공학박사학위논문

Molecular Property Prediction with Deep Learning

딥러닝 기반의 분자 특성 예측 연구

2021년 8월

서울대학교 대학원

협동과정 생물정보학

조 정 희

Molecular Property Prediction with Deep Learning

지도 교수 윤 성 로

이 논문을 공학박사 학위논문으로 제출함
2021 년 8 월

서울대학교 대학원
협동과정 생물정보학
조 정 희

조정희의 박사 학위논문을 인준함
2021 년 8 월

위 원 장 ____________________

부위원장 ____________________

위 원 ____________________

위 원 ____________________

위 원 ____________________

Abstract

Deep learning (DL) has been advanced in various fields, such as vision tasks, language processing, and natural sciences. Recently, several remarkable researches in computational chemistry were accomplished by DL-based methods. However, the chemical system consists of diverse elements and their interactions. As a result, it is not trivial to predict chemical properties which are determined by intrinsically complicated factors. Consequently, conventional approaches usually depend on tremendous calculations for chemical simulations or predictions, which are cost-intensive and time-consuming.

To address recent issues, we studied deep learning for computational chemistry. We focused on the chemical property prediction from molecular structure representations. A molecular structure is a complex of atoms and their arrangements. The molecular property is determined by the interactions from all these components. Therefore, molecular structural representations are the key factor in the chemical property prediction tasks. In particular, we explored public property prediction tasks in pharmacology, organic chemistry, and quantum chemistry. Molecular structures can be described as categorical sequences or geometric graphs. We utilized both representational formats for prediction tasks, and achieved competitive model performances. Our studies verified that the molecular representation is essential for various tasks in chemistry, and using appropriate types of neural networks for the representation type is significant to the model predictability.

Keywords: deep learning, molecular property prediction, graph neural networks

Student ID: 2016-30127

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

From early days, chemical modeling has been widely developed in various fields related with computational chemistry, biology, pharmacology, and others (Manly et al. 2001; Tropsha 2010; Cherkasov et al. 2014; Butler et al. 2018). By virtue of the rapid growth of computational resources, the chemical modeling methods have played an important role in more advanced research such as development of a drug or novel material (Sliwoski et al. 2014; Ward et al. 2016; Chen et al. 2018).

In chemical modeling research, the predictability of chemical properties is the most critical factor regardless of application fields. Recently, various researches such as the discovery of novel chemicals (Agatonovic-Kustrin and Beresford 2000; Ekins et al. 2019; Rifaioglu et al. 2019), repurposing of existing drugs (Zeng et al. 2019; Issa et al. 2021), and dynamic simulation (Ungerer et al. 2007; Wang and Hou 2011; Li et al. 2014) have utilized the predicted chemical properties in common.

The chemical properties are varied and correlated with each other. Most of the properties of a molecule are closely related with the different energy levels of local parts. These energies depend on the force fields from their corresponding local conformation (Maffucci and Contini 2015; Mieres-Perez and Sanchez-Garcia 2020). Therefore, the molecular properties can be induced by the atom types and geometry of a molecule (Cohen 1996; Bender and Glen 2004; Katritzky et al. 2010). In other words, the precise molecular structure representation is significant in the molecular property prediction tasks (Kozma et al. 2000; Schneider and Baringhaus 2008; Ferreira et al. 2015; Staker et al. 2020).

However, it is not trivial to make an accurate prediction of chemical properties even in small-sized molecules. As aforementioned, the molecular characteristics are determined by intricate relationships with multiple components. Conventional approaches such as density functional theory (DFT) require tremendous calculations to overcome the problem (Ruddigkeit et al. 2012; Montavon et al. 2013; Ramakrishnan et al. 2014; Chmiela et al. 2017, 2018).

To address the issue, several deep learning (DL)-based methods have been proposed for chemical modeling tasks. Several models (Hirohara et al. 2018; Shin et al. 2019; Zheng et al. 2019; Huang et al. 2020) used sequential molecular specification, which is called a simplified molecular-input line-entry system (SMILES) (Weininger 1988a). This format consists of discrete characters for atom charges and structural indications. Most of the previous works using SMILES (Goh et al. 2017; Honda et al. 2019; Wang et al. 2019a; Li and Fourches 2021) applied the methods for languages, because both SMILES and languages are sequentially represented. More recently, molecular graph representations are widely adopted (Duvenaud et al. 2015; Kearnes et al. 2016; Wang et al. 2019b), to represent molecular conformation at a lower level. The graph representation has an advantage over a sequential representation of molecules, because it can represent diverse conformations of a one molecule, while the sequential representations cannot.

