbors (KNN), and the number of neighbors of each data instances determined based on the fixed-radius nearest neighbors.
4. Kernel ridge regression performance

The features 1, 2 are used for evaluating the roughness of the surface that indicates the formation energy. The feature 3 is used for evaluating how dissimilarity measures preserve the distinction of objects in context-based comparison. The features 3, 4 show the performance of dissimilarity measures used in the instance-based learning methods. These features indeed affect the performance of instance-based learning methods.

To evaluate the roughness of the target surface indicating the formation energies, we estimate the sensitivity to the change of target values towards the change of instances by counting the number of neighboring regions of instances in which there has the significant change of target values. In addition, we evaluate the fluctuation amplitude of the energy surface towards a hyperplane through the global linear approximation using ridge regression.

We investigate the appropriateness of using dissimilarity measures in instance-based approximating the rough target surface by interpreting the accuracy obtained when using these measures in the $k-$nearest neighbors regression and kernel ridge regression. We examine various dissimilarity measures and kernel functions that are essentially constructed from dissimilarity measures. We consider the $k-$nearest neighbors regression and kernel ridge regression because these methods employ the local approximation of the target surface in which identifying proper neighbors plays an important role that depends on the dissimilarity measure or kernel function selection. Hence, we can evaluate the effectiveness of dissimilarity measures or kernel functions in the association with the learning method. By exploring the correlation between well-performing measures and their nature, we can

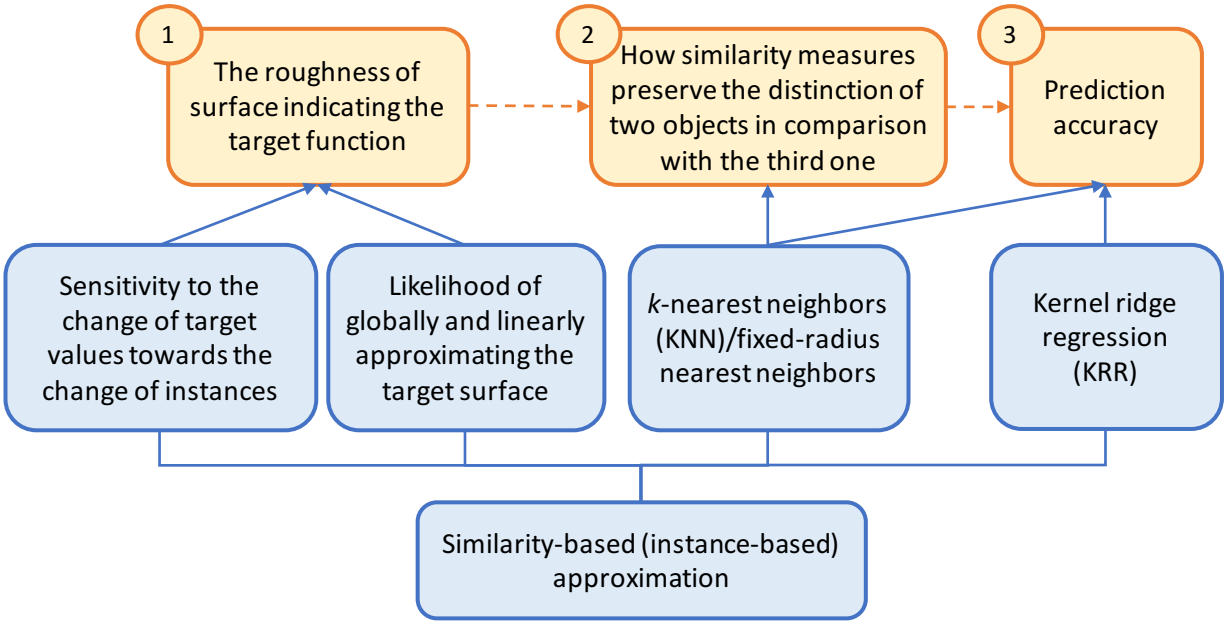

Figure 3.5: Protocol for examining the correlation among: the roughness of target surface indicating the material formation energies; the nature of dissimilarity measures as mentioned in the proposed criterion; and the formation energy prediction accuracy.

validate the proposed criterion. The detail of methods used in this protocol is provided in next sections.

3.5.5 Evaluating the roughness of the target surface

Sensitivity to the change of the target values towards the change of instances

Given a function $y = f(x)$, the derivative of the function f measures the slope of the surface indicating this function at each instance which is estimated by the ratio between the change of y with respective to the change of x. This is denoted by $\frac{\Delta y}{\Delta x}$. Inspired of this, we can investigate the change of target values in a neighboring region of each data instance. In this context, the change of x, Δx, is defined by the Euclidean distance from each data instance to its closest neighbor because this distance indicates the geometrical distance between two vectors. Let $(\boldsymbol{x}, y)$ with $\boldsymbol{x} \in \mathbb{R}^d$ be an instance in the dataset, and the instance $(\boldsymbol{x'}, y')$ be the closest neighbor of $\boldsymbol{x}$ according to the Euclidean distance. The ratio between the change of target values towards the change between $\boldsymbol{x}$ and $\boldsymbol{x'}$ as the following:

$$ChangeRatio(\boldsymbol{x}, \boldsymbol{x'}) = \frac{y' - y}{d_{euc}(\boldsymbol{x}, \boldsymbol{x'})}, \tag{3.14}$$

where $d_{euc}(\boldsymbol{x}, \boldsymbol{x'})$ is the distance from $\boldsymbol{x}$ to $\boldsymbol{x'}$.

Now we consider a neighboring region $\mathfrak{N}$ of $\boldsymbol{x}$ that additionally contains $\boldsymbol{x''}$ where $d_{euc}(\boldsymbol{x}, \boldsymbol{x'}) < d_{euc}(\boldsymbol{x}, \boldsymbol{x''})$. Next we estimate the $ChangeRatio(\boldsymbol{x}, \boldsymbol{x'})$ and $ChangeRatio(\boldsymbol{x}, \boldsymbol{x''})$, and then count the number of neighboring regions $\mathfrak{N}$ that satisfy one of following conditions:

- Change of sign: $ChangeRatio(\boldsymbol{x}, \boldsymbol{x'}) \times ChangeRatio(\boldsymbol{x'}, \boldsymbol{x''}) < 0$
- Change of magnitude:

$$\frac{ChangeRatio(\boldsymbol{x}, \boldsymbol{x'})}{ChangeRatio(\boldsymbol{x'}, \boldsymbol{x''})} \geq \zeta \text{ or } \frac{ChangeRatio(\boldsymbol{x'}, \boldsymbol{x''})}{ChangeRatio(\boldsymbol{x}, \boldsymbol{x'})} \geq \zeta$$

  We consider this change is significant if ζ is at least 10.

These conditions indicate the sensitivity to the change of target values with respect to the change of instances in $\mathfrak{N}$. Hence, given a dataset $\mathcal{D}$ with a representation r, we denote the number of neighboring regions of instances that satisfy one of above conditions by the $SensitivityToChange(\mathcal{D}, r)$.

3.5.6 Evaluating the likelihood of globally and linearly approximating the target surface

Suppose that the roughness of the surface is caused by adding noise to a hyperplane: $f(x) = linear(x) + \epsilon(x)$ where $\epsilon(x)$ is the noise. We locally approximate the function $f(x)$ by taking an average of target values at neighbors of each instance, or making a series of local linear approximation, as the following:

$$\hat{f}(x) = \frac{1}{N_k} \sum_{i \in N_k} y_i \approx \frac{1}{N} \sum_{i \in N_k} \langle w^{(N_k), x} \rangle, \tag{3.15}$$

where N_k is the set of nearest neighbors of x, and w^{N_k} is the linear coefficients obtained when linearly fitting neighbors in N_k.

The the fluctuation amplitude of $f(x)$ (as illustrated in Figure 3.6) values towards the hyperplane is small, local linear functions can be replaced by a global one. In contrast, we cannot find the global function that fits with the target surface. Hence, firstly we find the most likely hyperplane that fits with the target surface $f(x)$ using ridge regression. The fluctuation amplitude of $f(x)$ values towards this hyperplane is implicitly indicated through the prediction accuracy when performing the ridge regression.

Ridge regression is a parametric model that approximates the energy function by a linear function. In this method, the linear coefficients β are estimated to minimize the penalized residual sum of squares, as the following:

$$RSS(\lambda) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\beta^T\beta, \tag{3.16}$$

where the matrix $\boldsymbol{X}$ is the input data, and $\lambda \geq 0$ is a predefined parameter which indicates an amount of coefficient shrinkage towards zero (weight decay). Ridge regression has the following closed-form solution:

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{3.17}$$

where $\boldsymbol{I}$ is the identity matrix.

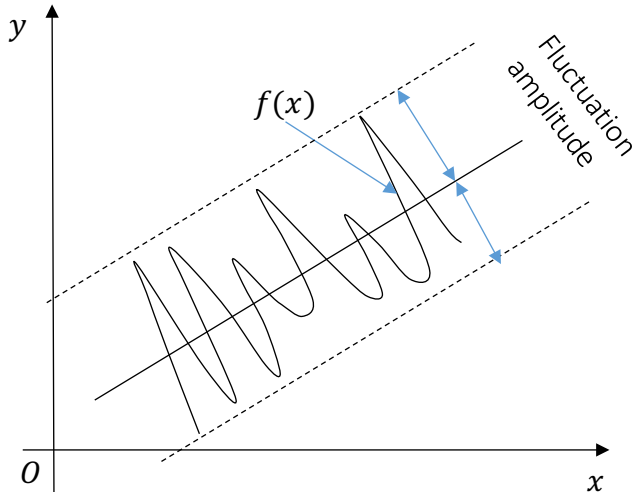

Figure 3.6: Illustration of the fluctuation amplitude of $f(x)$ values towards a hyperplane.

3.5.7 K-nearest neighbors regression

K-nearest neighbors (KNN) is known as a "lazy learning" algorithm that predicts a target value of an instance by averaging target values of nearest neighbors of this instance without any assumption of the relation between this instance and its target value. Let $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ be a sample data that is generated from a function $y = f(x)$. KNN locally approximates this function by predicting new instances x' as the following:

$$\hat{f}(x') = \frac{1}{|N_k|} \sum_{(x_i, y_i) \in N_k} y_i, \tag{3.18}$$

where $N_k \subset \mathcal{D}$ is the set of k nearest neighbors of x'.

Fixed-radius nearest neighbors − Instead of determining the number of nearest neighbors k, we can determine a neighboring region of each query point by a given predefined distance threshold. Data points which fall in this region are considered the neighbors of the query point. The target value at the query point is also estimated by taking an average of target values at these neighbors. This method is called the fixed-radius nearest neighbors regression [21].

3.5.8 Measuring the loss of instances distinction in their reference-based similarity evaluation when using similarity measures

In instance-based learning methods, we need to compare each data instance with the query instance to identify its neighbors. hence, we evaluate the similarity between two instances based on their similarity to the query instance. In fact, the query instance here is the reference for this comparison.

Inspired by the fixed-radius nearest neighbors method, and the intuition obtained by investigating the Manhattan and Euclidean distances, when using a similarity measure, we can measure the loss of the distinction between two instance where they are compared

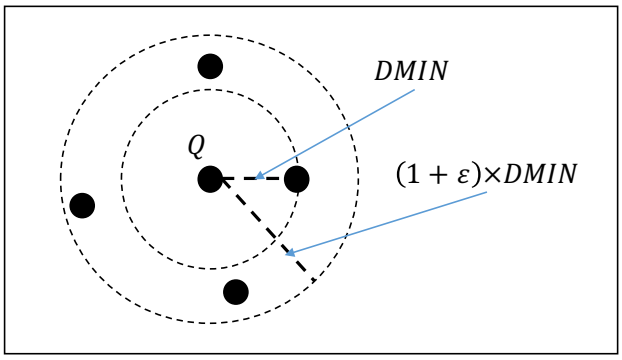

Figure 3.7: Illustration of the method used for estimating the loss of instances when they are compared using a specific instance as a reference, which corresponds to the use of each similarity measure.

using a referenced instance. Suppose that Q be an query instance that we aim to approximate the target value at. We determine the closest instance to the query instance by using a dissimilarity measure, and the distance (or dissimilarity) between these instances is denoted by $DMIN$. To determine other neighbors of Q, we enlarge a region surrounding Q by a radius $(1 + \varepsilon) \times DMIN$. This region is called the neighboring region of Q because data instances belonging to this region ae considered the neighbors of this query instance. The method is illustrated in Figure 3.7. In fact, different dissimilarity measures will determine different number of neighbors of a query instance given a ε. By this method, we can evaluate the effectiveness of dissimilarity measures based on the number of neighbors of the query instance they define.

Given a fixed value of ε for determining the neighboring region of the query instance Q, data instances are distinct when they are compared using the query instance as a a reference if there is a small number of instances that fall in such a region. On the other hand, if those instances are not distinct, we will find a large number of instances falling in this region. Inspired by this assessment, we propose a measure that quantifies how similarity measures preserve the instances distinction in their reference-based similarity evaluation. The loss of this distinction is defined by the average number of neighbors corresponding to each query instance when given a value of ε that is denoted by the $DLoss$. Taking a set of m random data samples $\boldsymbol{D} = \{\mathcal{D}_1, ..., \mathcal{D}_m\}$ into acount, the $DLoss$ is estimated as the following:

$$DLoss = \frac{1}{|\boldsymbol{D}|} \sum_{\mathcal{D}_i} \frac{1}{n_i} \sum_{j=1}^{n_i} \nu_j^{\varepsilon} \tag{3.19}$$

where n_i is the number of instances in the data sample $\mathcal{D}_i$, ν_j^{ε} is the number of instances falling in the neighboring region of each query instance x_j (in the data sample $\mathcal{D}_i$) that is determined by a given similarity measure and a value of ε. Note that the values produced by different similarity measures need to be normalized into the range of [0, 1] via dividing by the maximum value. Actually, the estimation of $DLoss$ depends on the distribution of a specific dataset. Therefore, to avoid this bias of the estimation, we use a set of random

data samples. Larger the average number of neighbors determined by ε is, greater the value of $DLoss$ is.

When the surface indicting the target function is rough, the small number of neighbors of an instance is preferred to use if this instance is an extreme point of the function. Hence, dissimilarity measures that have small $DLoss$ are appropriate in such a context. Although different dissimilarity measures result in different ranges of values, in this method, they share the common parameter ε. In other words, $1 + \varepsilon$ can be understood as the relative value of these measures. As such, it is possible to compare these measures based on their relative values.

3.5.9 Kernel ridge regression

Algorithm

Kernel ridge regression (KRR) is the dual form of the ridge regression solution (see the detail in Appendix A). KRR aims to improve the performance of linear methods by mapping instances from the original space (Hilbert space) to a higher-dimensional space to obtain linearly separable patterns. Let ϕ be the mapping function which transforms the data to the higher-dimensional space. In kernel method, because the computation of pairwise dot product of instances in the new space is intractable, this is approximated by kernel functions $K(x_i, x_j) \approx \langle \phi(x_i), \phi(x_j) \rangle$, which form kernel matrices $\boldsymbol{K}$ (i.e., Gram matrices).

KRR indeed makes a local approximation of the target function, in which kernel functions play the role of similarity measures. Given a kernel matrix $\boldsymbol{K}$, this method aims to estimate the dual coefficients α based on this matrix. In each row i^{th} of the kernel matrix, an element in this row measures the similarity between i^{th} instance (x_i) and another one (x_j), denoted by $s(x_i, x_j)$. The dual coefficients can be understood as the weights corresponding to each $s(x_i, x_j)$ which indicates how important to take x_j for approximating the target value at x_i. Hence, there exists a subset of instances that contributes to approximating the target value at the instance i^{th} more significantly than the others. The size of these subsets affects the locality level of the kernel matrix that depends on the used kernel function which is constituted by the similarity measure and the value of γ. Hence, we investigate how similarity measures preserving the distinction of objects in comparison with the third one affects the locality of kernel matrix.

Kernel function

Kernel functions play an essential role in KRR. The radial basis function (RBF) kernel and Laplacian kernel, which are constituted from the 2-norm and 1-norm distances, respectively, have been widely used in many applications. The formulas of these two kernels are as follows:

- *RBF kernel (K_{rbf})*: $K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|_2^2)$ where $\|x_i - x_j\|_2$ is the 2-norm distance between x_i and x_j, and γ is a predefined scalar.
- *Laplacian kernel (K_{lap})*: $K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|_1)$, where $\|x_i - x_j\|_1$ is the 1-norm distance between x_i and x_j.

In this work, we consider other kernel functions which are constructed from the 3-norm distance, cosine distance, B-C dissimilarity, Canberra distance, and Chebyshev distance, as follows:

- *3-norm-based kernel (K_{min3}):* $K(x_i, x_j) = exp(-\gamma||x_i - x_j||_3)$ where $||x_i - x_j||_3$ is the 3-norm distance between x_i and x_j.
- *cosine-based kernel (K_{cos}):* $K(x_i, x_j) = exp(-\gamma \times d_{cos}(x_i, x_j))$ where $d_{cos}(x_i, x_j)$ is the cosine distance between x_i and x_j.
- *B-C-based kernel (K_{bray}):* $K(x_i, x_j) = exp(-\gamma \times d_{bray}(x_i, x_j))$ where $d_{bray}(x_i, x_j)$ is the Bray-Curtis dissimilarity between x_i and x_j.
- *Canberra-based kernel (K_{can}):* $K(x_i, x_j) = exp(-\gamma \times d_{can}(x_i, x_j))$ where $d_{can}(x_i, x_j)$ is the Canberra distance between x_i and x_j.
- *Chebyshev-based kernel (K_{che}):* $K(x_i, x_j) = exp(-\gamma \times d_{che}(x_i, x_j))$ where $d_{che}(x_i, x_j)$ is the Chebyshev distance between x_i and x_j.

Model complexity

As mentioned above, inappropriately using representations (with the Euclidean distance) induces the roughness of the target surface, hence, to approximate this surface, we need high-complexity model. In fact, similarity measures used in kernel functions affect the model complexity in KRR. Hence, we investigate the relation between similarity measures and model complexity.

Model complexity is an important concept in model selection. In addition, it has the association with the nature of data. For example, if the real target function (e.g., the formation energy function) is rough with many local extreme points, the learning model should be a non-linear function to avoid underfitting. Of course, the non-linear model is more flexible and complex than simple linear ones. However, the high complexity of the learning model can cause overfitting.

There is no strict definition of model complexity. Simply, the model complexity will be related to the number of free parameters that the model requires for better fitting. A complex model often requires more free parameters than a simple one. However, this definition does not imply a one-to-one relationship between model complexity and the number of parameters because parameters are not necessary to be equally important. For example, in linear models, there is a subset of dimensions which are more important than the rest. The complexity results in the flexibility of model, hence, two terms complexity and flexibility can be exchanged.