Among several types of GNNs, a message passing neural network (MPNN) (Gilmer et al. 2017) is one of the most successful architectures for tasks on molecular graphs. In MPNNs, messages are created from each atom, and they are iteratively updated with neighboring environments. All messages at the final step are aggregated to output the predicted properties of a whole molecule. The intuition behind message passing is that the energy of a whole molecule is equal to the sum of the energy of each atom in a molecule. Many architectures that adopted MPNN as a base framework showed noteworthy model predictability.

Simultaneously, several public databases and standard benchmarks for various tasks were created. Tox21 (Mayr et al. 2016; Huang et al. 2016) is a challenge held in 2014 for toxicity predictions of target chemicals. RDKit (Landrum 2019), DeepChem (Ramsundar et al. 2019), and MoleculeNet (Wu et al. 2018) provide various molecular property prediction tasks on biophysics, physical chemistry, and quantum mechanics.

1.2 Contents of dissertation

The contents of this dissertation are as follows. In the front part, we briefly introduce our motivation and the history of related researches in Introduction in section 1. In the next Chapter 2, we first discuss the previous works of deep learning-based approaches in Chemistry. In particular, various computational researches related with chemicals, limitations of the conventional rule-based approaches, and recent success of DL-based works in computational chemistry. Next, we narrow down to specific issues on the molecular property predictions, which is the most widely discussed and practical problem in computational Chemistry. The last section in Background part 2, we divided the problems by specific subjects: pharmacology, biophysical, physiological, and quantum-mechanical properties prediction.

The following three Chapters are the application studies on different subjects. In Chapter 3, we introduce our study on drug classification problem with pairs of drug molecules and their corresponding target amino acids. The task is predicting therapeutic class of drugs with corresponding chemical representations and their target amino acid sequences. We extracted drug categories, drug molecular string and amino acid sequence data from DrugBank (Law et al. 2013). We adopted two types of deep neural networks: one-dimensional convolutional neural networks (1DCNN) and bidirectional long short-term memory networks (BiLSTM).

We evaluated our model with classification accuracy, area under the receiver operating characteristic curve (AUC-ROC). We also represented confucion matrices for each prediction class. We compared our results with existing methods and baseline models. Our study suggested that with molecular string representation and its binding target, pharmacological properties can be predicted without hand-craft features or complex preprocessings. This Chapter is based on the following research:

- **Jeonghee Jo**, Hyun-Soo Choi, and Sungroh Yoon. "Prediction of Drug Classes with a Deep Neural Network using Drug Targets and Chemical Structure Data." 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019.

In Chapter 4, we addressed biophysical and physiological molecular properties on different tasks. This task consists of multiple binary classifications on small molecules. The specific targets are related with drug actions: binding ability to specific protein, toxicity, permeability, and others. We used several datasets from the public benchmark (Wu et al. 2018). We used molecule strings predict all properties.

We assumed that molecular strings were analogous to sentences, so we adopted language-based preprocessing and training strategy. Based on this, we utilized following approaches: co-occurrence graph (Mihalcea and Tarau 2004), byte-pair encoding (Sennrich et al. 2016), and Word2Vec (Mikolov et al. 2013). We also adopted previously proposed architecture (Nikolentzos et al. 2019) for the word graphs. In these graphs, the words are the characters or byte-pair encodings from molecular strings.

We analyzed our model on various subtasks with different experimental settings. We found the best setting for each dataset, and concluded that the best performances of each dataset could be obtained from slightly different model parameters. Our model outperforms the previous models and baseline models in most of the tasks. To the best of our knowledge, it is the first research to train co-occurrence graphs on molecular string data, with graph neural networks. Our study showed that the molecular strings can be considered as a type of language, so many of the language-based approaches can be applied in other problems in Chemistry. This Chapter is involved in the following research:

- **Jeonghee Jo**, Bumju Kwak, Hyun-Soo Choi, Sungroh Yoon, "The message passing neural networks for chemical property prediction on SMILES." Methods 179 (2020): 65-72.

In Chapter 5, we explored more fundamental problems in Chemistry. We studied on quantum-mechanical properties prediction tasks on molecules, which are most significant factors in various chemical applications. Conventional approaches based on density functional theory or nuclear magnetic resonance require tremendous computations, which are time-consuming and cost-expensive. Therefore, it is the most widely studied field in chemistry in the latest days.

We used molecular graphs for representations different from two previous studies. The quantum-mechanical properties are formulated by consisting atom particles and their interactions, we considered graph formats are more suitable and reasonable for these tasks. We retrieved data from the public benchmarks for quantum-chemistry, QM9 (Ramakrishnan et al. 2014) and Molecular Dynamics (Chmiela et al. 2017, 2018). These datasets consist of 12 and 5 different regression tasks, respectively. The targets are continuous level of energies and other observed properties.