The complexity of the model can be quantitatively interpreted by the degrees of freedom. The degrees of freedom are denoted by *df* and defined as the number of freely varying parameters in the model (or function). In terms of model complexity, the greater the number of free parameters is, the more complex the model is. For computation, the degrees of freedom are defined as the trace of the first derivatives of $\hat{\boldsymbol{y}}$ according to $\boldsymbol{y}$ as follows:

$$df = tr\left(\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}}\right), \tag{3.20}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are the real target value and the estimated target value, respectively [55].

In KRR, because $\hat{\beta} = \boldsymbol{X}^T(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$ (see Equation A.7 in Appendix A), we have:

$$\begin{aligned} \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} &= \boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}, \\ &= \boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}. \end{aligned} \tag{3.21}$$

Therefore, the model degrees of freedom in KRR, $df(\lambda)$, are estimated as $tr\Big(\boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\Big)$. In fact, the locality of kernel matrix affects the model complexity that the increase of locality makes the increase of model complexity.

3.6 Experiments and discussion

In this section, firstly we introduce several well-known material datasets used in our work. This work focuses on predicting formation energy. Next, we demonstrate our investigations on the roughness of the energy surface over materials subject to material representations. We derive features as mentioned in the proposed protocol for validating the proposed hypothesis on the nature of dissimilarity measures. By using these features, we show the correlation among the roughness of the energy surface towards a representation, the need of preserving the instances distinction in their context-based comparison, and the high prediction accuracy. However, if the distinction of instances in context-based comparison is preserved too much, it induces the lower performance of similarity-based learning methods. Hence, we need to adjust $DLoss$ to find appropriate dissimilarity measures.

3.6.1 Material dataset

Open Quantum Materials Database

Open Quantum Materials Database (OQMD) [53, 89] is a well-known database of thermodynamic and structural properties of crystals that are calculated by using the Density Functional Theory (DFT). In this study, we use a sample of 5967 crystals extracted from this database that are magnetic materials based on rare earth-transition metal alloys. The structures of crystals containing rare-earth and transition elements are almost diverse that induces a diverse range of electronic properties on account of interval magnetic freedom [42, 67]. This dataset provides the information of chemical structure and lattice of crystals.

QM7 dataset

The QM7[2] is one of well-known dataset that is widely used for molecular machine learning [14, 88]. It consists of totally 7165 stable organic molecules which contain heavy atoms (e.g., C, N, O, S). This dataset provides the information of SMILES representation of molecules and 3D coordinates of each atom in these molecules.

[2]http://quantum-machine.org/datasets/

Table 3.1: Estimation of the $SensitivityToChange(\mathcal{D}), r$

	OQMD			QM7		
	OFM	CM	SOAP	OFM	CM	SOAP
$SensitivityToChange(\mathcal{D}, r)$	70.7%	75.9%	73.3%	70.7%	79.5%	69.9%

Table 3.2: Performance of ridge regression with the material datasets and representations of interest

Data	Desc	MAE	RMSE	R^2
OQMD	OFM	**0.188±0.007**	**0.06±0.01**	**0.968±0.004**
	CM	0.706±0.027	0.814±0.07	0.565±0.039
	SOAP	**0.201±0.012**	**0.094±0.026**	**0.95±0.015**
QM7	OFM	**8.546±0.429**	**134.824±22.973**	**0.997±0.001**
	CM	20.447±0.899	723.924±125.521	0.985±0.002
	SOAP	80.708±4.623	14168.311±2876.64	0.714±0.048

3.6.2 Evaluating the roughness of the energy surface subject to material representations

Sensitivity to the change of target values towards the change of instances

We estimate the value of $SensitivityToChange(\mathcal{D}, r)$ where $\mathcal{D}$ is the dataset of interest (OQMD, QM7) and r is the material representation of interest (OFM, CM, SOAP). The results are shown in Table 3.1.

Table 3.1 shows that most material datasets with representations of interest have large value of $SensitivityToChange(\mathcal{D}, r)$ (almost greater or equal than 70%). Hence, the energy surface are rough towards most material representations of interest.

Likelihood of globally and linearly approximating the target surface

Besides using the $SensitivityToChange(\mathcal{D}, r)$, we evaluate the fluctuation amplitude of the energy surface towards the hyperplane determined by approximating this surface by ridge regression. The likelihood of globally and linearly approximating the energy surface can be used for evaluating the roughness of this surface. This likelihood is assessed based on the prediction accuracy when performing ridge regression.

We perform the ridge regression with OQMD and QM7 datasets with the OFM, CM, and SOAP representations. The prediction accuracy is indicated through three evaluation metrics: the root mean squared error (RMSE), the mean absolute error (MAE), and the coefficient of determination (R^2). For an equitable assessment of the model and void the overfitting of model, we employ an ensemble approach for model assessment by repeating the cross-validation ten times with randomly generated samples. The most likely parameter in ridge model λ is chosen by performing grid search. The prediction accuracies obtained when performing ridge regression with the datasets and representations of interest are shown in Table 3.2.

The measure R^2 is not only useful for evaluating the linear fitting but also used as a common criterion for comparing the performance of ridge regression with different datasets

and representations. To qualitatively evaluate the fluctuation amplitude of materials formation energies towards the hyperplane learned by ridge regression, we classify material datasets with representations into three groups based on their corresponding R^2, as follows:

- Extremely high R^2: including the QM7-OFM and QM7-CM with R^2 of 0.997 and 0.985, respectively. The QM7-OFM is more likely to be linearly approximated than the QM7-CM because it results in much smaller MAE and RMSE than the QM7-CM.
- High R^2: including that OQMD-OFM and OQMD-SOAP with R^2 of 0.968 and 0.95, respectively.
- Low R^2: including the OQMD-CM and QM7-SOAP with R^2 of 0.565 and 0.714, respectively.

3.6.3 Estimating *DLoss*

As mentioned in Subsection 3.5.8, the measure *DLoss* is used for indicating how dissimilarity measures preserve the distinction of instances when comparing them based on a context (the query instance). We estimate the values of *DLoss* corresponding to each dissimilarity measure of interest with various values of ε, and then take an average over these values as shown in Table 3.3.

3.6.4 K-nearest neighbors performance

We examine the appropriateness of each dissimilarity measures in predicting materials' formation energies using KNN with various values k. For fairly assessing the model performance, we employ ten-times ten-fold cross-validation with random samples. For each material dataset and representation, we find the most likely dissimilarity measure which is shown in Table 3.4. In addition, we also demonstrate the dependency of the energy prediction accuracy obtained using KNN on the *DLoss* of dissimilarity measures, as shown in Figure 3.8.

Relying on Table 3.4 and Figure 3.8, we derive several assessments as follows:

- The function indicating the tendency of the dependency of KNN performance on the *DLoss* almost has an minimum with most material datasets and descriptors. As mentioned Subsection 3.4.2, we argue that exceedingly preserving the distinction of materials when comparing them using a referenced material can lead to overfitting, and result in poor predictive performance. In addition, the significant loss of this distinction also results in low performance where the target surface is rough. The tendency shows the high likelihood of this hypothesis.
- With pairs of datasets and representations OQMD-OFM, OQMD-CM, OQMD-SOAP, and QM7-SOAP, the fluctuation amplitude of energy values towards the hyperplane learned by ridge regression is fair large because of low or high R^2. For these datasets and representations, the tendency function has an extreme point at

Table 3.3: Estimation of *DLoss* corresponding to each dissimilarity measure of interest

Data & Desc	Dissimilarity measure	*DLoss*
OQMD-OFM	1-norm	2.01
	2-norm	2.46
	3-norm	2.92
	cosine	1.66
	B-C	2.06
	Canberra	1.84
	Chebyshev	4.35
OQMD-CM	1-norm	2.28
	2-norm	2.47
	3-norm	2.52
	cosine	1.38
	B-C	1.72
	Canberra	2.08
	Chebyshev	3.08
OQMD-SOAP	1-norm	3.3
	2-norm	4.02
	3-norm	4.76
	cosine	3.06
	B-C	3.3
	Canberra	1.74
	Chebyshev	5.44
QM7-OFM	1-norm	1.95
	2-norm	2.08
	3-norm	2.19
	cosine	1.49
	B-C	1.96
	Canberra	2.11
	Chebyshev	2.52
QM7-CM	1-norm	3.03
	2-norm	3.27
	3-norm	3.41
	cosine	1.93
	B-C	3.02
	Canberra	4.84
	Chebyshev	3.97
QM7-SOAP	1-norm	1.63
	2-norm	1.69
	3-norm	1.73
	cosine	1.36
	B-C	1.63
	Canberra	1.71
	Chebyshev	1.77

Table 3.4: The most likely dissimilarity measure for effectively performing KNN with material datasets and representations

Data	Descriptor	Dissimilarity measure
OQMD	OFM	1-norm, B-C
	CM	Canberra
	SOAP	B-C
QM7	OFM	2-norm
	CM	Canberra
	SOAP	1-norm, B-C

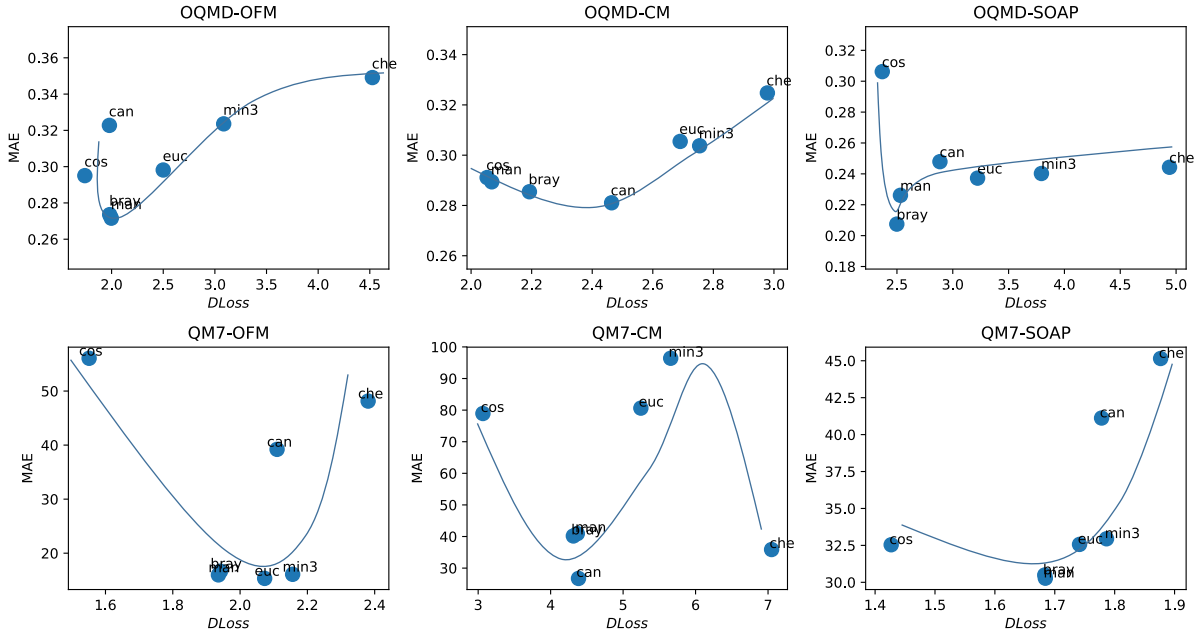

Figure 3.8: The dependency of KNN performance (MAE) on the *DLoss* of the 1-norm (man), 2-norm (euc), 3-norm (min3), cosine (cos), B-C (bray), Canberra (can), and Chebyshev (che). The solid blue line indicates the tendency of this dependency.

which, the prediction accuracy is largest. For example, the B-C dissimilarity and 1-norm distance almost result in the highest accuracy with most datasets and representations of interest. With OQMD-OFM, the Canberra dissimilarity is the most likely.

- With the QM7-OFM and QM7-CM, the fluctuation amplitude of energy values towards the hyperplane learned by ridge regression is small (extremely high R^2). For these datasets and representations, using dissimilarity measures with larger $DLoss$ result in better performance than the others. For example, with QM7-OFM, using the 2-norm distance with larger $DLoss$ than the 1-norm distance and B-C dissimilarity result in the highest accuracy. With the QM7-CM, the Canberra distance with the largest $DLoss$ results in the highest accuracy.

Relying on these assessments, we see that it is needful to make the trade-off between the preservation of material distinction in reference-based similarity evaluation and the loss of this distinction. When the target surface is rough with high fluctuation amplitude towards the based hyperplane, we need to minimize the loss of this distinction. For the surface with too small fluctuation amplitude, the preservation of this distinction is not compulsory.

Kernel ridge regression performance

As mentioned above, KRR aims to smooth the energy surface by using kernel functions which measure the similarity between data instance. Indeed, kernel functions are usually constructed based on (dis)similarity measures. Hence, we can investigate similarity measures based on the performance of their corresponding kernel functions used in KRR. To this end, we perform the KRR for material datasets and representations of interest with various kernel functions. We determine the most likely hyperparameters of the model λ and γ by performing grid search. To examine the likelihood of each pair of these hyperparameters, we also repeat cross-validation ten times with random samples. The full prediction results are shown in Table 3.5. In addition, we show the most likely kernel function when performing KRR with each dataset and representation in Table 3.6.

Similarly to KNN investigation, we examine the dependency of KRR performance on the $DLoss$ of each similarity measure used in each kernel function. The tendency of this dependency is shown in Figure 3.9. We also see that the function indicating the tendency with most material datasets and representations (excluding QM7-CM) has an unique minimum. This again shows the high likelihood of the hypothesis mentioned in Subsection 3.4.2.

As argued above, the use of inappropriate material representations (towards the Euclidean distance) induces the roughness of the energy surface over materials. Therefore, high-complexity models should be used for fitting this surface. To confirm this hypothesis, we show the dependency of KRR accuracy on the model complexity that is quantified by the degrees of freedom $df(\lambda)$ in Figure 3.10.

Figure 3.10 shows that for the OQMD-OFM, OQMD-CM, OQMD-SOAP, and QM7-SOAP, the model with high $df(\lambda)$ almost results in better performance than the others. Recalling that the energy surface of these datasets with these representations have large fluctuation amplitude towards the based hyperplane. Hence, we see that for rough target

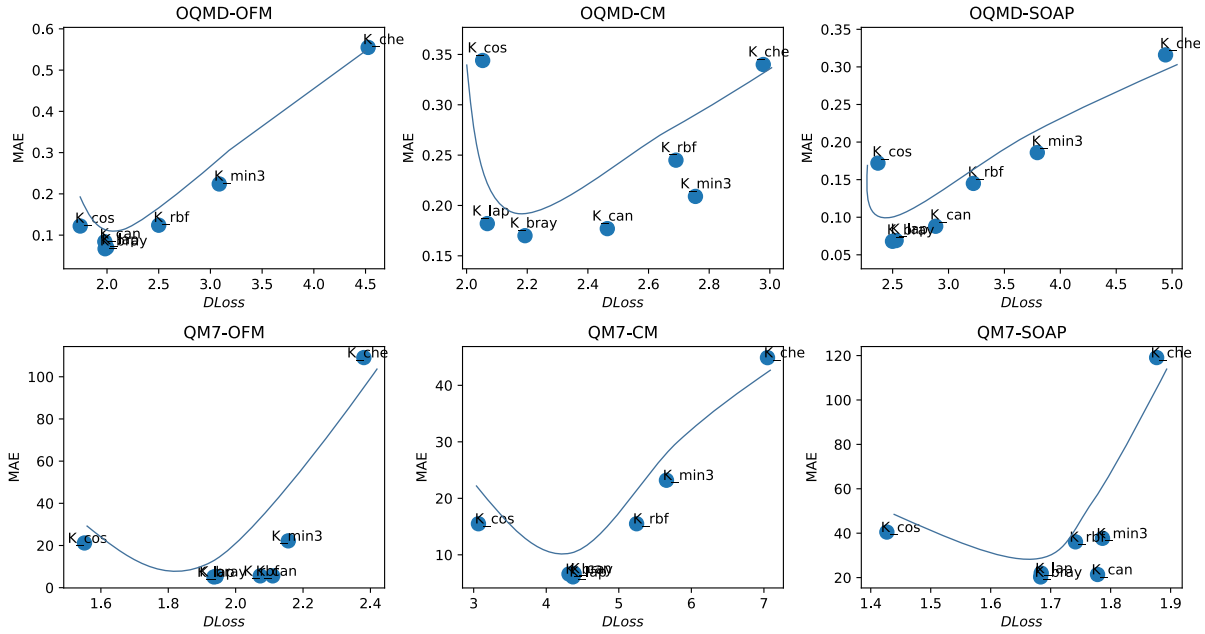

Figure 3.9: The tendency (blue line) of the dependency of KRR performance (MAE) on the $DLoss$ of similarity measures used in kernel functions, $K_{lap}, K_{rbf}, K_{min3}, K_{cos}, K_{bray}, K_{can}, K_{che}$.

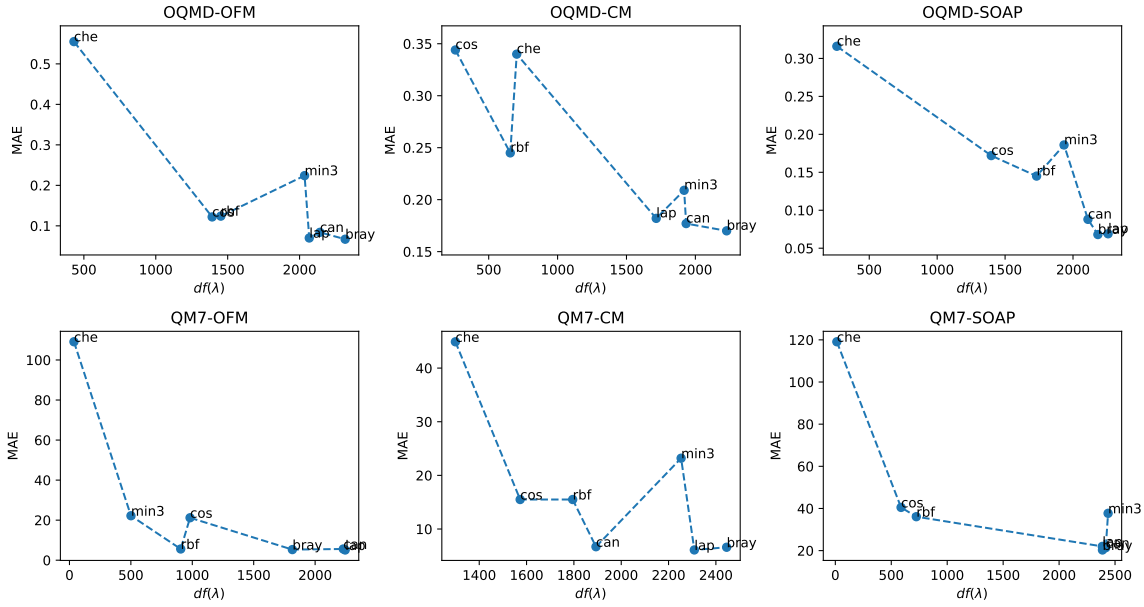

Figure 3.10: Dependency of KRR performance (MAE) on the model degrees of freedom $df(\lambda)$.

Table 3.5: Formation energy prediction performance using KRR with different kernel functions and descriptors for the OQMD and QM7 datasets, and the corresponding model complexity ($df(\lambda)$).

Data & Desc	Kernel	MAE	RMSE	R^2	$df(\lambda)$
OQMD-OFM	K_{rbf}	0.124±0.007	0.037±0.005	0.98±0.01	1452.3±16.6
	K_{lap}	**0.07±0.005**	**0.015±0.003**	**0.992±0.002**	2067.3±11.3
	K_{min3}	0.224±0.015	0.142±0.023	0.9±0.1	2034.7±22.1
	K_{cos}	0.122±0.007	0.033±0.006	0.982±0.003	1392±11
	K_{bray}	**0.067±0.006**	**0.015±0.003**	**0.992±0.002**	2317.1±20.9
	K_{can}	0.084±0.005	0.021±0.004	0.989±0.002	2141.3±13.7
	K_{che}	0.555±0.025	0.468±0.033	0.753±0.021	430.9±1.1
OQMD-CM	K_{rbf}	0.245±0.019	0.179±0.047	0.906±0.023	656.5±10.4
	K_{lap}	**0.182±0.014**	**0.099±0.018**	**0.945±0.012**	1715.9±13.1
	K_{min3}	0.209±0.018	0.142±0.031	0.922±0.018	1918.5±16.2
	K_{cos}	0.344±0.017	0.247±0.039	0.862±0.025	258.2±6.4
	K_{bray}	**0.17±0.02**	**0.106±0.024**	**0.944±0.013**	2227±17
	K_{can}	**0.177±0.016**	**0.105±0.022**	**0.941±0.014**	1933.1±7.4
	K_{che}	0.34±0.02	0.253±0.041	0.868±0.021	703.5±4.3
OQMD-SOAP	K_{rbf}	0.145±0.01	0.055±0.01	0.971±0.005	1731.9±11.2
	K_{lap}	**0.069±0.007**	**0.018±0.006**	**0.99±0.01**	2258.3±17.7
	K_{min3}	0.186±0.015	0.097±0.015	0.949±0.007	1934.1±13.1
	K_{cos}	0.172±0.009	0.068±0.009	0.964±0.005	1398.1±15.9
	K_{bray}	**0.068±0.006**	**0.017±0.004**	**0.99±0.01**	2183.2±12.3
	K_{can}	0.088±0.008	0.026±0.006	0.987±0.003	2110±15
	K_{che}	0.316±0.016	0.183±0.021	0.901±0.012	261.5±0.9
QM7-OFM	K_{rbf}	**5.6±0.4**	**68.2±16.9**	**0.999±0.001**	905.8±6.4
	K_{lap}	**5.1±0.4**	**78.5±24.2**	**0.998±0.001**	2244.8±12.6
	K_{min3}	22.2±1.1	956±116	0.982±0.002	500.7±19.4
	K_{cos}	21.2±1.7	1518.5±1116.1	0.969±0.028	981.8±9.6
	K_{bray}	**5.3±0.4**	**86.9±38.4**	**0.998±0.001**	1814.8±11.1
	K_{can}	5.6±0.6	114.5±94.3	0.997±0.003	2229.3±15.5
	K_{che}	109.1±4.9	17914.8±1284.1	0.633±0.036	36.5±1.9
QM7-CM	K_{rbf}	15.5±0.9	498.2±88.5	0.99±0.01	1794.9±10.9
	K_{lap}	**6.1±0.4**	**98.5±33.2**	**0.998±0.001**	2308.9±14.1
	K_{min3}	23.2±1.8	1494.9±603.3	0.969±0.011	2253±10
	K_{cos}	15.5±0.8	513±196	0.99±0.01	1572.9±4.7
	K_{bray}	6.6±0.6	133.7±69.2	0.997±0.002	2445.4±11.5
	K_{can}	**6.7±0.4**	**108.5±37.3**	**0.998±0.001**	1893.3±10.1
	K_{che}	44.9±2.8	5062.7±1165.2	0.898±0.021	1299.3±4.9
QM7-SOAP	K_{rbf}	36.1±4.3	7862.5±3835.2	0.839±0.108	724±8
	K_{lap}	22±3	3928.1±1648.9	0.922±0.035	2386.2±15.1
	K_{min3}	37.7±6.2	11630.6±3713	0.758±0.078	2438±18
	K_{cos}	40.5±4	7968±2407	0.84±0.04	587.6±8
	K_{bray}	**20.3±3.1**	**3463.1±1548.5**	**0.932±0.033**	2387.7±19.9
	K_{can}	**21.4±2.7**	**3524±1923**	**0.927±0.031**	2416.8±15
	K_{che}	119.1±5.4	24764.4±3089.4	0.491±0.037	13.7±0.1

Table 3.6: The most likely kernel function for effectively performing KRR with material datasets and representations

Dataset	Representation	Kernel function
OQMD	OFM	K_{lap}, K_{bray}
	CM	K_{lap}, K_{bray}, K_{can}
	SOAP	K_{lap}, K_{bray}
QM7	OFM	K_{rbf}, K_{lap}, K_{bray}
	CM	K_{lap}, K_{can}
	SOAP	K_{bray}, K_{can}

Table 3.7: Combining derived features from empirical experiments for validating the effectiveness of the proposed criterion

Data	Desc	Surface roughness		$DLoss^*$	Model complexity*
		$SensitivityToChange(\mathcal{D}, r)$	R^2		
OQMD	OFM	large	high	small	high
	CM	large	low	small	high
	SOAP	large	high	small	high
QM7	OFM	large	extremely high	large	low or high
	CM	large	extremely high	large	low or high
	SOAP	large	low	small	high

surface, using kernel functions which induce the high model complexity is the most likely for fitting. In addition, these kernel functions are often constructed from dissimilarity measures with small $DLoss$.

For the QM7-OFM and QM7-CM, the energy surface have too small fluctuation amplitude towards the based hyperplane, hence, using kernel functions which induce the small model complexity can provide the better performance. For example, for QM7-OFM, RBF kernel is the most likely, and for QM7-CM, using the K_{can} that makes low-complexity model provides the highest accuracy.

3.6.5 Combining derived features and make induction rule for effectively using dissimilarity measures for material datasets

A series of our investigations based on the proposed protocol (as presented above) indeed derives qualitative features indicating the roughness of the energy surface, the nature of dissimilarity measures as mentioned in the proposed criterion, and the prediction accuracy. We combine these features as shown in Table 3.7. Relying on this table, we draw the tree indicating the rule for appropriately using dissimilarity measures for fitting the energy surface, as shown in Figure 3.11.

The combination of derived features show the correlation among the roughness of target surface given a representation, the need of preserving objects distinction in context-based comparison when using dissimilarity measures, and the high prediction accuracy. In other words, it shows the high potential of our proposed hypothesis on the nature of dissimilarity measures and the importance of considering this nature in similarity-based

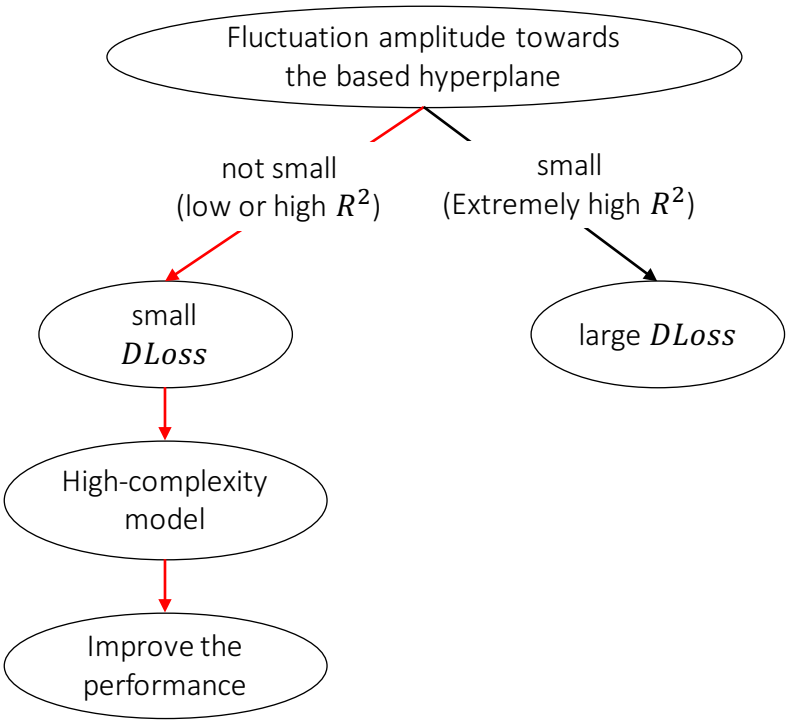

Figure 3.11: Derived rules for appropriately using dissimilarity measures for predicting formation energies.

learning.

3.7 Learning the distance between materials

To deal with the problem that measuring the similarity based on data representation is not compatible with the real similarity of target values, we use the distance metric learning (DML) to learn a suitable Mahalanobis distance between instances by taking the target variable into account. By using this measure, close instances (according to this similarity measure) in the representation space will have the similar target values, and vice versa. Learning the Manhalanobis distance is the original space is equivalent to learning a linear transformation of original instances. In other words, we expect that the Euclidean distance between two transformed instances will be consistent with the difference/similarity of their target values.

The DML learns the distance that fits with the relation between instances in a specific dataset, and aims to improve the performance of similarity-based learning models for this dataset. Therefore, we can use DML to confirm whether finding an appropriate similarity measure between materials for effectively predicting their formation energies will result in the high complexity of learning model.

3.7.1 Introduction to distance metric learning

DML is a branch is machine learning that aims to explore an appropriate distance between instances in a given dataset. It is widely used for improving the performance of similarity-based methods and dimensionality reduction [24]. The motivation of DML is that the distance measured instances is not compatible with the difference of their target values, so it attempts to enhance the consistency between them. The idea of DML originates from the definition of the Mahalanobis distance that is defined as the following:

$$d_M(x, y) = \sqrt{(x - y)^T M (x - y)} \tag{3.22}$$

where M is a positive definite (or semidefinite) matrix. Note that when $M = I$, the distance $d_M(x, y)$ is the Euclidean. Because M is symmetric positive semidefinite matrix, it can be decomposed into its Cholesky factors as $M = L^T L$. Hence, alternatively, we can re-write the Equation 3.22 as $d_L(x, y) = ||L(x - y)||_2$. Given an objective, DML can learn an appropriate distance between instances for fitting with this objective by adjusting the matrix L.

There are many methods for learning the Mahalanobis distance between instances in a given dataset that depend on a specific purpose. The original method to learn the Mahalanobis distances proposed by Xing *et al.*, which is called probabilistic global distance metric learning [101]. For improving the performance of k-nearest neighbors classification, Goldberger *et al.* proposed the neighborhood component analysis (NCA) that attempts to optimize the leave-one-out error on the training set [35]. In addition, Wang *et al.* proposed the method called average neighborhood margin maximization (ANMM) [96], and Weinberger *et al.* proposed the method called large margin nearest neighbors (LMNN) [98]. Recently, Ying *et al.* formulated the DML as an eigenvalue optimization problem [103].

In this study, we investigate two well-known DML methods: neighborhood component analysis (NCA); and large margin nearest neighbors (LMNN). These methods are used for improving the performance of k-nearest neighbors.

3.7.2 Neighborhood component analysis (NCA)

NCA [35] is a distance metric learning algorithm which aims to minimize the leave-one-out error expected by the nearest-neighbors classification. We consider the training set $X = \{x_1, x_2, .., x_N\} \subset \mathbb{R}^d$ with labels $y_1, ..., y_n$. The goal of NCA is to determine a linear transformation L to optimize the leave-one-out error. However, the leave-one-out error is non-smooth function subject to the linear transformation L, so it is difficult to minimize this function directly. Alternatively, NCA solve this problem in a stochastic way. Given two instances $x_i, x_j \in X$, the probability that x_i has x_j as its nearest neighbor is defined via their distance as the following:

$$\begin{aligned} p_{ij}^L &= \frac{exp(-||Lx_i - Lx_j||^2)}{\sum_{k \neq j} exp(-||Lx_i - Lx_k||^2)} \quad (j \neq i), \\ p_{ii}^L &= 0 \end{aligned} \tag{3.23}$$

The probability that x_i is correctly classified is defined as the sum of probabilities that x_i and its nearest neighbors have the same label, that is:

$$p_i^L = \sum_{j \in C_i} p_{ij}^L, \text{ where } C_i = \{j \in \{1, ..., N\} : y_j = y_i\}. \tag{3.24}$$

We define the expected number of correctly classified instances, and try to maximize the this function as the following:

$$f(L) = \sum_{i=1}^{N} p_i^L = \sum_{i=1}^{N} \sum_{j \in C_i} p_{ij}^L = \sum_{i=1}^{N} \sum_{j \in C_i} \frac{exp(-||Lx_i - Lx_j||^2)}{\sum_{k \neq i} exp(-||Lx_i - Lx_k||^2)} \tag{3.25}$$

This function is differentiable, and we can compute its first derivative as the following:

$$\bigtriangledown f(L) = 2L \sum_{i=1}^{N} \Big(p_i^L \sum_{k=1}^{N} p_{ik}^L O_{ik} - \sum_{j \in C_i} p_{ij}^L O_{ij} \Big), \tag{3.26}$$

where $O_{ij} = (x_i - x_j)(x_i - x_j)^T$. Since the gradient is known, we can optimize the objective function using a gradient descent method.

3.7.3 Large margin nearest neighbors (LMNN)

Large margin nearest neighbors (LMNN) [98] is a metric learning algorithm that learns a Mahalanobis distance metric for improving the accuracy of k-nearest neighbor classification. As mentioned above, learning the Mahalanobis distance is equivalent to learning a projection matrix L ($M = L^T L$) of the original data. To improve the performance of KNN, k nearest neighbors are expected to share the same label, thus, LMNN targets to learn a distance between instances to maximize the number of instances which share its label with as many neighbors as possible.

Let $X = \{x_1, x_2, ..., x_N\} \subset \mathbb{R}^d$ be a dataset with corresponding labels $y_1, y_2, ..., y_N$. Given a instance $x_i \in \mathcal{X}$, considering k nearest neighbors of x_i (N_k), if a neighbor $x_j \in N_k$ has the same label with x_i, it is called a target neighbor of x_i, denoted by $j \rightsquigarrow i$. In such neighbors, a neighbor x_l is called an impostor of x_i and x_j if it has different label from x_i and x_j ($y_l \neq y_i = y_j$), and satisfies the constraint $||x_i - x_l||^2 \leq ||x_i - x_j||^2 + 1$ where $j \rightsquigarrow i$. In fact, each sample x_i has a margin for its neighbors which are expected to have the sample label with x_i, but impostors invade this margin. The target neighbors and impostors of a given instance are illustrated in Figure 3.12.

There are two optimization goals in LMNN. The first goal is to minimize the Mahalanobis distance between an instance of interest and its target neighbors as the following:

$$\begin{aligned} \varepsilon_{pull}(M) &= \sum_{i, j \rightsquigarrow i} \mathcal{D}_M(x_i, x_j) \\ \Leftrightarrow \varepsilon_{pull}(L) &= \sum_{i, j \rightsquigarrow i} ||L(x_i - x_j)|| \end{aligned} \tag{3.27}$$

where, $M = L^T L$, and $\mathcal{D}_M(x_i, x_j)$ denotes the squared distance with respect to the Mahalanobis metric M. The second goal is to penalize small distances between differently

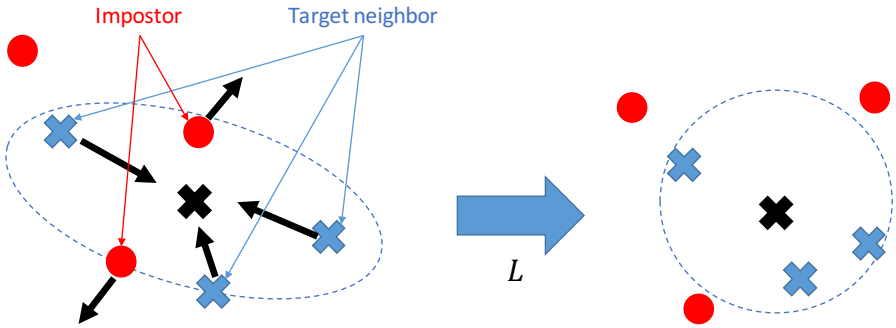

Figure 3.12: Graphical description of target neighbors and imposters in the large margin nearest neighbors algorithm.

labeled instances. That means increasing the distance between the instance of interest and its impostors. The second objective is formulated as follows:

$$\begin{aligned} \varepsilon_{push}(M) &= \sum_{i,j \rightsquigarrow i} \sum_{l} (1 - y_{il}) \Big[1 + \mathcal{D}_M(x_i, x_j) - \mathcal{D}_M(x_i, x_l) \Big]_+ \\ \Leftrightarrow \varepsilon_{push}(L) &= \sum_{i,j \rightsquigarrow i} \sum_{l} (1 - y_{il}) \Big[1 + ||L(x_i - x_j)||^2 - ||L(x_i - x_l)||^2 \Big]_+ \end{aligned} \tag{3.28}$$

where the term $[z]_+ = \max(z, 0)$, $y_{il} = 1$ if and only if $y_i = y_l$, and $y_{il} = 0$ otherwise. The objective function of LMNN is the combination of two goals, as follows:

$$\begin{aligned} \varepsilon(M) &= (1 - \mu)\varepsilon_{pull}(M) + \mu\varepsilon_{push}(M) \\ &= (1 - \mu) \sum_{i,j \rightsquigarrow i} \mathcal{D}_M(x_i, x_j) + \mu \sum_{i,j \rightsquigarrow i} \sum_{l} (1 - y_{il}) \Big[1 + \mathcal{D}_M(x_i, x_j) - \mathcal{D}_M(x_i, x_l) \Big]_+ \end{aligned} \tag{3.29}$$

where $\mu \in [0, 1]$ is the weight parameter. For minimizing $\varepsilon(M)$, we can solve by using a semidefinite program (SDP) that is a linear programming incorporating an additional constraint on a symmetric positive semidefinite matrix. To this end, the optimization in Equation 3.29 needs to be formulated as a standard form of SDP by additionally using slack variables $\{\xi_{ijl}\}$ for all triplets of target neighbors $(j \rightsquigarrow i)$ and impostors. This minimizes the following function:

$$(1 - \mu) \sum_{i,j \rightsquigarrow i} (x_i - x_j)^T M (x_i - x_j) + \mu \sum_{i,j \rightsquigarrow i,l} (1 - y_{il}) \xi_{ijl}, \tag{3.30}$$

subject to:

1. $(x_i - x_l)^T M (x_i - x_l) - (x_i - x_j)^T M (x_i - x_j) \geq 1 - \xi_{ijl}$
2. $\xi_{ijl} \geq 0$
3. $M \succeq 0$

Table 3.8: Formation energy prediction performance by using KNN for the original and LMNN-based and NCA-based transformed data.