For molecular graphs, we adopted the message passing-based approach, which is a type of graph convolutional neural networks. The strategy of message passing is that each atom representation is updated based on its neighboring atoms and the representations from previous time step (layer). It is closely related with the basic concept of molecular energy calculation.

We assumed that each type of quantum-mechanical property has different relationship with molecular attributes. Furthermore, we considered that message-passing framework has a risk of over-mixing atom representations, which can cause over-smoothing problem in deep graph neural networks (Li et al. 2018). Therefore, we add two types of novel operations to message passing neural networks: (1) for atom-wise multilayer perceptron (MLP) based pathway and (2) trainable scalars for differentiating relative importances on heterogeneous molecular attributes.

Our experimental results showed remarkable performances on two benchmark datasets. These results validated our assumptions that heterogenous features contribute to differently for each target. We did ablation studies for various experimental setting, which is not shown in the preprint. This Chapter is based on the following research in preprinting:

- **Jeonghee Jo**, Bumju Kwak, Byunghan Lee, Sungroh Yoon. "Flexible dual-branched message passing neural network for quantum mechanical property prediction with molecular conformation." arXiv preprint arXiv:2106.07273 (2021).

In Chapter 6, we discuss overall contents of this dissertation. In particular, we briefly summarize our experiments, and discuss the significance and limitations of our approaches with respect to performances and generalizability. We also suggest the future study to overcome current limitations of our studies.

2 Background

2.1 Deep learning in Chemistry

For several years, machine learning (ML) and deep learning (DL)-based research has increased rapidly in both natural and applied sciences (Angermueller et al. 2016; Min et al. 2017; Dimiduk et al. 2018; Cova and Pais 2019). In general, prediction and simulation analyses in biology, chemistry, or physics require cost-intensive computations (Michel et al. 2010; Najafabadi et al. 2015). In some cases, target solutions cannot be solved because of intractable formulation or unknown factors in the process of calculations.

To overcome the limitations of conventional methodologies, many researchers adopted learning-based approaches for various scientific fields. In particular, molecules are fundamental subjects in most of the fields. Molecules are extremely diverse, interact with each other, and have complex properties. Consequently, many DL-based models have been applied to research on molecules (Lorenz et al. 2004; Behler and Parrinello 2007; Smith et al. 2017; Zhang et al. 2018).

In particular, ML and DL methods have played a significant role in unknown property prediction tasks from known molecular structure data. For example, molecular structure analysis is essential for development of a novel drug or material (Chen et al. 2018; Lavecchia 2019; Rifaioglu et al. 2019), repurposing of existing drugs (Moridi et al. 2019; Zeng et al. 2019), medicinal effect or toxicity prediction (Mayr et al. 2016; Karim et al. 2019; Benning et al. 2020), or simulation in special environments (DeFever et al. 2019; Noé et al. 2020; Doerr et al. 2021).

There are many types of molecular structure representation. These representation types can be divided into two main categories. One is sequence-based and the other is graph-based format. Examples of sequential representations are a simplified molecular-input line-entry system (SMILES) (Anderson et al. 1987), international chemical identifier (InChI) (Heller et al. 2013), morgan fingerprints (Glen et al. 2006), and others (Gobbi and Poppinger 1998; Rogers and Hahn 2010). These types of representations should contain structural characteristics such as bond types. Graph-formatted representations have two types of components; atom types and their corresponding positions. Bond type or other structural information is not necessary in this format, because it can be estimated by the distance between any two atoms with their atom types.

We will briefly discuss previous approaches according to molecular representation types and corresponding DL frameworks.

2.2 Deep Learning for molecular property prediction

A molecule consists of multiple atoms located in the unique arrangement. Each atom interacts with each other based on their spatial relationships within a molecule. If two atoms are engaged in a close distance, they will share electrons to minimize their energy. The entire molecular structure is shaped by the sum of all local energies to be minimized.

Therefore, most of the molecular properties including reactivity, solubility, and energies have been analyzed based on its structural characteristics from the early days. Density functional theory (DFT) and nuclear magnetic resonance (NMR) are well known conventional approaches for analyzing molecular structures. However, exact molecular structure determination is still challenging because atomic interactions are complicated and hard to formulate (Reichenbächer and Popp 2012). Moreover, the conventional approaches require cost-intensive and time-consuming calculations (Gauss 1995; Peterson et al. 2012).