Dataset & descriptor		X	X_{LMNN}	X_{NCA}
OQMD-OFM	MAE	0.286±0.003	**0.277±0.003**	**0.266±0.003**
	R^2	0.898±0.003	**0.903±0.003**	**0.911±0.003**
OQMD-CM	MAE	0.265±0.004	**0.261±0.004**	0.277±0.004
	R^2	0.904±0.004	**0.907±0.004**	0.896±0.004
AB Compound	MAE	0.325±0.026	**0.283±0.018**	**0.278±0.028**
	R^2	0.81±0.03	**0.835±0.021**	**0.832±0.043**

3.7.4 Model complexity investigation with the learned distance

Notation

As mentioned above, learning a Mahalanobis distance for instances in a dataset is equivalent to learning a projection matrix L of these instances. Hence, we perform KNN and KRR for transformed data then compare the performance with the original one. Let X be a data sample. The instances in this sample are transformed to a new space by using LMNN and NCA, and then we denote transformed sample by X_{LMNN} and X_{NCA}, respectively.

Experiments

We use the OQMD dataset with the OFM and CM descriptors, which are denoted by OQMD-OFM and OQMD-CM, respectively. In addition, we take the AB compound dataset [48] into account that consists of 239 binary AB materials (A elements are metallic atoms and B elements are metalloids and non-metallic atoms) described by 17 features.

Because the LMNN and NCA work with discrete target values (labels), we divide formation energies (continuous variable) into bins, and then assign a label for each bin. For OQMD dataset, we divide the formation energies into 500 bins. For the AB compound dataset, we divide energies into 20 bins. We project X on a new space with preserving the original number of dimensions.

To evaluate the appropriateness of learned Mahalanobis distance, we perform KNN (with using the Euclidean distance) for the transformed data, and then evaluate based on the accuracy. The accuracies when using KNN for X, X_{LMNN}, and X_{NCA} are shown in Table 3.8. This table shows that in general, the LMNN and NCA make the improvement in predicting formation energy by using KNN.

Next, we perform the KRR using RBF kernel for transformed data X_{LMNN} and X_{NCA}, and then also compare their prediction accuracy with that of the original data. The accuracies obtained when performing KRR for the original data and transformed data are shown in Table 3.9. In addition, we measure the model complexity of KRR with these samples by estimating model degrees of freedom (denoted by $df(\lambda)$). The table shows that with the Mahalanobis distance learned by using LMNN, we obtain the higher prediction accuracy. Meanwhile, we obtain a lower accuracy by using NCA.

Besides showing the improvement of prediction accuracy when using KRR with LMNN, Table 3.9 also shows that LMNN implies the increase of model complexity with all datasets

Table 3.9: Formation energy prediction performance using KRR for the original and LMNN-based and NCA-based transformed data.

Dataset & descriptor		X	X_{LMNN}	X_{NCA}
OQMD-OFM	MAE	0.113±0.001	**0.109±0.001**	1.077±0.01
	R^2	0.987±0.001	**0.987±0.001**	-0.075±0.019
	$df(\lambda)$	2132.017	2679.793	5953.459
OQMD-CM	MAE	0.245±0.006	**0.212±0.005**	0.769±0.005
	R^2	0.916±0.008	**0.93±0.01**	0.278±0.007
	$df(\lambda)$	711.061	1041.639	4501.203
AB Compound	MAE	0.179±0.017	**0.156±0.011**	0.256±0.017
	R^2	0.938±0.013	**0.945±0.007**	0.869±0.024
	$df(\lambda)$	128.059	234.003	227.273

of interest. Obviously, the learned distance, used for constructing the kernel matrix, induces the increase of model complexity and is appropriate for approximating the energy surface of materials. It is consistent with our investigations as presented in previous sections.

Extremely high complexity of model may cause overfitting that results in a poor prediction accuracy. The experiment shows that the use of NCA makes the model complexity exceedingly high, so KRR poorly performs with X_{NCA}.

3.8 Chapter summary

We address an important problem in machine learning that the similarity measure based on objects' representation is not consistent (compatible) with the similarity of their target values. It makes the roughness of target function. In case the representation is not ideal, selecting appropriate similarity measures can help to improve the performance in predicting target values. To effectively fitting rough target function, we hypothesize that similarity measures need to preserve the distinction of two objects in comparison with the third one. In other words, the distinctiveness of pairwise comparison in triplet of objects needs to be preserved. We demonstrate and verify this hypothesis by carrying out a study on measuring similarity between materials serving the formation energy prediction. We employ a protocol that incorporates various methods for investigating the roughness of target function and similarity measures, and the distance metric learning to verify this hypothesis. In addition, various material dataset and descriptors are also taken into account for verifying the generalization of the proposed hypothesis. The experimental results indicates high likelihood of our proposed hypothesis. Relying on this, we establish general principles for effectively using similarity measures for mining material data, which do not depend on a specific learning model.

Chapter 4

Reference Diversification in Analogy-based Causality Inference

4.1 Introduction

In this chapter, we present our work on using analogy-based approach for causality inference through a study on assessing the cause of adverse drug reactions (ADRs) by using the analogy criterion - one of nine criteria proposed by Bradford Hill for causality assessment in medicine. This criterion states that similar drugs can cause similar ADRs, hence, if we observe an ADR that has the association with two similar drugs, it is likely to be caused by these drugs. In this context, we use similar drugs as a reference for confirming the causal relation between the drug of interest with an ADR. By using the similarity-based causality inference, we have chance to recognize confounding factors that are caused by polypharmacy. Confounding factor here are non-causal drug-ADR pairs, which are coincidentally and frequently observed in the treatment period.

In this work, we present a drug by two main attributes: (i) its mechanism of actions and targets; and (ii) list of associated ADRs that is resulted from the use of this drug in practical treatment. We model the analogy criterion as the voting process of similar drugs for the existence of causal relationship between a pair of drug-ADR of interest. The conflict in voting, which is exploited for eliminating non-causal pairs, is resulted from the the difference of similar drugs according to the second attribute. Hence, we propose methods for selecting groups of similar drugs that maximize the diversity of these groups according to the second attribute.

4.2 Overview of pragmatic clinical trials

Clinical trials that are often carried out before a drug is approved to be marketed cannot cover all possible responses to this drug because these trials are conducted on non-representative patient cohorts and under ideal and controlled conditions. In practical treatment, the drug-drug interaction and overdose can lead to unpredictable effects. Therefore, to enrich knowledge about these effects as well as to manage the drug risks, the need of a long-term administration on the entire population is inevitable. This administration is called post-marketing pharmaceutical surveillance, a.k.a. pragmatic clinical

trials (PCT) [16, 31, 43, 113].

PCT has been promoting a new paradigm shift in epidemiology and drug safety [75]. The main purpose of PCT is to relatively evaluate drug effectiveness in real-world treatment where the diversity and evolution of circumstances may lead to an occurrence of unintended adverse drug reactions (ADRs). The risk of drugs is evaluated by determining the association between these drugs and observed ADRs during the treatment period.

Most studies on PCT have been based on textual data which includes clinical notes and patients' reports. Several spontaneous reporting systems (SRS) were early established to collect information about ADRs from patients, physicians, and pharmacists. These systems help to accelerate the process of detecting unknown ADRs with an effective cost [23, 40, 46]. However, spontaneous reports, which are almost collected from patients, do not fully meet requirements for effectively assessing drug-ADR causality since ADRs described by patients are bias and incomplete.

Recently, the use of electronic medical records (EMRs) has been encouraged, which is expected to overcome limitations of SRS. This data source provides objective descriptions about patient treatment progress that helps to improve the quality of PCT [23, 94]. In addition, EMRs can help to discover unknown ADRs. However, because this data describes real-world treatments with the co-occurrence of multiple events, it can makes confusions when assessing the causal relationship between drugs and ADRs. This raises a big challenge in PCT when using EMRs.

The key problem in PCT is to assess the causality ADRs based on reasonable evidences [32]. This helps to estimate the risk of drugs for early preventing the recurrence of medical failure in the future [81]. For assessment, it is needful to estimate the likelihood a drug will be responsible to an ADR [30, 60].

Essentially, the causality assessment is PCT is a phenotype-based (or observational) study. Hence, this cannot guarantee the truth of discovered causal relations between drugs and ADRs because we cannot clarify the biologically causal mechanism under these relations from textual data. In vivo and in vitro test are further required for intensively confirming drug-ADR relations found by PCT. However, PCT is needful because people have not known all ADRs and have not fully understood how ADRs are caused, so PCT helps to highlight suspicious drug-ADR pairs which are likely to have causal relationship. Drug-ADR causal relations recommended by PCT can help to reduce cost of confirmation steps.

For treatment, particularly, for elderly patients, polypharmacy is usually required for the longer life expectancy and co-morbidity [34]. In several cases, the polypharmacy can be used for reducing side effects of several medications [66, 44]. However, the unnecessary polypharmacy can result in negative consequences such adverse effects, drug-drug and drug-disease interactions [90, 38, 71, 39]. When combining multiple medications, adjusting doses of medications plays an important role. In fact, physicians may reduce doses of drugs [65, 61, 62]. However, medical errors can result in the overdose of medications which can lead to appearance of ADRs [15]. Polypharmacy makes a big challenge in detecting drugs which actually cause observed ADRs when multiple drugs are prescribed simultaneously for treating co-morbidities.

4.3 Confounding caused by polypharmacy

4.3.1 The importance of considering confounding for avoiding bias in medicine

Much of epidemiology and social science research is devoted to estimation of causal effects and testing causal hypothesis using non-experimental data. In such effort, issues of confounding will invariably arise [37]. Confounding refers to the bias in estimating causal effects which is informally described as a mixing of effects of extraneous factors (called confounders) with the effect of interest. This definition has been widely used in studies in epidemiology and sociology. Confounding factors may mask an actual association, so it causes the false estimation of the treatment-outcome association when there is no real association between them. The existence of confounding factors in observational studies makes a difficulty to establish a clear causal link between treatment and outcome. In [92], several general characteristics of confounding factors were discussed that include: (i) a confounding factor is predictive a the outcome, even in the absence of the exposure; (ii) a confounding factor has the association with the exposure being studied but is not the result of the exposure; and (iii) a confounder cannot be an intermediate between the exposure and the outcome.

In general, the concept of confounding is mathematically quantified by considering the distribution of the outcome in a specific population. Suppose that the objective here is to determine the effect of applying a treatment or exposure x_1 for the population A with the distribution of the outcome parameterized by μ_{A1}. In addition, another treatment x_0 is also applied for the population A with the outcome distribution denoted by μ_{A0}. For example, the population A could be a cohort of breast-cancer patients, treatment x_1 could be a new hormone therapy, and x_0 could be a placebo therapy. The outcome distribution μ could be the expected survival or the five-year survival probability in the cohort. The causal effect of x_0 relative to x_1 is defined as the change from μ_{A1} to μ_{A0} that could be measured by $\mu_{A0} - \mu_{A1}$ (or by $\frac{\mu_{A0}}{\mu_{A1}}$). Suppose, however, additionally considering a population B under the treatment x_0, we expect that the outcome distribution of x_0 on these populations is the same, i.e., $\mu_{A0} = \mu_{B0}$. We say confounding is present if we observe that $\mu_{A0} \neq \mu_{B0}$. The population B is called a control or reference population.

The confounding should be considered and controlled in design and implementation of study. To avoid confounding, a common approach is to use a reference population B, however, such a population may be difficult or impossible to find, so the constructions of this are demanded that is called design-based methods. Besides, confounding can be controlled by analytic adjustments which are based on observed covariate distributions in the compared populations [37]. Inspired by this, we need the reference in assessing the cause of ADRs with the aim of controlling confounding.

4.3.2 Definition of drug-ADR association

Before addressing the issue of confounding factors in drug-ADR causality assessment, we define the association between drugs and ADRs. Note that the association between drugs and ADRs does not means that they have the causal relationship. In fact, there is only a subset of these associations that have causal relation.

Table 4.1: An example of prescriptions in EMRs.

SUBJECT_ID	HADM_ID	START_DATE	END_DATE	DRUG
57139	155470	12/27/85	1/11/86	Heparin
57139	155470	12/28/85	1/11/86	Rifaximin

Let X and Y be a set of drugs prescribed for a specific patient cohort and a set of ADRs observed, respectively. For treating a given patient, a drug $x \in X$ is being prescribed from t_x^{start} (the time of starting using the drug) to t_x^{stop} (the time of stopping using the drug). An ADR, denoted by $y \in Y$, then is observed at the time t_y. The drug x is suspected to cause the ADR y, denoted by $x \rightarrow y$ (x is followed by y), if $t_y \in (t_x^{start}, t_x^{stop})$. In other words, we say that the drug x and the ADR y have an association. Suppose, we do not know the biologically causal mechanism under such an association, so we predict the causative relationship between this drug-ADR pair based on the frequency of its co-occurrence although the co-occurrence frequency does not reflect the causality. Each drug-ADR association is considered as a candidate for the causality assessment.

4.3.3 Polypharmacy-induced confounding definition

In practical treatment, patients are often admitted to the hospital with a presence of several diseases (co-morbidity) that requires the use of several drugs for curing at the same time. The use of multiple drugs can be risky because of accidents caused by the drug-drug or drug-disease interaction, and overdose. The prescriptions, treatment progress, and observed adverse events can be noted in EMRs that can help with taking a chance to find the cause of such adverse events. However, we just observe a mixture of drugs' effects which are noted in clinical notes. Tables 4.1 and 4.2 show an example of prescriptions and clinical notes which contain treatment information of a patient whose id is 57139. Clinical notes mention about ADRs observed during the treatment. Note that the date in the tables has the format of mm/dd/yy where the year are encoded for de-identification. Relying on the information in these tables, we can draw a diagram indicating the progress of treatment by using Heparin and Rifaximin with the time of observing ADRs ("abnormally deep breathing", "coughing", and "erythema") as shown in Figure 4.1.

As illustrated in Figure 4.1, a mixture of drugs' effects are observed when multiple drugs are being prescribed, which raise the problem of confounding in assessing the causal relation between drugs and effects. In fact, we have no information of single drug used in EMRs and may lack the domain knowledge to predict causal drug-ADR pairs. Hence, there is a huge space of all possible drug-ADR associations for assessment. For example, since Heparin and Rifaximin are co-prescribed, there are totally six associations between these drugs and observed ADRs (candidates) for assessment. These candidates include non-causative associations, e.g., Rifaximin and erythema because erythema is not caused by Rifaximin. The lack of strict evidences to distinguish causal drug-ADR pairs from non-causal ones, and the coincidental and frequent co-occurrence of non-associated pairs make most of associated-based methods unsuccessful in causality assessment.

Non-causal drug-ADR pairs, which frequently co-occur, are called polypharmacy-induced confounding factors [77]. Confounding factors make a distortion in measuring

Table 4.2: An example of clinical notes in EMRs. Terms indicating ADRs are italic.

SUBJECT_ID	HADM_ID	DATE	CATEGORY	TEXT
57139	155470	12/30/85	Nursing	His symptoms gradually worsened with *erythema*.
57139	155470	12/29/85	Nursing	Pt has episode of *coughing* with *abnormal deep breathing*. *Erythema* was observed.

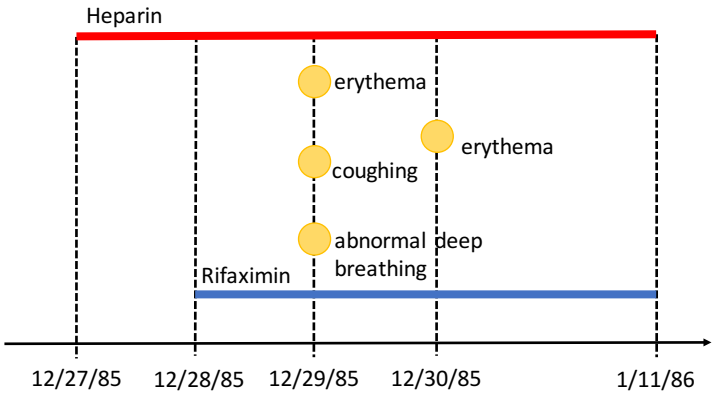

Figure 4.1: The diagram indicating the treatment progress which is extracted from Tables 4.1 and 4.2. The red and blue lines indicate the period that Heparin and Rifaximin are being prescribed.

the drug-ADR causality [68, 95].

In principle of causality perception, an action has a causal effect on a response if the response changes when the action changes while everything else remains unchanged [4]. Therefore, isolating causal mechanism is the basis to draw conclusions about the causal relationship between a presume cause and an effect. In pre-marketing clinical trials, it is feasible to isolate the interaction between a drug of interest and an ADR of interest because we can design ideal circumstances with selective and controlled conditions. Nevertheless, no information of single drug utilization is given in EMRs that makes isolating the causal mechanism in PCT infeasible. Hence, we need mechanisms to control and avoid confounding factors for effectively predicting the causal relationship.

4.4 Previous studies on ADR causality assessment

The methods for assessing causality of ADRs so far vary from the expert-judgment-based methods to the probabilistic and algorithmic methods [1, 3]. Several algorithms were early developed to assess the cause of ADRs, which measure the association between drugs and ADRs based on designing a series of criteria or questionnaires and counting yes/no answers. These algorithms were reviewed in [1, 29] in which the Naranjo algorithm has been disseminated for evaluating the drug-outcome causality [78]. Additionally, the genetic algorithm was designed as a quantitative method and utilized in several ADR assessment systems [54]. The methods following the algorithmic approach can reduce the disagreement and uncertainty in the assessment, but narrow down the searching space for possible causal relations, particularly, for searching unknown relations.

To overcome the limitation of the algorithmic approach, methods, which estimate drug-ADR associations by using the co-occurrence-based statistics from textual data, aim to enlarge the search space of drug-ADR causal relations and capture the uncertainty. So far, various measures have been developed to quantify the drug-ADR associated strength that were reviewed in [82]. Liu *et al.* measured the drug-ADR associations by using the log-likelihood ratio on drug reviews [64]. Several methods base on the contingency table such as the χ^2 test [20, 97], the reporting odd ratio (i.e., ROR, ROR_{05}) [112]. For longitudinal databases, temporal association rules can be used for representing drugs followed by ADRS in a predefined time interval. The rule strength can be calculated via several measures commonly used in association rule mining such as: confidence (i.e., conf), leverage, unexlev (in the MUTARA algorithm), and RankRatio (in the HUNT algorithm) [49, 50, 51]. In addition, Harpaz *et al.* proposed a novel measure called the relative reporting ratio (i.e., RR) [41]. Noren *et al.* developed the observed to expected ratio algorithm with the information component measure (i.e., IC) [79]. There are two main drawbacks of methods following this approach such that: (i) non-associated drug-ADR pairs, which are coincidentally and frequently observed, will be wrongly identified because of the polypharmacy-induced confounding; and (ii) these methods cannot retrieve infrequently observed drug-ADR associations.