To relieve the problem, various deep neural networks have been proposed. In early stages, many types of molecular features such as bond types, valences, and aromaticities are used (Duvenaud et al. 2015; Ryu et al. 2018; Coley et al. 2019). Typically, one-dimensional convolutional neural networks (1DCNNs) (Hirohara et al. 2018; Wang et al. 2019b), long short-term memory networks (LSTMs) (McCarthy and Lee 2020), or models for languages have been adopted for this data (Goh et al. 2017; Honda et al. 2019; Wang et al. 2019a). These approaches were usually applied in property prediction tasks of known chemicals.

Afterwards, other approaches have arisen in this field, which use less types of molecular features. SchNet (Schütt et al. 2017, 2018) used only two types of features: atom types and atom pair distances, followed by PhysNet (Unke and Meuwly 2019) and DimeNet (Klicpera et al. 2020b). Note that molecular geometries such as distances or angles have continuous values. Therefore, categorical features such as type information are embedded as discrete vectors while geometric features are expanded by continuous functions. To handle geometric attributes, most of the models adopted graph neural networks (GNNs) (Kearnes et al. 2016; Liao et al. 2019; Wang et al. 2019b) or message passing neural networks (MPNNs) (Gilmer et al. 2017; Chen et al. 2019; Unke and Meuwly 2019; Tang et al. 2020) as their base frameworks. An MPNN is a type of graph convolutional neural networks (GCNNs), so it can also be included in the GNN category.

As mentioned in the previous section, a molecule can be described in two types of formats: a sequence and a graph. We adopt two types of molecular representations, one is a sequential format and the other is graph-based format. In particular, we used a simplified molecular-input line-entry system (SMILES) as sequential format, while a graph is constructed by rule-based preprocessing.

Each representation type has its own advantages. Sequential representations can indicate various categorical feature types such as bond types or aromaticity of a molecule. However, a sequential representation cannot preserve molecular geometry, because a sequence cannot handle continuous values and orders can be arbitrary. Meanwhile, a graph representation can preserve entire molecular geometry, however, it cannot keep non-geometric properties. In summary, a sequential representation is more suitable for high-level prediction tasks such as drug-drug interactions (DDI) or drug-target interactions (DTI). If the target property is low-level, such as thermodynamic or energy-related properties, a graph format would be better.

In this work, we used both of the molecular representation formats: SMILES and graphs. In particular, we used SMILES representation in drug classification task and biophysical property prediction task, and graphs in quantum-mechanics property prediction task. For these tasks, we adopted 1DCNN, BiLSTM, and two types of MPNN frameworks.

We briefly introduce three types of our DL-based models for chemical property prediction tasks. First, we adopted a one-dimensional convolutional neural network (CNN) and bidirectional long short-term memory networks (BiLSTM) for strings. Next, we obtained co-occurrence graphs from SMILES strings, and applied an MPNN to predict molecular properties. Finally, we applied other MPNN directly on molecular geometries. Note that two types of MPNN are not similar with each other; one is for abstract graphs from strings and the other is for geometric graphs in Euclidean space.

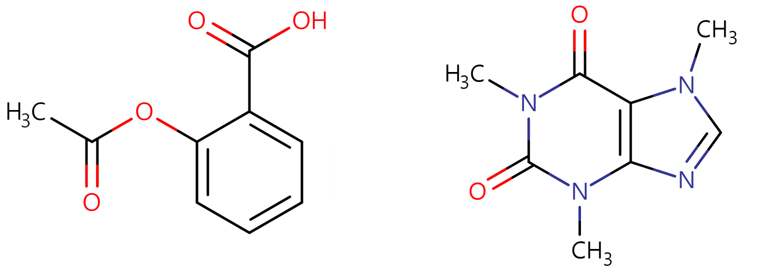
Our main contributions are as follows: We explored various problems and data types in chemical property prediction tasks in public benchmarks. We proposed novel architectures for each prediction task, and reached outstanding performances compared with previous or baseline models. Inspired by scientific developments of molecules, we implemented our methodology with the intuition behind scientific principles. We validated that our experimental results accorded with the known chemical principles.

2.3 Approaches for molecular property prediction

2.3.1 Sequential modeling for molecular string

We introduce the SMILES string, which is the most widely used among several types of strings for identifying molecules. SMILES is a specification for describing molecular structures with ASCII characters. It is one of the most widely used sequential formats for molecular structures, because it is more human-readable and specified than other sequential formats (O'Boyle 2012). The examples of SMILES are described in Figure 2.1.