Besides using the temporal association rule, Bayesian network was also utilized to produce probabilities which indicate the likelihood that a drug-ADR pair is associated for assessing the drug-ADR causal relationship [11, 87]. In this method, the network structure needs to be declared beforehand by experts. Zitnik *et al.* utilized the graph convolutional

networks for modeling polypharmacy side effects based on constructing a multimodal graph of protein-protein, drug-protein, and drug-drug interactions [111]. Bansal *et al.* addressed computational challenges in predicting the activity of pairs of compounds [8]. For the drug-disease association prediction, Zhang *et al.* proposed a similarity constrained matrix factorization method that uses known drug-disease associations, drug features and disease semantic information [107]. In addition, multiple kernel learning was used for identifying the drug-side-effect association [28]. For improving efficacy and reducing side effects, Huang *et al.* developed a novel method for predicting combinations of drugs (i.e., drug co-prescription) [44].

Bradford Hill criteria have been widely used in many areas such as epidemiology, genetics, molecular biology, and toxicology. Several attributes based on Bradford Hill criteria were investigated for predicting drug-ADR causal relations, which showed that the temporality and specificity are useful for causal inference [83]. However, most of associations found in longitudinal observational databases are non-causal because of confounding that was addressed in [84]. Therefore, the consistency evaluation is needful for improving the classification performance. A feature set imitating the strength, specificity, temporality, biological gradient, and experimentation was constructed to enable applying supervised learning methods to detect ADRs [85].

Most existing work is poor to deal with polypharmacy-induced confounding, and can leave out causal drug-ADR pairs that are infrequently observed. Hence, this motivates our study on drug-ADR causality assessment with focusing on reducing the bad effects of polypharmacy-induced confounding on the assessment performance.

4.5 Objectives and ideas

4.5.1 Objectives

Most previous studies assess the causality between a drug and an ADR based on their co-occurrence, so they are poor to deal with the polypharmacy-induced confounding factors. Hence, it motivates our work. As the process of unsupervised causal relation recognition, associations of drugs and ADRs are ranked according to the likelihood that they have causal relation. Therefore, appropriately measuring the likelihood that the drug and ADR have causal relation plays an essential role. To reduce the bad effect of confounding factors on the performance of causal relation detection, we target to find an appropriate measure that can distinguish causative drug-ADR pairs from non-causative pairs.

4.5.2 Ideas

Inspired by the use of reference population for controlling confounding, we use similar drugs as the reference for assessing the causal relation between the drug of interest with ADRs based on the analogy criterion. This criterion states that similar drugs may cause similar ADRs. Hence, the drug is believed more to cause the ADR if we find other drugs which are similar to the drug of interest and also have association with the ADR. In other words, the likelihood that a drug-ADR association is causal will increase if we observe the association between similar drugs and ADRs.

We propose a novel semi-supervised model called the analogy-based active voting (AAV) that represents the analogy criterion as a voting process of similar drugs. Similar drugs vote for the existence of causal relationship in the drug-ADR association of interest if they have association with the ADR. We call a set of similar drugs as a *committee*. The voting rate of similar drugs in the committee is used as the likelihood measure for assessing the causal relation of this association.

We represent each drug x_i by two features: (i) the mechanism of actions and targets; and (ii) the list of associated ADRs that are extracted from clinical notes, denoted by $F_{x_i} = y|x_i \rightarrow y$. The first feature is used for grouping similar drugs to establish committee, while the second one is used for the voting process of similar drugs. To push non-causal associations down in rank list, we select similar drugs (towards x_i) whose lists of associated ADRs are different from that of x_i. As such, ADRs, which have association with most drugs in the committee, will get a higher voting rate than ADRs that have association with a small number of drugs in the committee. Note that under the analogy criterion, ADRs that are associated with most drugs in the committee are likely to be caused by these drugs. We propose methods for measuring the difference among lists of associated ADRs of similar drugs that are useful for selecting similar drugs to establish a good committee.

Essentitally, the difference of F_{x_i} is resulted from the diversity of pharmaceutical therapies. In fact, the diversity of co-morbidity presenting in patients results in the diversity of treatment because the treatment need to adapts for each specific case (case-based treatment). That means different combinations of drugs are prescribed for different patients. The explanation in detail of how the diversity in treatment implies the difference of F_{x_i} will be mentioned in Subsection 4.9.5.

The analogy criterion indeed cannot cover all cases that each drug causes its own ADRs which its similar drugs do not cause. Therefore, in our model, we incorporate the voting rate of each drug-ADR pair with its association strength as the likelihood for assessment.

4.6 Electronic medical record Data

In this case study, we utilize the MIMIC-III database[1] for assessing the causal relationship between drugs and ADRs observed in practical treatment. The term MIMIC-III stands for the Medical Information Mart for Intensive Care III [52], which contains demographics, laboratory tests, clinical notes of more than forty thousand patients n the Beth Israel Deaconess Medical Center for supporting a wide range of studies in medicine. In this work, we utilize clinical notes and prescriptions that are available in the "NOTEEVENTS" and "PRESCRIPTIONS" tables, respectively. The prescriptions provide the information of drug names with starting and ending dates when drugs and being prescribed. The clinical narratives provide the information of patients' symptoms, adverse events, and abnormalities occurring during the treatment process.

[1]https://mimic.physionet.org

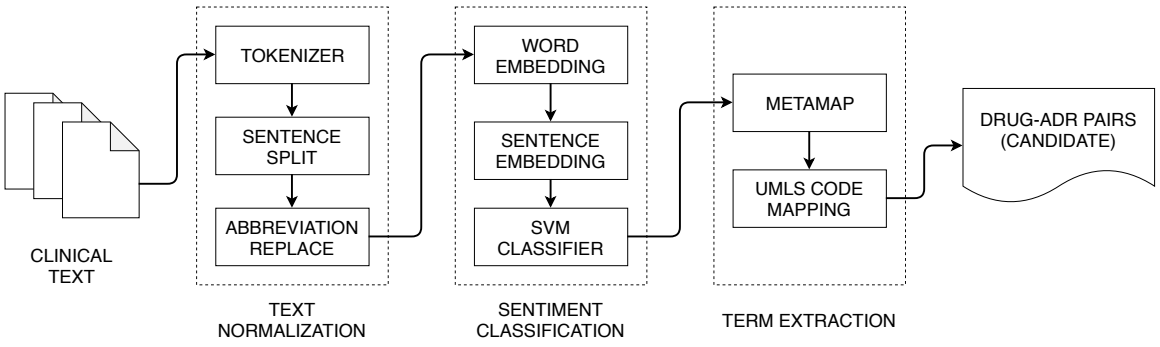

Figure 4.2: Data preprocessing pipeline.

4.7 Data preprocessing

The data preprocessing aims to extract all possible drug-ADR pairs where the ADR is suspected to be caused by the drug from clinical texts. The pipeline of this process is illustrated in Figure 4.2.

4.7.1 Text normalization

In the first step of data preprocessing, we utilize the word tokenizer, which is available in the NLTK package[2], for separating words from punctuations, commas, and semicolons, etc. Next, we split clinical notes, which are documents, into sentences and remove punctuations, commas, etc. We replace special tokens by their corresponding label. For example, the number in the texts is replaced by the label "#NUM". Additionally, we substitute abbreviations such that "SOB" is replaced into "short of breath" by utilizing a dictionary of 800 abbreviations with their unique term for replacing.

4.7.2 Sentiment classification

In clinical texts, words and phrases which indicate ADRs can be mentioned in both positive and negative sentences. The negative sentences mention the appearance of adverse events or reactions during the treatment process. Meanwhile, the positive ones mention the improvement in the treatment or effectiveness of drugs although they may contain terms indicating ADRs. For example, we consider the following sentence:

"The patient is less nauseous than previous".

In spite of containing the negative word "nauseous", this sentence mentions the drug efficacy for curing diseases (or symptoms) rather than side effects. This work focuses on detecting ADRs caused by inappropriately prescribing drugs, hence, eliminating positive sentences helps to avoid wrongly selecting drug-ADR pairs which are extracted from these sentences.

Utilizing machine learning approach for sentiment classification on clinical texts must be faced with the problems of implicit sentiment, various negation terms, and short

[2]https://www.nltk.org

texts [25]. The various negation terms make a difficulty in identifying the sentiment orientation of sentences, and the short text makes the measurement of semantic similarity between sentences imprecise. Hence, we solve these problems by learning embedding vectors for sentences in clinical notes.

Before learning the embedding vector for each sentence, we present each word in a sentence by its own distributed vector. The word distributed vector is generated by taking the context (surrounding words) of each word into account, which helps to reflect the word semantic. We produce words' vectors by utilizing the Word2Vec model [73]. We train this model implemented in the Gensim package[3] with approximately 260,000 clinical notes (documents) in the MIMIC-III database. The vectors of sentences are generated by summing up vectors of words in these sentences.

To classify sentences into positive ones and negative ones, we use the binary support vector machine (SVM) classifier. For training, we feed to this model 7000 sentences which are extracted from clinical notes and annotated with two labels "1" (indicating positive and neural sentences) and "-1" (indicating negative sentences). In the training set, the number of positive and negative sentences is approximately equal. We present each sentence by its corresponding vector by using the embedding method mentioned above. To evaluate the performance of the classifier system as well as select the most likely hyperparameters (e.g., kernel function, kernel parameter) for the model, we randomly divide the annotated data into ten folds in which six folds are used for training and the rest is used for testing. We repeat this process in ten times and then take the precision average over iterations. The accuracy attains approximately 86% with the RBF kernel.

4.7.3 Term extraction

After determining negative sentences by using the sentiment classifier, we use the MetaMap[4] to extract terms which indicate ADRs [6, 7]. The MetaMap is a well-known Natural Language Processing tool for biomedical texts whose main functions include: parsing an input sentence or paragraph into words and phrases; assigning an appropriate semantic label for each word and phrase. To select terms indicating ADRs, we base on three labels "Acquired Abnormality", "Finding", and "Sign and Symptom". The extracted terms are mapped to their corresponding unique concepts and IDs which are defined in Unified Medical Language System (UMLS). For example, the UMLS ID of term "abdominal cramps" is C0000729. After extracting ADR terms, we determine all possible drug-ADR pairs (candidates) for assessing their causal relations based on the starting and ending time when drugs are being prescribed and the creating time of clinical notes.

4.8 Preliminaries

Our proposed model for drug-ADR causality assessment is based on the analogy criterion which is one of nine Bradford Hill criteria, and incorporates several existing association strength measures. Hence, in this section, we make a brief introduction about nine Bradford Hill criteria and several association measures.

[3]https://radimrehurek.com/gensim

[4]https://metamap.nlm.nih.gov/

4.8.1 Bradford Hill criteria

Nine Bradford Hill criteria provide general epidemiological principles for inferring the causal relationship between interventions and effects, which have been widely applied in many medical studies. Given a drug-ADR pair, the causality assessment for this pair according to these principles is specified as follows:

1. Association strength: a measure of the dependence between the drug and the ADR.
2. Consistency: is the association between the drug and the ADR found in different databases?
3. Specificity: is the association between drugs and ADRs unique?
4. Temporality: does ADRs occur after prescribing drugs?
5. Biological gradient: the influence of the drug on biological factors that cause the ADR.
6. Plausibility: is it possible to exist a causal mechanism under the co-occurrence of the drug and the ADR?
7. Coherence: does the drug causing the ADR make sense or contradict to known knowledge?
8. Experimentation: does the observation of the ADR start and stop in synchronizing with the drug?
9. Analogy: Could similar drugs cause similar ADRs?

So far, the first eight criteria, which support the direct assessment on drug-ADR pairs, have been exploited in most of existing work. In these criteria, evaluating drug-ADR pairs based on their association strength has been widely used so that many measures have been developed for quantifying the association strength, as presented in Subsection 4.8.2. Meanwhile, the analogy criterion, which demands the consideration of similar drugs for evaluation, has not been exploited. Hence, this motivates our work to investigate whether this criterion is applicable for assessing the drug-ADR causality.

4.8.2 Drug-ADR association measurement

Various statistical measures have been developed and used for measuring strength of drug-ADR associations. This section makes a brief introduction regarding how these measures characterize the likelihood that a drug-ADR pair (i.e., candidate, denoted by $x \rightarrow y$) has causal relationship. Furthermore, these measures are used as a part in our proposed method.

Several measures are based on contingency tables which provide probabilities that a drug and an ADR co-occur and do not co-occur. Table 4.3 shows an example of a 2×2 contingency table in which we have:

a: the number of patients who used the drug x and the ADR y was observed.

Table 4.3: Contingency table of two random variables (x and y)

	$y = yes$	$y = no$
$x = yes$	a	b
$x = no$	c	d

b: the number of patients who used the drug x, but the ADR y was not observed.

c: the number of patients who did not use the drug x, but the ADR y was observed.

d: the number of patients who did not use the drug x and the ADR y was not observed.

$n = a + b + c + d$: the total number of patients under consideration.

The independence between x and y can be tested by using the χ^2 test that relies on the contingency table. This is calculated as the following:

$$\chi^2 = \frac{n \times (a \times d - b \times c)^2}{(a+b) \times (c+d) \times (b+d) \times (a+c)} \tag{4.1}$$

This value is used to test two hypotheses: one is that x and y are not associated (null hypothesis); the other is that they are associated. To determine whether the null hypothesis can be rejected, we compare this value to a critical value which is estimated from the χ^2 distribution with a given degree of freedom and a level of significance. We reject the null hypothesis if the χ^2 value is greater than the critical value.

Not only χ^2 test, relative odds ratio (i.e., ROR) and 90% confidence interval of ROR (i.e., ROR_{05}) are also based on the contingency table as follows:

$$ROR(x \to y) = \frac{a/c}{b/d} \tag{4.2}$$

$$ROR_{05}(x \to y) = \exp\left(\ln\left(\frac{a/c}{b/d}\right) - 1.645 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right) \tag{4.3}$$

Other measures have been also used to compute drug-ADR association strength such as relative reporting ratio (i.e., RR), confidence (i.e., $conf$), leverage, unexlev, RankRatio, and information component (i.e., IC). The formulations of these measures are a follows:

$$RR(x \to y) = \frac{n \times \text{supp}(x \to y)}{\text{supp}(x) \times \text{supp}(y)} \tag{4.4}$$

$$\text{conf}(x \to y) = \frac{\text{supp}(x \to y)}{\text{supp}(x)} \tag{4.5}$$

$$\text{leverage}(x \to y) = \text{supp}(x \to y) - \text{supp}(x) \times \text{supp}(y) \tag{4.6}$$

$$\text{unexlev}(x \hookrightarrow y) = \text{supp}(x \hookrightarrow y) - \text{supp}(x) \times \text{supp}(\hookrightarrow y) \tag{4.7}$$

$$\text{RankRatio}(x \hookrightarrow y) = \frac{\text{rank}_{\text{leverage}(x \to y)}}{\text{rank}_{\text{unexlev}(x \hookrightarrow y)}} \tag{4.8}$$

where

n: the total number of patients

$\text{supp}(x \to y)$: the proportion of patients who are prescribed the drug x and the ADR y is observed.

$\text{supp}(x)$: the proportion of patients who are prescribed the drug x.

$\text{supp}(y)$: the proportion of patients who have the ADR y after being prescribed any drug.

$\text{supp}(x \hookrightarrow y)$: the proportion of patients who have the ADR y when the drug x is being prescribed, and do not have the ADR y in a defined time period before the first time the drug x is used. We set the length of time period prior to the first time each drug is used to two days.

$\text{supp}(\hookrightarrow y)$: the proportion of patients who have never used the drug x and have the ADR y, plus with $\text{supp}(x \hookrightarrow y)$.

$$\begin{aligned} \text{IC} &= \frac{n_{xy}^t + 1/2}{E_{xy}^t + 1/2} \\ E_{xy}^t &= n_{x.}^t \frac{n_{.y}^t}{n_{..}^t} \end{aligned} \tag{4.9}$$

where

$n_{x.}^t$: the number of patients who use the drug x for the first time with an active follow up in time period t.

$n_{.y}^t$: the number of patients who are prescribed any drug for the first time and have event (or ADR) y within time period t.

$n_{..}^t$: the number of patients who use any drug for the first time with an active follow up in time period t.

E_{xy}^t: the expected number of patients who use the drug x and then have event y in time period t.

n_{xy}^t: the number of patients who use the drug x for the first time and event y occurs within time period t.

4.9 Analogy-based active voting

We propose a novel model, called the analogy-based active voting (AAV), for detecting drugs which are assessed to cause ADRs because of unnecessary polypharmacy. This model is based on the analogy criterion–one of nine Bradford Hill criteria, and applied for EMR data. For predicting drug-ADR causal relations, the model plays the role as the first screening process that only takes into account the information of the drug-ADR co-occurrence during the treatment and the drug mechanism of actions without any additional information such as dose, etc.

4.9.1 Do similar drugs cause similar ADRs?

We need to evaluate the feasibility of applying the analogy criterion for assessing the cause of ADRs. In other words, we need to investigate how likely that similar drugs cause similar ADRs is in terms of medical domain knowledge. Furthermore, this investigation poses criteria for selecting similar drugs in our proposed model.

Firstly, we consider whether two drugs, which have the same mechanism of action, are likely to cause similar ADRs. For example, Fluvastatin and Rosuvastatin are cholesterol-lowering statin drugs that target to inhibit the hepatic enzyme hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase. Because HMG-CoA reductase converts HMG-CoA to mevalonate in cholesterol synthesis, this inhibition results in the decrease in hepatic cholesterol levels. The change of membrane cholesterol leads to the change of membrane fluidity in tissues of skeletal muscles which can affect ion channels and modify muscle membrane excitability. Therefore, this change may cause several side effects such as myositis, myalgia, and rhabdomyolysis.

Secondly, we investigate whether two drugs, which have different mechanisms of action, can cause similar ADRs. For example, we consider two drugs: Nifedipine and Nitroglycerin. These drugs are often utilized for curing hypertension. The mechanisms of action of these two drugs are described in detail in Table 4.4. The table shows that although two drugs Nifedipine (whose essence is a calcium blocker) and Nitroglycerin (whose essence is a nitrate) act in different ways, both drugs target to relax the smooth muscle cells and dilate the coronary for reducing blood pressure. These actions result in several side effects such as headache, dizziness, and nausea because of reducing the blood pressure.

Through two investigations mentioned above provide us several viewpoint and criteria for determining similar drugs. Furthermore, they indicate the feasibility of applying the analogy principle for assessing drug-ADR causality. In general, two drugs are likely to be similar in terms of causing similar ADRs if:

- The drugs have a similar mechanism of action.
- The target of drugs' actions is similar.

4.9.2 Model intuition

As mentioned in Section 4.5, the analogy criterion is the groundwork of our proposed model which tightens up the constraint for inferring and selecting causative drug-ADR pairs. Concretely, although non-causative drug-ADR pairs can be frequently co-occurred,

Table 4.4: Nifedipine and Nitroglycerin mechanisms of action

Drug	Mechanism of action
Nifedipine	• Decreasing arterial smooth muscle contractility and vasoconstriction by inhibiting the influx of calcium ions through L-type calcium channels (i.e., the calcium blocker). • Calcium ions entering the cell through these channels bind to calmodulin. • Calcium-bound calmodulin activates myosin light chain kinase (MLCK). • Activated MLCK catalyzes the phosphorylation of myosin light chain, which leads to muscle contraction. • inhibition of the influx of calcium inhibits the contractile process of smooth muscle cells, which causes the coronary dilation and increased oxygen delivery to the myocardial tissue. • Nifedipine results in the decrease in blood pressure.
Nitroglycerin	• Nitroglycerin is converted to nitric oxide (NO) which activates the enzyme guanylate cyclase. • This stimulates the synthesis of cyclic guanosine 3', and 5'-monophosphate (cGMP), which results in the dephosphorylation of the myosin light chain of the smooth muscle fiber. • This causes the relaxation of the smooth muscle cells and vasodilation.

we have chance to recognize and eliminate these pairs when these pairs have not association with other similar drugs (towards the drug of interest). This may help to avoid confusions in assessing ADR causality which are induced by the use of multiple drugs. In fact, each drug has its own ADRs which its similar drugs have not. Our proposed model just prioritizes common ADRs of similar drugs by assigning the high score for them.

Given a drug-ADR association, we model the analogy criterion for assess whether this association has causal relation as a voting process of similar drugs. The likelihood that this association is causative is estimated by the number of similar drugs which are associated with the ADR. Relying on medical domain knowledge, we determine similar drugs in terms of their mechanism of action, and then establish groups of such similar drugs that are called committees. We denote a committee by the set $C = x_1, x_2, ..., x_n$ where x_i indicates a drug and $x_i \sim x_j$ with $i \neq j$. Besides the domain knowledge, by taking the clinical notes and prescriptions into account, we can extract drug-ADR association where the ADR is observed after prescribing the drug, which is denoted by $x_i \rightarrow y$. According to the analogy criterion, the drug x_i is believed to cause the ADR y if its similar drugs $x_j \in C$ also has the association with y.

Generally, the AAV model includes two main steps:

1. Establishing committees based on expertise knowledge, as presented in SubSection 4.9.3.

2. Estimating the voting rate of each drug-ADR association which is used for ranking associations according to their likelihood of causality, as presented in Subsection 4.9.4. Relying on the ranked list of associations, we can select associations (at the top of the list) which are likely to have the causal relationship.

As mentioned in Section 4.5, the conflict in the voting of similar drugs is exploited for eliminating non-causative drug-ADR pairs. The conflict is characterized through the difference of sets F_{x_i} extracted from the EMR data. It is needful to measure the diversity of drugs in committees which is presented in detail in Subsection 4.9.5. The measure will help to select similar drugs in each committee that maximize the committee's diversity.

4.9.3 Establishing committee for voting

Relying on two criteria for selecting similar drugs as mentioned in Subsection 4.9.1, we can group drugs to establish committees based on the pharmacological expertise knowledge. In this work, similar drugs are grouped based on the information of their mechanism of action which is available in the DrugBank[5]. We consider four committees C1, C2, C3, C4 which are briefly presented in Table 4.5.

4.9.4 Estimating voting rate of drug-ADR pairs

As mentioned in Section 4.5, For each drug x_i in the committee C, $F_{x_i} = \{y_j | x_i \rightarrow y_j\}$ is the set of ADRs which have the association with x_i (these ADRs are observed during period that the drug x_i is being prescribed). Let $F = F_{x_1} \cup F_{x_2} \cup ... \cup F_{x_n}$ be the set of all

[5]https://www.drug-bank.ca/

Table 4.5: Expertise-based committee establishment

Committee	Drug	Common indication	Mechanism of action
C1	Cefuroxime, Valganciclovir, Ribavirin, Meropenem, Cefazolin, Oseltamivir, Albendazole, Miconazole	Antibiotic, treat bacterial infections, antiviral, antiworm, antifungal	Binding to protein, RNA to inhibit bacterial/viral/fungal cell synthesis. Battling with bacteria may result in adverse effects.
C2	Fluvastatin, Ezetimibe, Rosuvastatin	Lowering cholesterol in blood, reducing cholesterol absorbed by the body	Inhibiting the hydroxylmethylglutaryl-coenzyme A reductase or cholesterol transport protein to reduce cholesterol. The change of cholesterol membrane can lead to adverse effects of muscle.
C3	Guanfacine, Timolol, Lisinopril, Diltiazem, Nicardipine, Labetalol, Metoprolol, Valsartan, Nifedipine, Nitroglycerin	Treat hypertension	Drugs in this group act in different ways but all of them aim to reduce blood pressure by dilating the coronary or reducing heart rate. Overdose can cause ADRs regarding low blood pressure.
C4	Lorazepam, Alprazolam, Diazepam	Benzodiazepine, treat anxiety, panic disorders	All of drugs in this group are benzodiazepine. Adverse effects can result from inhibiting neurotransmitter.

ADRs observed when all drugs in the committee C are being prescribed. We represent drugs in the committee with their associated ADRs by a voting matrix which is denoted by $\boldsymbol{V}_{|C|\times|F|}$. The voting matrix $\boldsymbol{V}$ is a binary matrix in which each row corresponds to each drug in the committee, and each column corresponds to each ADR $y_j \in F$. The value in each cell of this matrix gets one if there exists the association $x_i \rightarrow y_j$, otherwise it gets zero.

From the voting matrix $\boldsymbol{V}$, the likelihood that a drug-ADR association ($x_i \rightarrow y_j$ where $x_i \in C$ and $y_j \in F$) has causal relationship according to the analogy criterion is estimated as the following:

$$\begin{aligned} Vote(x_i \rightarrow y_j|C) &= exp\Big(v(C, y_j)\Big) \times assoc(x_i \rightarrow y_j), \\ v(C, y_j) &= \sum_{i=1}^{n} \boldsymbol{V}_{ij}, \end{aligned} \tag{4.10}$$

where

- $v(C, y_j)$ is the number of drugs in the committee C which vote for the causal relationship between x_i and y_j. This value is calculated by summing up all cells in the j^{th} column of the voting matrix $\boldsymbol{V}$. The exponential function is utilized with the aim of emphasizing on the order of drug-ADR associations according to the voting rate $v(C, y_j)$ in the ranked list.

- The $assoc(x_i \rightarrow y_j)$ measures the association strength between x_i and y_j. This measure is added to assessed two drug-ADR associations which have the same value of $v(C, y_j)$. We can used measures of association mentioned in Subsection 4.8.2 for estimating the $assoc(x_i \rightarrow y_j)$, however, only measures, which produce only positive values such as ROR, RR, conf, and RankRatio, are preferable to preserve the proximity (order) among associations formed by the $v(C, y_j)$. Indeed, drugs can cause their particular ADRs which their similar drugs do not cause, thus, we also incorporate the drug-ADR association strength for recognizing these ADRs.

4.9.5 Evaluating the committee diversity

Non-causative drug-ADR associations, which coincidentally and frequently observed when using multiple drugs, can be recognized by exploiting the disagreement of voters (drugs in a committee). This is indicated by the difference of sets F_{x_i}.

Indeed, the difference of F_{x_i} is resulted from the difference of drugs which used with similar drugs in a committee for treatment, which is illustrated by a simple example shown in Figure 4.3. The explanation is as follows:

- Let x_1, x_2, and x_3 be three similar drugs, denoted by $x_1 \sim x_2 \sim x_3$. Two drugs x_4 and x_5 are drugs which are co-prescribed with such three similar drugs.

- Five drugs are used for treating three patients, and then we obtain associations between such drugs and ADRs (including y_1 and y_2) as shown in the figure.

- From obtained associations, we can export sets F_{x_i} with $i = 1, ..., 5$ for each drug.

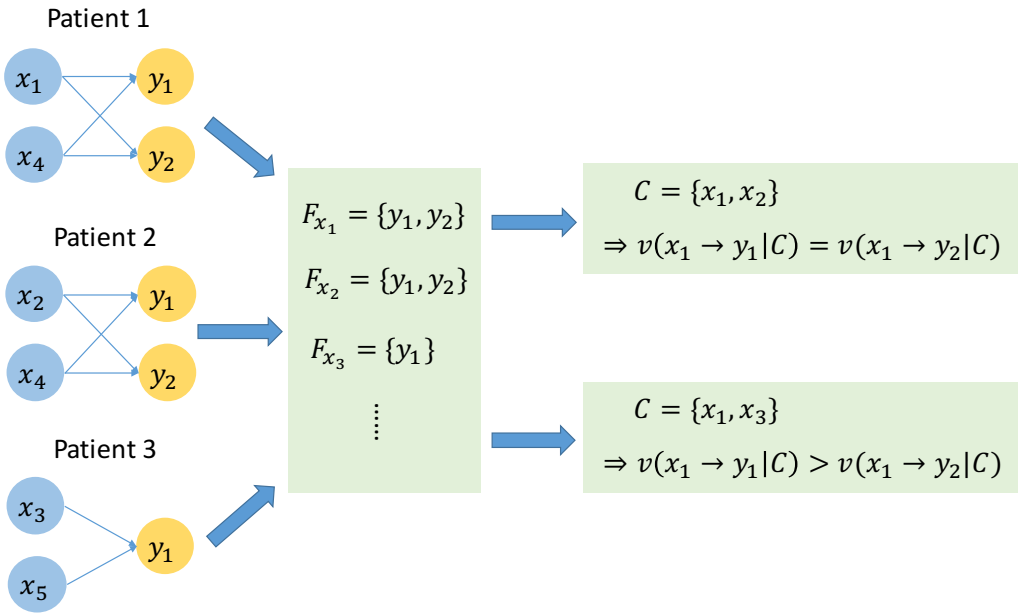

Figure 4.3: An example illustrating that the difference of drugs, co-prescribed with similar drugs, results in the diversity in voting for drug-ADR causality in committees. Note that three drugs x_1, x_2, x_3 are similar: $x_1 \sim x_2 \sim x_3$.

- Suppose, the ground truth is that: (1) x_1 causes y_1 and does not cause y_2; (2) x_4 causes y_2.
- If the committee C includes drugs x_1 and x_2: $C = \{x_1, x_2\}$, the voting rate for the causative relation between x_1 and y_1 is equal to that between x_1 and y_2 because both drugs are prescribed with the same drug x_4.
- If $C = \{x_1, x_3\}$, the voting for the causal relation between x_1 and y_1 is greater than that between x_1 and y_2. This reflects the ground truth. Indeed, drugs x_1 and x_3 are prescribed with different drugs, so we do not see the occurrence of y_2 after prescribed x_3.

Evaluating the diversity of committees is not straightforward that requires the consideration in multiple views. Because in this case, the committee diversity is indicated through the difference of F_{x_i}, we can evaluate this by relatively comparing the intersection of F_{x_i} towards their union. Thus, the committee diversity can be measured by estimating the proportion between the number of ADRs voted by all drugs in the committee (the intersection) and the total number of ADRs (the union). We denote this measure by $div_i(C)$. The voters in the committee are more divergent if the value of $div_i(C)$ is small.

$$div_i(C) = \frac{|F_{x_1} \cap F_{x_2} \cap ... \cap F_{x_n}|}{|F|} \tag{4.11}$$

In other view, the committee diversity can be evaluated by examining the contradiction of each drug to the majority in the committee. To this end, we measure the disagreement between the votes of each drug for ADRs and the overall votes of the rest for such ADRs.

The measure of disagreement is denoted by $dis(x_i, C/\{x_i\})$. We first determine the voting consensus of drugs in $C/\{x_i\}$ as follows:

- Let $\boldsymbol{V} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_n]^T$ where $\boldsymbol{v}_i$ is a row of the voting matrix $\boldsymbol{V}$ that indicates the votes of the drug x_i for all ADRs in F_{x_i}.
- Deriving a matrix $\boldsymbol{V}^{(i)}$ from the matrix $\boldsymbol{V}$ by eliminating i^{th} row.
- Obtaining the voting consensus of the drugs in the set $C/\{x_i\}$ by extracting a vector $\boldsymbol{r}^{(i)}$ based on the matrix $\boldsymbol{V}^{(i)}$. The value at j^{th} element in the $\boldsymbol{r}^{(i)}$ (a binary value) is the value which accounts for the highest proportion in cells at j^{th} column of $\boldsymbol{V}^{(i)}$. We ignore columns that we cannot find the major value in these columns. Vector $\boldsymbol{r}^{(i)}$ is a one-hot vector.

This disagreement between the drug x_i and the drugs in the set $C/\{x_i\}$ is characterized by the normalized Hamming distance between the row $\boldsymbol{v}_i$ and the vector $\boldsymbol{r}^{(i)}$ as the following:

$$dis(x_i, C/\{x_i\}) = \frac{ham(\boldsymbol{v}_i, \boldsymbol{r}^{(i)})}{\max_{i=1,...,n} ham(\boldsymbol{v}_i, \boldsymbol{r}^{(i)})} \tag{4.12}$$

where $ham(\boldsymbol{v}_i, \boldsymbol{r}^{(i)})$ is the Hamming distance between $\boldsymbol{v}_i$ and $\boldsymbol{r}^{(i)}$. As noted above, if j^{th} column of $\boldsymbol{V}^{(i)}$ is ignored because of having no major value, j^{th} element in $\boldsymbol{v}_i$ is also ignored in estimating the Hamming distance between $\boldsymbol{v}_i$ and $\boldsymbol{r}^{(i)}$.

A diverse committee needs to includes many drugs which have a conflict to the majority when voting for ADRs. Therefore, the diversity of the committee here, denoted by $div_m(C)$, is measured as a proportion between the number of drugs x_i that satisfy the constraint of $dis(x_i, C/\{x_i\}) \geq 0.5$ (that means x_i conflict to the rest of the committee) and the total number of drugs in the committee, as the following:

$$div_m(C) = \frac{|\{x_i | x_i \in C; dis(x_i, C/\{x_i\}) \geq 0.5\}|}{n} \tag{4.13}$$

This assessment method is just applicable for committees which contain more than three drugs because the set $C/\{x_i\}$ must contain at least three drugs for finding the major value of each column.

The assessment of the committee diversity is based on both measures $div_i(C)$ and $div_m(C)$. According to the $div_i(C)$ measure, a committee is more diverse if its value is small while according to the $div_m(C)$ measure, its value is large. The rank of each committee according to each measure will be summed up for comparing committee.

4.10 Results and discussion

4.10.1 Data preparation and ground truth

We export the prescription and clinical notes of 8000 patients who are prescribed the drugs listed in Table 4.5. These clinical notes and prescriptions are processed to extract associations between drugs and ADRs from EMRs data, as presented in Section 4.7. The SIDER[6] [56], which is a well-known database of side effects, is utilized as the ground truth for confirming drug-ADR relations extracted from EMRs.

[6]http://sideeffects.embl.de

Table 4.6: Precision (%) obtained by using the existing measures for ranking drug-ADR associations.

	Top	χ^2	ROR	ROR_{05}	RR	Conf	Leverage	Unexlev	RankRatio	IC
C1	5%	5.88	1.85	11.11	1.85	**25.93**	9.26	5.56	3.7	7.41
	10%		7.41	5.56	2.78	**25.0**	8.33	3.7	3.7	3.7
	15%		5.52	4.29	4.91	**22.09**	7.98	3.07	3.07	4.29
	20%		5.07	5.07	4.15	**18.43**	5.99	2.76	2.76	5.07
C2	5%	6.38	0.0	0.0	0.0	**13.33**	**13.33**	**13.33**	**13.33**	0.0
	10%		3.33	0.0	0.0	**16.67**	13.33	6.67	6.67	0.0
	15%		2.22	6.67	2.22	**15.56**	13.33	6.67	4.44	4.44
	20%		5.0	6.67	5.0	**16.67**	13.33	6.67	5.0	5.0
C3	5%	11.29	2.8	6.54	0.93	**24.3**	13.08	4.67	3.74	0.93
	10%		3.27	7.94	1.87	**25.7**	8.39	4.04	2.8	3.42
	15%		4.35	9.63	3.42	**25.16**	8.39	4.04	2.8	3.42
	20%		5.83	10.96	4.66	**23.31**	7.23	3.5	2.8	5.36
C4	5%	10.17	2.63	5.26	2.63	**13.16**	5.26	0.0	0.0	2.63
	10%		2.6	7.79	2.6	**18.18**	6.49	2.6	3.9	2.6
	15%		2.59	5.17	2.59	**21.55**	8.62	3.45	2.59	4.31
	20%		4.52	7.1	3.87	**20.0**	7.1	3.23	1.94	4.52

4.10.2 Evaluation metric

By using measures for quantifying the likelihood that a drug-ADR association has causal relationship, we can rank such associations. Drug-ADR associations, which have the causative relation, are expected to be arranged in the top of the ranked list. Hence, to compare the effectiveness of such measures, we investigate how they highlight causative drug-ADR associations. We estimate the precision at the top of K associations in the ranked list. In this work, we examine the precision at the top of 5%, 10%, 15%, and 20% of associations in the ranked list.

4.10.3 Comparing the AAV with existing methods

Table 4.6 shows the precision at the top of 5%, 10%, 15%, and 20% of ranked drug-ADR associations by using existing association measures which are presented in Subsection 4.8.2. The experiments show that most of existing measures give low accuracies. In these measures, the confidence (Conf) significantly outperforms the others. Noting that the method for evaluating the χ^2 test is different from the others because we retrieve associations the reject the null hypothesis then estimate the precision of such associations without considering the top of K associations.

Table 4.7 shows the accuracy of our proposed measure for ranking drug-ADR associations. Four baseline measures, ROR, RR, Conf, and RankRatio are integrated to the voting rate for ranking associations. These measures are selected because they produce only positive values. The integrated measures are denoted by Vote_{ROR}, Vote_{RR}, Vote_{Conf}, and $\text{Vote}_{RankRatio}$, respectively. The table shows that the measure Vote_{Conf} outperforms the rest.