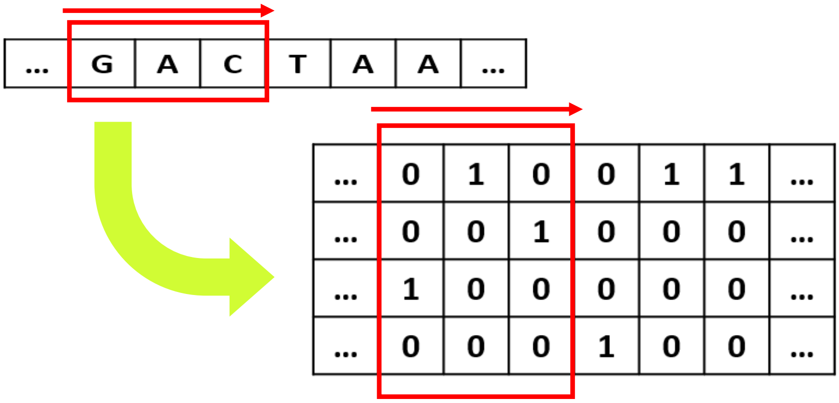
Previous works used SMILES data for several studies, such as drug-drug interaction (DDI) analysis Lee et al. 2019a; Huang et al. 2020, drug-target interaction (DTI) analysis Monteiro et al. 2020; Wang et al. 2020; Lee et al. 2019b, medicinal response Chang et al. 2018; Manica et al. 2019, toxicity Chen et al. 2020, or other biophysical properties prediction tasks Zheng et al. 2019; Hu et al. 2020; Tang et al. 2020; Kurotani et al. 2021. Several models Lee et al. 2019a; Manica et al. 2019; Huang et al. 2020 used encoder-decoder frameworks.



CC(=O)OC1=CC=CC=C1C(O)=O
Aspirin

CN1C=NC2=C1C(=O)N(C)C(=O)N2C
Caffeine

Figure 2.1 Examples of SMILES: Aspirin and Caffeine



...	G	A	C	T	A	A	...

...	0	1	0	0	1	1	...
...	0	0	1	0	0	0	...
...	1	0	0	0	0	0	...
...	0	0	0	1	0	0	...

Figure 2.2 Examples of one-dimensional convolutional neural networks for DNA sequences

One-dimensional convolutional neural networks (1DCNNs)

The most widely used model architecture for SMILES data is one-dimensional convolutional neural networks (1DCNNs). While standard two-dimensional convolutional filters move along two perpendicular axes, the one-dimensional convolutional filters move along only one axis. Therefore, 1DCNNs have been adopted for sequential data, such as time-series or text. The example of 1DCNN on characterized DNA sequences is represented in Figure 2.2.

Several works adopt a one-dimensional convolutional neural network (1DCNN) for sequential representations Chang et al. 2018; Lee et al. 2019b; Monteiro et al. 2020; Chen et al. 2020; Hu et al. 2020.

Long short-term memory (LSTM) networks

Long short-term memory (LSTM) networks is a type of recurrent neural networks (RNNs). LSTM was proposed to overcome the gradient vanishing problem of vanilla RNNs. An LSTM unit consists of a cell, an input gate, an

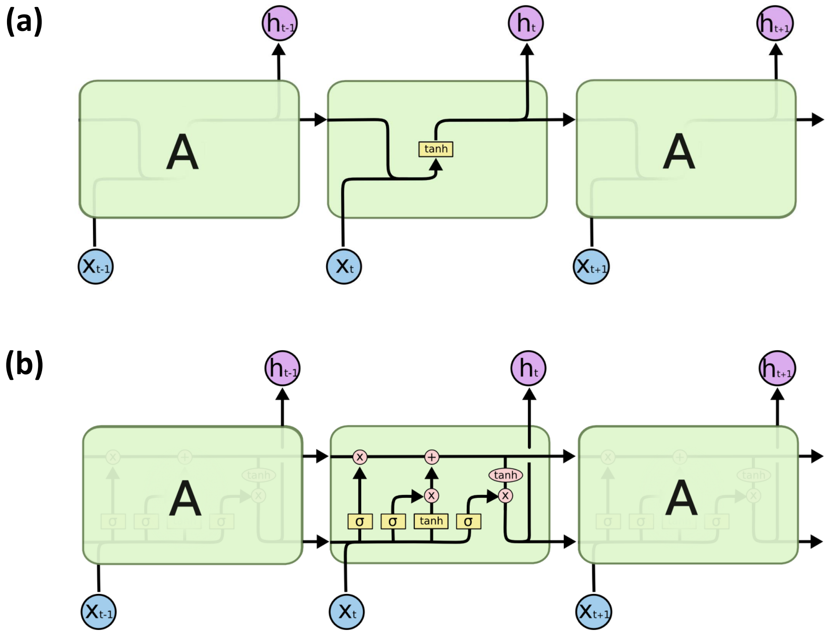


Figure 2.3 Comparison between (a) the standard RNN and (b) LSTM network (Olah 2015)

output gate, and a forget gate. These gates keep gradient flow consistently. The conceptual comparison between the standard RNN and LSTM is described in Figure 2.3.