Table 4.7: Precision (%) obtained by using the proposed measure for ranking drug-ADR associations

	Top	Vote_{ROR}	Vote_{RR}	Vote_{Conf}	$\text{Vote}_{RankRatio}$
C1	5%	22.22	22.22	**37.04**	18.52
	10%	22.22	24.07	**28.7**	19.44
	15%	19.63	20.86	**22.09**	17.79
	20%	16.13	16.13	**18.43**	15.21
C2	5%	6.67	6.67	**13.33**	6.67
	10%	6.67	6.67	**13.33**	10.0
	15%	6.67	4.44	**15.56**	8.89
	20%	5.0	5.0	**16.17**	6.67
C3	5%	18.69	20.56	**29.91**	22.43
	10%	19.16	20.56	**25.23**	14.95
	15%	11.29	17.7	**25.47**	12.42
	20%	16.78	17.72	**22.84**	12.35
C4	5%	5.26	5.26	**13.16**	7.89
	10%	9.09	9.09	**18.18**	5.19
	15%	10.34	8.62	**21.55**	3.45
	20%	9.03	9.03	**20.0**	4.52

Relying on Tables 4.6 and 4.7, we compare the performance of detecting drug-ADR causal relations obtained by using our proposed method with that obtained by using existing methods in four committees. The comparison is showed in Figure 4.4. The figure shows that our proposed measure significantly outperforms three measures ROR, RR, and RankRatio. However, the Vote_{Conf} measure just outperforms the Conf measure at the top of 5% and 10% of ranked associations in the committee C1, and at the top of 5% of ranked associations in the committee C3. For committees C2 and C4, there is no improvement when using the proposed method.

In fact, the performance improvement when using the AAV model reflects its ability for dealing with the polypharmacy-induced confounding factors. Indeed, this improvement is resulted from the committee diversity that will be discussed in next section.

4.10.4 Association between the committee diversity and AAV performance

We aim to clarify the role of the committee diversity towards the performance of detecting drug-ADR causal relations as well as the ability of the AAV model for dealing with the polypharmacy-induced confounding. In addition, this analysis can give an explanation for the fact that the Vote_{Conf} just makes an improvement of performance when detecting ADRs caused by each drug in the committees C1 and C3.

As presented in Section 4.9.5, the diversity of each committee is quantitatively evaluated through two measures $div_i(C)$ and $div_m(C)$, which are presented in Table 4.8. We can rank four committees based on their diverse levels according to each criterion corresponding to each measure. The most diverse committee according each criterion of

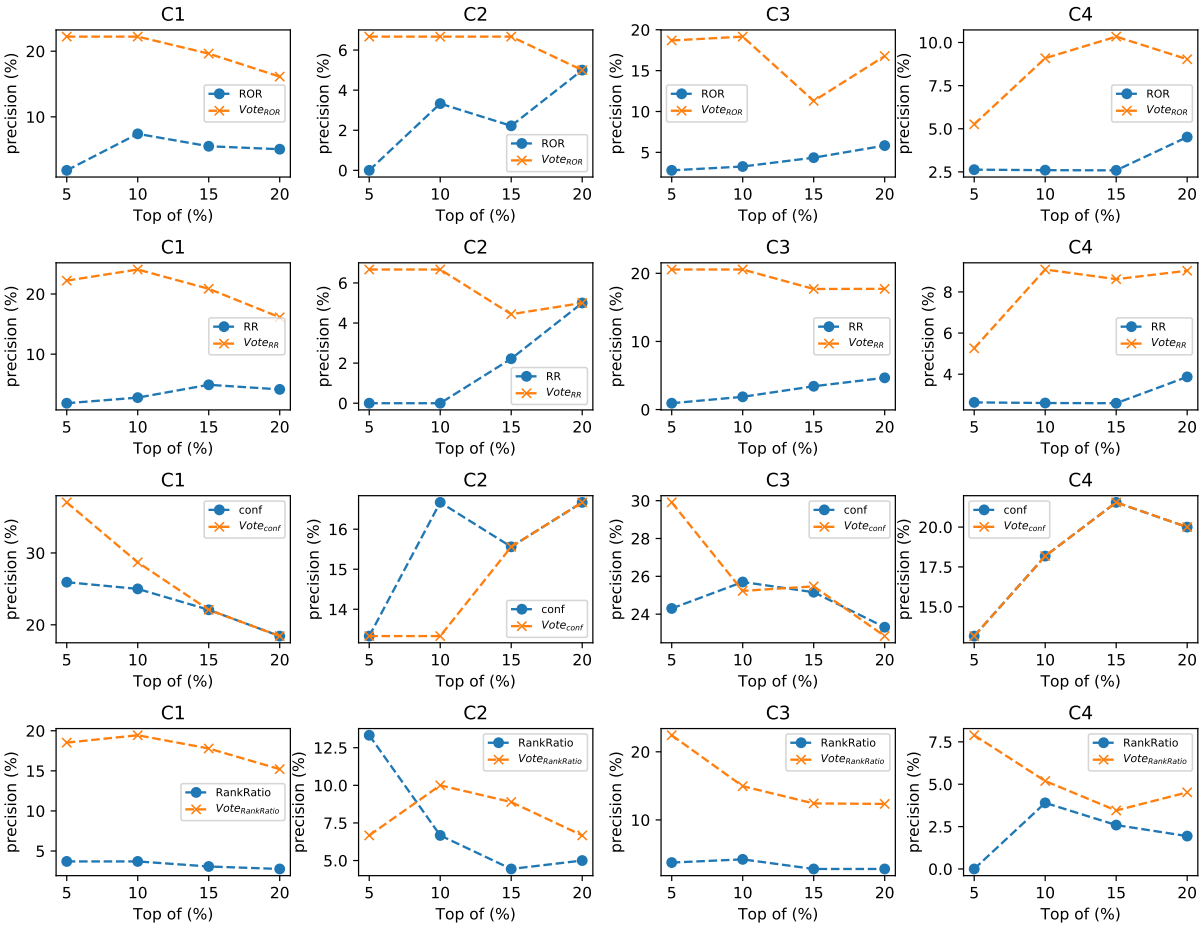

Figure 4.4: Comparing the proposed measures with existing (baseline) measures.

Table 4.8: Estimation of the $div_i(C)$ and $div_m(C)$ for committees C1, C2, C3, and C4 (the notation $\downarrow$ indicates the smaller value is better, and the notation $\uparrow$ indicates the larger value is better). The rank assigned for each committee is in the parentheses.

Committee	$div_i(C) \downarrow$	$div_m(C) \uparrow$
C1	0.003 (1)	0.75 (1)
C2	0.037 (3)	-
C3	0.02 (2)	0.2 (2)
C4	0.431 (4)	-

measures (indicated by $\downarrow$ and $\uparrow$) will get the first rank which is shown in the parentheses in Table 4.8.

Relying on the values of $div_i(C)$, we see that committees C1 and C3 are more diverse than committees C2 and C4 because of smaller values. In addition, we get an increase in performance of assessing drug-ADR causality when utilizing the AAV model for committees C1 and C3 while there has no improvement when using this model for committees C2 and C4. This evaluation shows an evidence for that the diversity of committees results in the better performance in drug-ADR causality assessment.

The $div_m(C)$ measure is not applicable for assessing the diversity of committees C2 and C4 because these committees do not meet the requirement of the number of drugs. Relying on the values provided by this measure, we see that the committee C1 is more diverse than the committee C3. The experiment also shows that utilizing the Vote_{Conf} measure for the committee C1 makes the accuracy increase from 25.95% to 37.04% (increases about 11%) at the top of 5% of ranked associations. Meanwhile, utilizing this measure for the committee C2 just makes the accuracy increase from 24.3% to 29.91% (increases about 5.5%). In addition, when considering the top of 10% of ranked associations, we do not gain any improvement by using the Vote_{Conf} measure for committee C3 while we still obtain the improvement by using this measure for C1 (the accuracy increases from 25.0% to 28.7%). The committee diversity evaluation based on the $div_m(C)$ also supports our hypothesis of the influence of the committee diversity on the improvement of the drug-ADR causality assessment performance.

4.10.5 Detecting infrequently observed drug-ADR causal relations

We perform the ability of the AAV model for detecting uncommon causative drug-ADR pairs. By using measures which are based on the frequency of drug-ADR co-occurrence, these pairs can be left out because of having a weak association strength.

Basing on the analogy criterion, causative drug-ADR associations, in which the drug is uncommonly used, can be recognized if the drug belongs to a large committee. The voting rate by a large number of similar drugs can help to highlight uncommon associations. Of course, only causative associations between uncommon drugs and common ADRs (i.e., ADRs are observed in many different treatment) can be recognized. The AAV model cannot recognize causative associations where both the drug and the ADR are rarely used and observed.

Table 4.9: An example of recognizing uncommon causative drug-ADR associations by the AAV model (considered at the top of 5% of ranked associations).

Causative drug-ADR association	ADR name	Rank by Vote_{Conf}	Rank by Conf
Ribavirin → C0042963	vomiting	7	55
Meropenem → C0042963	vomiting	9	77
Oseltamivir → C0042963	vomiting	10	85
Valganciclovir → C0042963	vomiting	14	90
Miconazole → C0042963	vomiting	15	91
Cefazolin → C0042963	vomiting	19	133
Lisinopril → C0027497	nausea	59	115
Metoprolol → C0027497	nausea	61	117

To examine this ability of the AAV model, we consider causative drug-ADR associations which are retrieved by selecting top of 5% ranked associations by using the Vote_{Conf} measure. Next, we consider the order of retrieved causative associations in the ranked list constructed based on values of the confidence (Conf) measure. We present the order of such associations according to the Vote_{Conf} and Conf measures in Table 4.9.

Table 4.9 shows that many causative drug-ADR associations, which have the low rank according to the Conf measure (hence, cannot be recognized), can be highlighted (have higher rank) according to the Vote_{Conf} measure. However, we can only recognize causative relations between uncommonly used drugs and commonly observed ADRs such as vomit, nausea, etc.

4.11 Chapter summary

This chapter aims to clarify the role of preserve the diversity in designing reference factors for avoiding confounding in analogy-based causality inference. To demonstrate that, we carry out a study on drug-ADR causality assessment with focusing on dealing with confounding caused by polypharmacy. In this study, we propose a semi-supervised model that for improving the performance of drug-ADR causality assessment. This method is called the analogy-based active voting (AAV). This model is based on the analogy criterion for causality inference that is proposed by Bradford Hill. In this work, we model the analogy-based causality inference as a voting process of similar drugs. The conflict in the voting, which is resulted from the diversity of drugs used for practical treatment, is exploited to eliminate non-causal drug-ADR pairs. The experimental results show the improvement in detecting causal drug-ADR pairs from EMRs when taking the diversity into account in constructing the set of reference drugs (similar drugs) for inferring.

Chapter 5

Conclusions and Future Work

5.1 Summary

In this dissertation, we concentrate on elucidating the role of diversity preservation in similarity-based inference. To the best of our knowledge, this is a novel and important problem in machine learning. However, this problem has not been intensively discussed in most existing studies. We demonstrate the necessary of preserving the diversity in similarity-based inference via two studies: (1) preserving the distinction of pairwise comparison in triplet of objects when measuring the similarity; (2) reference diversification in analogy-based causality inference. Each study provides a perspective of the problem of interest.

In both studies, we focus on model interpretation and explanation based on performance of models. In the first study, we explore characteristics of similarity measures that significantly affect the performance of approximating rough target functions. Relying on this, we establish general principles for effectively using similarity measures for mining material data. In the second one, we clarify that establishing committee in which similar drugs have different lists of associated ADRs can help to control polypharmacy-induced confounding as well as improve the accuracy in detecting causal relations between drugs and adverse reactions.

In the first study, we address the problem that measuring similarity of objects based on their representation is not consistent with the similarity of their target values as using non-ideal representations. It induces the roughness of target function (or surface) subject to these representations. For effectively approximating rough target functions, we hypothesize that it is needful to preserve the distinction of pairwise comparison in the triplet of objects when using similarity measures. In other words, by using these similarity measures, two objects are distinct in comparison with the third one. To verify this hypothesis, we investigate the appropriateness of similarity measures for fitting rough target function based on locally approximating this function. The main criterion for investigation is the number of neighbors of each data point determine by each measure in a predefined scope. We employ a protocol for the investigation that includes two main steps: (i) estimating the surface roughness; and (ii) evaluating similarity measures in context of fitting rough target function via $k-$nearest neighbors and kernel ridge regression. The experimental results show the high likelihood of our proposed hypothesis. Furthermore, relying in-

vestigations of similarity measures, we establish general principles for effectively using similarity measures that do not depend on a specific dataset and representation method.

In the second study, we concentrate on an important problem in post-marketing pharmaceutical surveillance – drug-ADR causality assessment. The main issue in this problem is to deal with confounding factors caused by polypharmacy. Confounding factors here are non-causal drug-ADR pairs that coincidentally and frequently co-occur. To control confounding factors, we employ the similarity-based causality inference method for inferring drug-ADR causality based on the analogy criterion (one of nine Bradford Hill criteria). This criterion states that similar drugs may cause similar ADRs. We mimic the criterion by a voting process of similar drugs for the existence of causal relation in the drug-ADR association of interest. Relying on that, we propose a novel model for drug-ADR causality inference, called analogy-based active voting. In this model, groups of similar drugs used for voting are called committees. We represent each drug by two main features: (i) mechanism of actions and targets; and (ii) list of associated ADRs extracted from clinical notes. The first feature is used for establishing committees, and the second one is used for voting of drugs in these committees. Our work shows that diversifying drugs in committees according to the second feature can induce the conflict in voting process of similar drugs. This plays an important role for recognizing non-causal drug-ADR pairs and improving performance of detecting drug-ADR causal relations.

5.2 Contributions to knowledge science

In this dissertation, we address a novel problem in machine learning – preserving the diversity in similarity-based inference. The concept of diversity preservation in similarity-based inference is elucidated through concrete case studies. Through each study, we provide a view of this concept. In this work, we hypothesize and verify properties of operators an representations that strongly affect performance of learning models such as: how similarity measures preserve the distinction of pairwise comparison in a triplet of objects; and diversifying reference in similarity -based inference. We conceptualize these properties by providing quantitative definitions. Based on these definitions, we can explain the performance of learning models. In the first study on measuring the similarity of materials, we establish general principles for effectively using similarity measures for mining material data that are generated from making the induction of investigations on similarity measures performance. In conclusion, our work focus on conceptualizing the diversity in similarity-based inference, and attempt to enrich knowledge about this.

5.3 Future work

Our work derives knowledge about diversity preservation in similarity-based inference based on the induction of observations on learning models performance with specific datasets. In the current work, we just verify our hypotheses by using limited data resources. To enhance the stability and generalization of these hypotheses, it is needful to verify them with more data sources in the future work.

Preserving the diversity is an important concept in machine learning. It helps to

improve the performance of machine learning models. In this dissertation, we just provide perspective of this concept in terms of similarity-based inference models. Hence, it demands to investigate this concept in other contexts. This motivates our work in the future.

Appendices

Appendix A

Kernel ridge regression - dual form of ridge regression

We rewrite the optimization problem for ridge regression as

$$\begin{aligned} \underset{\beta,\mathbf{r}}{minimize} \quad & \frac{1}{2}\left(||\mathbf{r}||_2^2 + \lambda||\beta||_2^2\right) \\ \text{subject to} \quad & \mathbf{r} = \mathbf{X}\beta - \mathbf{y} \end{aligned} \tag{A.1}$$

The solution is equivalent to

$$\begin{aligned} &\min_{\beta,\mathbf{r}} \max_{\alpha} L(\beta, \mathbf{r}, \alpha) \\ =&\min_{\beta,\mathbf{r}} \max_{\alpha} \left(\frac{1}{2}||\mathbf{r}||_2^2 + \frac{\lambda}{2}||\beta||_2^2 + \alpha^T(\mathbf{r} - \mathbf{X}\beta + y)\right), \end{aligned} \tag{A.2}$$

where $L(\beta, \mathbf{r}, \alpha)$ is the Lagrangian function. We solve the minimization problem by setting to zero the first derivatives of the Lagrangian function according to β and $\mathbf{r}$:

$$\begin{aligned} &\frac{\partial L}{\partial \beta}(\beta, \mathbf{r}, \alpha) = 0 \Rightarrow \lambda\beta - \mathbf{X}^T\alpha = 0 \Rightarrow \hat{\beta} = \frac{1}{\lambda}\mathbf{X}^T\alpha \\ &\frac{\partial L}{\partial \mathbf{r}}(\beta, \mathbf{r}, \alpha) = 0 \Rightarrow \mathbf{r} + \alpha = 0 \Rightarrow \hat{\mathbf{r}} = -\alpha \end{aligned} \tag{A.3}$$

Plugging $\hat{\beta}$ and $\hat{\mathbf{r}}$ into the Lagrangian function obtains

$$\begin{aligned} L(\hat{\beta}, \hat{\mathbf{r}}, \alpha) &= \frac{1}{2}||\alpha||_2^2 + \frac{1}{2\lambda}||\mathbf{X}^T\alpha||_2^2 + \alpha^T(-\alpha - \frac{1}{\lambda}\mathbf{X}\mathbf{X}^T\alpha + \mathbf{y}) \\ &= -\frac{1}{2}||\alpha||_2^2 - \frac{1}{2\lambda}\alpha^T\mathbf{X}\mathbf{X}^T\alpha + \alpha^T\mathbf{y} \end{aligned} \tag{A.4}$$

Now, the dual problem is $\max_\alpha L(\hat{\beta}, \hat{\mathbf{r}}, \alpha)$, which is equivalent to the following (noting that $\lambda \geq 0$):

$$\min_{\alpha} \left(\frac{1}{2}\alpha^T(\mathbf{K} + \lambda\mathbf{I})\alpha - \lambda\alpha^T\mathbf{y}\right), \tag{A.5}$$

where $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ is called the kernel matrix. To obtain α, we also set the first derivatives of the dual objective function to zero, to obtain

$$\begin{aligned} (\mathbf{K} + \lambda\mathbf{I})\alpha - \lambda\mathbf{y} &= 0 \\ \Rightarrow \alpha &= \lambda(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \end{aligned} \tag{A.6}$$

Based on Equation A.3, we obtain

$$\beta = \mathbf{X}^T(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \tag{A.7}$$

Bibliography

[1] Taofikat B Agbabiaka, Jelena Savović, and Edzard Ernst. Methods for causality assessment of adverse drug reactions. *Drug safety*, 31(1):21–37, 2008.

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the second ACM international conference on web search and data mining*, pages 5–14. ACM, 2009.

[3] Yannick Arimone, Bernard Bégaud, Ghada Miremont-Salamé, Annie Fourrier-Réglat, Mathieu Molimard, Nicholas Moore, and Françoise Haramburu. A new method for assessing drug causation provided agreement with experts' judgment. *Journal of clinical epidemiology*, 59(3):308–314, 2006.

[4] Elja Arjas and Jan Parner. Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics*, 31(2):171–187, June 2004.

[5] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[6] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[7] Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[8] Mukesh Bansal, Jichen Yang, Charles Karan, Michael P Menden, James C Costello, Hao Tang, Guanghua Xiao, Yajuan Li, Jeffrey Allen, Rui Zhong, et al. A community computational challenge to predict the activity of pairs of compounds. *Nature biotechnology*, 32(12):1213, 2014.

[9] Frédérique Barbosa and Dragos Horvath. Molecular similarity and property similarity. *Current topics in medicinal chemistry*, 4(6):589–600, 2004.