Several studies Zheng et al. 2019; Wang et al. 2020 have adopted LSTM networks on DTI tasks with SMILES string.

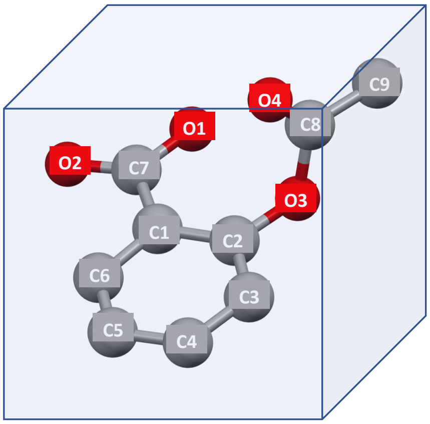
2.3.2 Structural modeling for molecular graph

Although many researches showed advanced performances, sequential format is somehow limited in terms of representative ability. First, it cannot handle deformed structure (in non-equilibrium state) or multiple forms of one molecule (isomers). Therefore, it is not possible for a simulation of molecular dynamics (MD) using sequential representations. Second, the diverse atom-atom interactions are too simplified in this format. In theory, intramolecular energies are the continuous function of atom-atom distances and their charges. Sequences are defined by discrete strings, so they cannot cover those continuous numerical properties. On the other hand, molecular graphs can represent intact three-dimensional structures. Therefore, they can preserve continuous properties obtained from geometry.

For these reasons, more recent works have utilized molecular geometries directly in their analyses. Most of the previous works using molecular geometries adopt a graph neural network (GNN) as their base framework. Duvenaud et al. 2015 adopted a graph convolution for end-to-end learning of molecular graphs. Message passing neural network (MPNN) (Gilmer et al. 2017) proposed a novel framework, which iteratively aggregates and updates the atomic message defined by its neighbors. SchNet (Schütt et al. 2017, 2018) proposed a continuous-valued filter (CF) for atom-atom distances, so called a radial basis. Afterwards, many researches such as MegNet (Chen et al. 2019), MCGN (Lu et al. 2019), PhysNet (Unke and Meuwly 2019), and DimeNet (Klicpera et al. 2020b) have explored combining CFs and MPNN. For more specific geometric details, several works additionally adopted angle features or roto-translational group equivariances (Thomas et al. 2018; Anderson et al. 2019; Klicpera et al. 2020b,a; Anderson et al. 2019; Fuchs et al. 2020).

Message passing neural networks (MPNNs)

Message passing neural networks (MPNNs) (Gilmer et al. 2017) aim for localized convolution on graphs. With this property, MPNNs can be regarded as spatial-based graph convolutional neural networks (GCNNs). In MPNNs, each node is individually represented, and initial messages are created from node embeddings with neighboring attributes. During the training stage, a block



	x	*y*	*z*
C1	1.636	3.690	0.755
C2	2.757	3.181	0.084
...			
O4	4.486	5.104	0.378

Figure 2.4 Example of molecular representation in three-dimensional space. Hydrogens are not shown for simplicity.

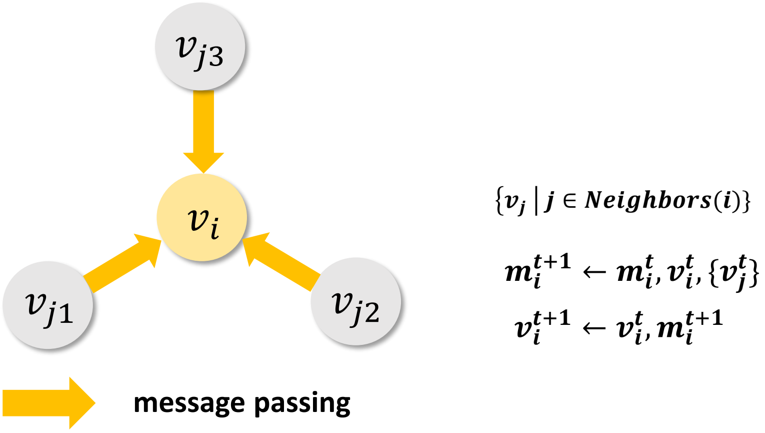

v_{j3}
v_i
v_{j1}
v_{j2}
$\{v_j \mid j \in Neighbors(i)\}$
$m_i^{t+1} \leftarrow m_i^t, v_i^t, \{v_j^t\}$
$v_i^{t+1} \leftarrow v_i^t, m_i^{t+1}$
message passing

Figure 2.5 The schema of a message passing process of a sample graph.

of message passing process consists of two types of consecutive functions: a *message passing* and an *update* function. Message passing function produces a *message*, which is a combined representation of each node and its neighbor's features. Messages and node features of the previous block are used to update current node features. Both of the message passing and update functions can be trainable layers or non-trainable operations such as a summation operator. The concise schema of massage passing process is described in 2.5.

$$m_i^{t+1} = M^t(h_i^t,\ h_{j,j\in N(i)}^t) \quad (t = 1, ..., T) \qquad \text{(message passing)} \tag{2.1}$$

$$h_i^{t+1} = U^t(h_i^t,\ m_i^{t+1}) \qquad \text{(update)} \tag{2.2}$$

M^t and U^t are the functions for message and update, respectively. h_i and m_i mean the hidden representation and message of node i, respectively. $N(i)$ means the neighboring nodes of node i, and t means the time step. The *readout* function followed by the last message passing block. It aggregates all messages at the final time step T, and yields the global representation of a graph G. Readout function R can be trainable or non-trainable, similar to message passing.

$$\hat{y} = R(i \in G \mid H^T) \qquad \text{(readout)} \tag{2.3}$$

Continuous filter convolution (CFC)

Continuous filter (CF) is a type of continuous function for scalar values, proposed in SchNet. Different from standard convolutions which operate matrix multiplication between feature maps and discrete filters, continuous filter convolutions (CFCs) do an element-wise operation between the features and the evaluations of radial basis functions of atom-atom distances. CF has an advantage over standard discrete convolution filters in that it is more suitable for features defined . A general form of continuous filters are described as

$$h_{ij} = CF(i, j \in G \mid D_{ij}) \qquad \text{(continuous filter)} \tag{2.4}$$

Radial basis functions embed the scalar-valued measures of molecular geometry to vectors. In general, radial basis functions are made up of different orthogonal functions. For example, SchNet adopted multiple Gaussian functions $g_{\mu_k,\sigma_k}(D_{ij})$ for embedding atom-atom distances. After SchNet, many studies utilized CFs to describe geometric measures of molecules. PhysNet (Unke and Meuwly 2019) used polynomial equations, MegNet (Chen et al. 2019) used *set2set* network (Vinyals et al. 2015) and DimeNet (Klicpera et al. 2020b) used Bessel functions.

2.4 Tasks on molecular properties

2.4.1 Pharmacological tasks

In pharmacology, DL-based methods have widely applied in developing novel drugs or repurposing existing drugs. In particular, DL-based methods can help explore the candidate chemicals in the initial process. Examples of these candidates are another drugs, proteins, or other biological subjects. In addition, medicinal efficacy or side effects can be prediction targets.

Various types of tasks are commonly involved in the interaction with other subjects. Therefore, most of the tasks are binary classifications: for predicting the presence of an interaction with the target. In general, these classification tasks are extremely imbalanced, because the molecular interactions are highly specific and most of pairs have no interaction with each other. In these cases, AUC-ROC or AUC-PRC are more recommended for estimating model performances. (Wu et al. 2018).

2.4.2 Biophysical and physiological tasks

DL methods can also be applied for biophysical or physiological properties of chemicals, which is important in developing synthetic materials as well as drug development. These tasks are generally based on regression problems, because the targets are continuous levels of properties, such as solubility, lipophilicity, permeability, and binding activity. Consequently, root mean-squared error (RMSE) or mean-squared error (MSE) are generally adopted as the evaluation methods of the model performances. Otherwise, if there are the criteria of target levels, then the problem is formulated as a binary classification.

2.4.3 Quantum-mechanical tasks

In the quantum chemical field, DL can predict quantum-mechanical properties of molecules more cost-effectively than conventional density functional theory (DFT) do. MD simulation based only on DFT is extremely time-consuming. DL can predict molecular properties without tremendous theoretical formulations more quickly. In the same with the biolphysical or physiological tasks, the quantum-mechanical properties usually consist of continuous numeric values. Related with the standard formula in quantum-chemical theories, the evaluation metric is usually a mean absolute error (MAE). In addition, if the prediction target is a type of energy, the loss function would be combined with additional loss terms for predicting forces. It is because the energy is a differential equation of forces with atom positions, by definition.

$$Loss(E, \hat{E}) = c_e \times |E - \hat{E}| + c_f \times \sum |F - \frac{\partial}{\partial \mathbf{r}} \hat{E}(\{\mathbf{r}\})| \tag{2.5}$$

The above equation is the commonly form of the loss functions in quantum-chemical property prediction tasks (Schütt et al. 2017, 2018; Chmiela et al. 2018; Klicpera et al. 2020b). E and $\hat{E}$ mean the true and predicted energy, respectively. The first term of the above equation is the MAE of energy prediction, and the second term is the MAE of forces prediction. r in the second term means positional attributes of each atom in a molecule. c_e and c_f are the coefficient for each loss term.