[10] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[11] Andrew Bate, Marie Lindquist, IR Edwards, Sten Olsson, Roland Orre, Anders Lansner, and R Melhado De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European journal of clinical pharmacology*, 54(4):315–321, 1998.

[12] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7):074106, 2011.

[13] Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–3218, 2004.

[14] L. C. Blum and J.-L. Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.*, 131:8732, 2009.

[15] Amy S.B. Bohnert, Maureen A. Walton, Rebecca M. Cunningham, Mark A. Ilgen, Kristen Barry, Stephen T. Chermack, and Frederic C. Blow. Overdose and adverse drug event experiences among adult patients in the emergency department. *Addictive Behaviors*, 86:66–72, November 2018.

[16] Timothy Brewer and Graham A. Colditz. Postmarketing surveillance and adverse drug reactions. *JAMA*, 281(9):824, March 1999.

[17] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1548–1560, 2010.

[18] Gabriele Capannini, Franco Maria Nardini, Raffaele Perego, and Fabrizio Silvestri. Efficient diversification of web search results. *Proceedings of the VLDB Endowment*, 4(7):451–459, 2011.

[19] Jaime G Carbonell and Jade Goldstein. The use of mmr and diversity-based reranking for reodering documents and producing summaries. 1998.

[20] Elizabeth S Chen, George Hripcsak, Hua Xu, Marianthi Markatou, and Carol Friedman. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association*, 15(1):87–98, 2008.

[21] George H Chen, Devavrat Shah, et al. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends® in Machine Learning*, 10(5-6):337–588, 2018.

[22] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM, 2008.

[23] Preciosa M. Coloma, Gianluca Trifirò, Vaishali Patadia, and Miriam Sturkenboom. Postmarketing safety surveillance. *Drug Safety*, 36(3):183–197, February 2013.

[24] John P Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research*, 16(1):2859–2900, 2015.

[25] Tran-Thai Dang and Tu-Bao Ho. Mixture of language models utilization in score-based sentiment classification on clinical narratives. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 255–268. Springer, 2016.

[26] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[27] Thomas G Dietterich et al. Ensemble learning. *The handbook of brain theory and neural networks*, 2:110–125, 2002.

[28] Yijie Ding, Jijun Tang, and Fei Guo. Identification of drug-side effect association via semi-supervised model and multiple kernel learning. *IEEE journal of biomedical and health informatics*, 2018.

[29] Martin J Doherty. Algorithms for assessing the probability of an adverse drug reaction. *Respiratory Medicine CME*, 2(2):63–67, 2009.

[30] MaulikS Doshi, PrakrutiP Patel, SamidhP Shah, and RamK Dikshit. Intensive monitoring of adverse drug reactions in hospitalized patients of two medical units at a tertiary care teaching hospital. *Journal of Pharmacology and Pharmacotherapeutics*, 3(4):308, 2012.

[31] Ian Ford and John Norrie. Pragmatic trials. *New England Journal of Medicine*, 375(5):454–463, August 2016.

[32] Shayne Cox Gad. *Drug Safety Evaluation*. John Wiley & Sons, Inc., October 2016.

[33] Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(Jul):2001–2049, 2010.

[34] Negar Golchin, Lisa Isham, SharonB Meropol, April Vince, and ScottH Frank. Polypharmacy in the elderly. *Journal of Research in Pharmacy Practice*, 4(2):85, 2015.

[35] Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Ruslan R Salakhutdinov. Neighbourhood components analysis. In *Advances in neural information processing systems*, pages 513–520, 2005.

[36] Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. *arXiv preprint arXiv:1807.01477*, 2018.

[37] Sander Greenland, James M Robins, Judea Pearl, et al. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999.

[38] Emily R. Hajjar, Angela C. Cafiero, and Joseph T. Hanlon. Polypharmacy in elderly patients. *The American Journal of Geriatric Pharmacotherapy*, 5(4):345–351, December 2007.

[39] Joseph T. Hanlon, Richard J. Sloane, Carl F. Pieper, and Kenneth E. Schmader. Association of adverse drug reactions with drug–drug and drug–disease interactions in frail older outpatients. *Age and Ageing*, 40(2):274–277, December 2010.

[40] L. Härmark and A. C. van Grootheest. Pharmacovigilance: methods, recent developments and future perspectives. *European Journal of Clinical Pharmacology*, 64(8):743–752, June 2008.

[41] Rave Harpaz, Krystl Haerian, Herbert S Chase, and Carol Friedman. Statistical mining of potential drug interaction adverse effects in fda's spontaneous reporting system. In *AMIA Annual Symposium Proceedings*, volume 2010, page 281. American Medical Informatics Association, 2010.

[42] Yukio Hinatsu. Diverse structures of mixed-metal oxides containing rare earths and their magnetic properties. *Journal of the ceramic society of Japan*, 123(1441):845–852, 2015.

[43] Tu-Bao Ho, Ly Le, Dang Tran Thai, and Siriwon Taewijit. Data-driven approach to detect and predict adverse drug reactions. *Current Pharmaceutical Design*, 22(23):3498–3526, June 2016.

[44] Hui Huang, Ping Zhang, Xiaoyan A. Qu, Philippe Sanseau, and Lun Yang. Systematic prediction of drug combinations based on clinical side-effects. *Scientific Reports*, 4(1), November 2014.

[45] SUG Hyontai. Performance of machine learning algorithms and diversity in data. In *MATEC Web of Conferences*, volume 210, page 04019. EDP Sciences, 2018.

[46] Pedro Inácio, Afonso Cavaco, and Marja Airaksinen. The value of patient reporting to the pharmacovigilance system: a systematic review. *British Journal of Clinical Pharmacology*, 83(2):227–246, October 2016.

[47] Olexandr Isayev, Denis Fourches, Eugene N Muratov, Corey Oses, Kevin Rasch, Alexander Tropsha, and Stefano Curtarolo. Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chemistry of Materials*, 27(3):735–743, 2015.

[48] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013.

[49] Yanqing Ji, Hao Ying, Peter Dews, Ayman Mansour, John Tran, Richard E Miller, and R Michael Massanari. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Transactions on Information Technology in Biomedicine*, 15(3):428–437, 2011.

[50] Yanqing Ji, Hao Ying, Peter Dews, John Tran, Ayman Mansour, Richard E Miller, and R Michael Massanari. An exclusive causal-leverage measure for detecting adverse drug reactions from electronic medical records. In *2011 Annual Meeting of the North American Fuzzy Information Processing Society*, pages 1–6. IEEE, 2011.

[51] Huidong Jin, Jie Chen, Hongxing He, Chris Kelman, Damien McAullay, and Christine M O'Keefe. Signaling potential adverse drug reactions from administrative health databases. *IEEE Transactions on knowledge and data engineering*, 22(6):839–853, 2010.

[52] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[53] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1:15010, 2015.

[54] Yvonne Koh, Chun Wei Yap, and Shu Chuen Li. A quantitative approach of using genetic algorithm in designing a probability scoring system of an adverse drug reaction assessment system. *International journal of medical informatics*, 77(6):421–430, 2008.

[55] Nicole Krämer and Mikio L Braun. Kernelizing pls, degrees of freedom, and efficient model selection. pages 441–448, 2007.

[56] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079, October 2015.

[57] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

[58] AV Kupriyanov and DV Kirsh. Estimation of the crystal lattice similarity measure by three-dimensional coordinates of lattice nodes. *Optical Memory and Neural Networks*, 24(2):145–151, 2015.

[59] Tien Lam Pham, Hiori Kino, Kiyoyuki Terakura, Takashi Miyake, Koji Tsuda, Ichigaku Takigawa, and Hieu Chi Dam. Machine learning reveals orbital interaction in materials. *Science and technology of advanced materials*, 18(1):756–765, 2017.

[60] Amanda Hanora Lavan and Paul Gallagher. Predicting risk of adverse drug reactions in older adults. *Therapeutic Advances in Drug Safety*, 7(1):11–22, November 2015.

[61] Ching-Hua Lin, Chao-Chan Kuo, Li-Shiu Chou, Yeng-Hung Chen, Cheng-Chung Chen, Kuo-Hao Huang, and Hsien-Yuan Lane. A randomized, double-blind comparison of risperidone versus low-dose risperidone plus low-dose haloperidol in treating schizophrenia. *Journal of Clinical Psychopharmacology*, 30(5):518–525, October 2010.

[62] Ching-Hua Lin, Fu-Chiang Wang, Shih-Chi Lin, Yu-Hui Huang, Cheng-Chung Chen, and Hsien-Yuan Lane. Antipsychotic combination using low-dose antipsychotics is as efficacious and safe as, but cheaper, than optimal-dose monotherapy in the treatment of schizophrenia. *International Clinical Psychopharmacology*, 28(5):267–274, September 2013.

[63] Wen-Yan Lin, Siying Liu, Jian-Huang Lai, and Yasuyuki Matsushita. Dimensionality's blessing: Clustering images by underlying distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5784–5793, 2018.

[64] Jingjing Liu, Alice Li, and Stephanie Seneff. Automatic drug side effect discovery from online patient-submitted reviews: Focus on statin drugs. In *Proceedings of First International Conference on Advances in Information Mining and Management (IMMM), Barcelona, Spain*, pages 23–29. Citeseer, 2011.

[65] Mong-Liang Lu, Hsien-Yuan Lane, Kun-Po Chen, Michael W. Jann, Muh-Hwan Su, and Wen-Ho Chang. Fluvoxamine reduces the clozapine dosage needed in refractory schizophrenic patients. *The Journal of Clinical Psychiatry*, 61(8):594–599, August 2000.

[66] Mong-Liang Lu, Hsien-Yuan Lane, Shih-Ku Lin, Kun-Po Chen, and Wen-Ho Chang. Adjunctive fluvoxamine inhibits clozapine-related weight gain and metabolic disturbances. *The Journal of Clinical Psychiatry*, 65(6):766–771, June 2004.

[67] David W Lynch and RD Cowan. Effect of hybridization on 4d 4f spectra in light lanthanides. *Physical Review B*, 36(17):9228, 1987.

[68] A. F. Macedo, F. B. Marques, C. F. Ribeiro, and F. Teixeira. Causality assessment of adverse drug reactions: comparison of the results obtained from published decisional algorithms and from the evaluations of an expert panel, according to different levels of imputability. *Journal of Clinical Pharmacy and Therapeutics*, 28(2):137–143, April 2003.

[69] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jurgen Bajorath. Molecular similarity in medicinal chemistry: miniperspective. *Journal of medicinal chemistry*, 57(8):3186–3204, 2013.

[70] Gerald M Maggiora and Veerabahu Shanmugasundaram. Molecular similarity measures. pages 1–50, 2004.

[71] Robert L Maher, Joseph Hanlon, and Emily R Hajjar. Clinical consequences of polypharmacy in elderly. *Expert Opinion on Drug Safety*, 13(1):57–65, September 2013.

[72] Ana G Maldonado, JP Doucet, Michel Petitjean, and Bo-Tao Fan. Molecular similarity and diversity in chemoinformatics: from theory to applications. *Molecular diversity*, 10(1):39–79, 2006.

[73] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[74] Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V Lilienfeld, and Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. In *Advances in Neural Information Processing Systems*, pages 440–448, 2012.

[75] Thomas J. Moore and Curt D. Furberg. Electronic health data for postmarket surveillance: A vision not realized. *Drug Safety*, 38(7):601–610, May 2015.

[76] Sean A Munson, Daniel Xiaodan Zhou, and Paul Resnick. Sidelines: An algorithm for increasing diversity in news and opinion aggregators. In *Third International AAAI Conference on Weblogs and Social Media*, 2009.

[77] Hiroki Murayama, Mio Sakuma, Yuri Takahashi, and Takeshi Morimoto. Improving the assessment of adverse drug reactions using the naranjo algorithm in daily practice: The japan adverse drug events study. *Pharmacology Research & Perspectives*, 6(1):e00373, January 2018.

[78] Cláudio A Naranjo, Usoa Busto, Edward M Sellers, P Sandor, I Ruiz, EA Roberts, E Janecek, C Domecq, and DJ Greenblatt. A method for estimating the probability of adverse drug reactions. *Clinical Pharmacology & Therapeutics*, 30(2):239–245, 1981.

[79] G Niklas Norén, Johan Hopstadius, Andrew Bate, Kristina Star, and I Ralph Edwards. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery*, 20(3):361–387, 2010.

[80] 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, 2018.

[81] H.S. Rehan, Deepti Chopra, and Ashish Kumar Kakkar. Physicians guide to pharmacovigilance: Terminology and causality assessment. *European Journal of Internal Medicine*, 20(1):3–8, January 2009.

[82] Jenna Reps, Jonathan M Garibaldi, Uwe Aickelin, Daniele Soria, Jack E Gibson, and Richard B Hubbard. Comparing data-mining algorithms developed for longitudinal observational databases. In *2012 12th UK Workshop on Computational Intelligence (UKCI)*, pages 1–8. IEEE, 2012.

[83] Jenna Reps, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack E. Gibson, and Richard B. Hubbard. Attributes for causal inference in electronic healthcare databases. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. IEEE, June 2013.

[84] Jenna M. Reps and Uwe Aickelin. Incorporating spontaneous reporting system data to aid causal inference in longitudinal healthcare data. In *2014 IEEE International Conference on Data Mining Workshop*. IEEE, December 2014.

[85] Jenna Marie Reps, Jonathan M. Garibaldi, Uwe Aickelin, Jack E. Gibson, and Richard B. Hubbard. A supervised adverse drug reaction signalling framework imitating bradford hill's causality considerations. *Journal of Biomedical Informatics*, 56:356–368, August 2015.

[86] Sereina Riniker and Gregory A Landrum. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of cheminformatics*, 5(1):43, 2013.

[87] Pedro Pereira Rodrigues, Daniela Ferreira-Santos, Ana Silva, Jorge Polónia, and Inês Ribeiro-Vaz. Implementing guidelines for causality assessment of adverse drug reaction reports: A bayesian network approach. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 55–64. Springer, 2017.

[88] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[89] James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65(11):1501–1509, 2013.

[90] Bhavik M. Shah and Emily R. Hajjar. Polypharmacy, adverse drug reactions, and geriatric syndromes. *Clinics in Geriatric Medicine*, 28(2):173–186, May 2012.

[91] Lei Shi and Yi-Dong Shen. Diversifying convex transductive experimental design for active learning. In *IJCAI*, pages 1997–2003, 2016.

[92] Andrea Skelly, Joseph Dettori, and Erika Brodt. Assessing bias: the importance of considering confounding. *Evidence-Based Spine-Care Journal*, 3(01):9–12, February 2012.

[93] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[94] T.-P. v. Staa, B. Goldacre, M. Gulliford, J. Cassell, M. Pirmohamed, A. Taweel, B. Delaney, and L. Smeeth. Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ*, 344(feb07 1):e55–e55, February 2012.

[95] Fabiana Rossi Varallo, Cleopatra S. Planeta, Maria Teresa Herdeiro, and Patricia de Carvalho Mastroianni. Imputation of adverse drug reactions: Causality assessment in hospitals. *PLOS ONE*, 12(2):e0171470, February 2017.

[96] Fei Wang and Changshui Zhang. Feature extraction by maximizing the average neighborhood margin. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[97] Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337, 2009.

[98] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[99] Peter Willett. The calculation of molecular structural similarity: principles and practice. *Molecular informatics*, 33(6-7):403–413, 2014.

[100] David H Wolpert, William G Macready, et al. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

[101] Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003.

[102] Zhengwei Yang, Ada Wai-Chee Fu, and Ruifeng Liu. Diversified top-k subgraph querying in a large graph. In *Proceedings of the 2016 International Conference on Management of Data*, pages 1167–1182. ACM, 2016.

[103] Yiming Ying and Peng Li. Distance metric learning with eigenvalue optimization. *Journal of machine Learning research*, 13(Jan):1–26, 2012.

[104] Xinge You, Ruxin Wang, and Dacheng Tao. Diverse expected gradient active learning for relative attributes. *IEEE Transactions on Image Processing*, 23(7):3203–3217, 2014.

[105] Cong Yu, Laks Lakshmanan, and Sihem Amer-Yahia. It takes variety to make a world: diversification in recommender systems. In *Proceedings of the 12th international conference on extending database technology: Advances in database technology*, pages 368–378. ACM, 2009.

[106] Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

[107] Wen Zhang, Xiang Yue, Weiran Lin, Wenjian Wu, Ruoqi Liu, Feng Huang, and Feng Liu. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC bioinformatics*, 19(1):233, 2018.

[108] Ping Zhong, Zhiqiang Gong, Shutao Li, and Carola-Bibiane Schönlieb. Learning to diversify deep belief networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3516–3530, 2017.

[109] Li Zhu, Maximilian Amsler, Tobias Fuhrer, Bastian Schaefer, Somayeh Faraji, Samare Rostami, S Alireza Ghasemi, Ali Sadeghi, Migle Grauzinyte, Chris Wolverton, et al. A fingerprint based metric for measuring similarities of crystalline structures. *The Journal of chemical physics*, 144(3):034203, 2016.

[110] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32. ACM, 2005.

[111] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

[112] Ivan Zorych, David Madigan, Patrick Ryan, and Andrew Bate. Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Statistical methods in medical research*, 22(1):39–56, 2013.

[113] Mira G.P. Zuidgeest, Iris Goetz, Rolf H.H. Groenwold, Elaine Irving, Ghislaine J.M.W. van Thiel, and Diederick E. Grobbee. Series: Pragmatic trials and real world evidence: Paper 1. introduction. *Journal of Clinical Epidemiology*, 88:7–13, August 2017.

Publications

1. Dang TT, Ouankhamchan P, Ho TB. Detection of new drug indications from electronic medical records. In 2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF) 2016 Nov 7 (pp. 223-228). IEEE. DOI: 10.1109/RIVF.2016.7800298.

2. Dang TT, Ho TB. Sequence-based measure for assessing drug-side effect causal relation from electronic medical records. In International Symposium on Knowledge and Systems Sciences 2017 Nov 17 (pp. 53-65). Springer, Singapore. DOI: https://doi.org/10.1007/978-981-10-6989-5_5.

3. Tran-Thai Dang, Tien-Lam Pham, Hieu-Chi Dam. Measuring dissimilarity between materials with an emphasis on identity of atomic orbital hybridization. Asian Consortium on Computational Materials Science Theme Meeting on "Multi-scale Modelling of Materials for Sustainable Development" (ACCMS-TM 2018) 7th to 9th, Hanoi, Vietnam. (poster)

4. Tran-Thai Dang, Thanh-Hang Nguyen, and Tu-Bao Ho. Causality assessment of adverse drug reaction: controlling confounding induced by polypharmacy. Current Pharmaceutical Design, 2019. DOI: 10.2174/1381612825666190416115714

5. Tran-Thai Dang, Tien-Lam Pham, Hiori Kino, Takashi Miyake, and Hieu-Chi Dam. Measuring the Similarity between Materials with an Emphasis on the Material's Distinctiveness. Science and Technology of Advanced Materials, 2019. (**submitted**)