3 Chapter 1. Drug class classification with molecular string

3.1 Introduction

Drug actions are the combination of various biological and chemical reactions, and the drug class is determined by drug's target systems and environment. Thus the prediction of drug class is most significant in discovering property of drug candidates. Accurate and precise prediction method for drug class discovery can conserve time and costs in drug repositioning and de novo drug designs.

A drug molecule's structure have its own local site for binding its target proteins. Two different drugs having similar binding site would react with similar target proteins, and have similar mechanisms of action (MoA). Modern chemoinformatic researchers adopt this assumption and attempt to find and interpret local structure of specific target detection for drug designing and repositioning. Previous approaches including statistical model or machine learning models have no ability of selecting more important feature automatically, as we mentioned above. So manual feature selection methods, such as similarity or distance measuring were must required before main algorithm. However, deep learning can learn feature importance and complex itself, so it can reduce cumbersome data handling processes at the beginning. In addition, it allows computational models composed of multiple layers to learn representations of data with multiple levels of abstraction

Therefore, medicines can be categorized into different groups, according to the target organ or system on which they act and pharmacological and physiological properties. One of the drug classification methods is the anatomical therapeutic chemical (ATC) classification system, controlled by the World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC) (*WHO Collaborating Centre for Drug Statistics Methodology*). This system groups drugs into fourteen categories, which are coded by a single letter to one character as shown in Table 3.1. The ATC codes are closely related to a drug's indication and mechanism of action (MoA). Therefore, the model predictability is significant to novel drug design or repurposing.

Previous approaches based on statistical learning or ML-based models required many types of features (Chen et al. 2012; Chen and Jiang 2015; Wang et al. 2013; Liu et al. 2015). It is not desirable property, because it is not always possible to collect all types of features in practice. On the contrary, DL-based methods (Ertl et al. 2017b; Gupta et al. 2018-01; Lusci et al. 2013) can automatically learn relative importances of features, and their complex relations by virtue of multiple non-linear operations.

Recently, several ML-based models have been proposed to drug classification tasks using molecular structural and biochemical properties. Shortest path and random walk are adopted to predict ATC code in (Öztürk et al. 2016). In more recent days, DL-based methods have arisen for these pharmaceutical tasks. CNNs were used in several works (Ertl et al. 2017b; Gupta et al. 2018-01; Lusci et al. 2013; Graves and Schmidhuber 2005; Schuster and Paliwal 1997; Lipton 2015; Hochreiter and Schmidhuber 1997) for biochemical property predictions.

The other work (Wang et al. 2014) adopted CNN to predict ATC code, however, the predictability of this model could not exceed the performance from using a non DL-based algorithm. (Graves and Schmidhuber 2005) also applied CNN, while requiring a complex engineering and pretraining processes. We believed that these characteristics of deep learning is more proper and reasonable in biologically and pharmacologically for ATC code prediction task.

We briefly introduce two supervised learning methods used in this work. Convolutional neural network (CNN), a class of deep, feed-forward artificial neural networks (ANN), learns visual patterns directly from raw pixel images Paisitkriangkrai et al. 2015. CNN is mainly used for vision and natural language processing researches for capturing spatial patterns in restricted region of input data. Protein, gene, and chemical structures can be represented as linear sequence data as sentences, and thus the CNN model can also learn local features in these data. Recurrent neural network (RNN), also a kind of deep learning algorithms, can "memorize" their internal status and use it as past context to process sequences of inputs. It has widely applied in mainly speech recognition or text generation tasks. Amino acids and atoms also arranged in sequence, so they can be used an input of an RNN algorithm.

In this respect, it is expected that deep learning would learn local site of a drug molecule which plays a more important role in binding target protein without any idea of pharmacodynamics. We believed that these characteristics of deep learning is more proper and reasonable in biologically and pharmacologically for ATC code prediction task. So we constructed two kinds of deep learning models for predicting the drug's ATC code via drug chemical structure data and amino acid sequences of the drug's target genes. Both models predicted the ATC code of unseen data more accurately than when other machine learning methods were used.

Table 3.1 Anatomical therapeutic chemical (ATC) classification system

Code	Contents
A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dematologicals
G	Genito-urinary system and sex hormones
H	Systemic hormonal preparations, excluding sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

In this work, we constructed two types of neural networks: CNN and BiLSTM. We used two types of inputs: molecular string data of drugs and their corresponding target amino acid sequences. With these input pairs, we predict the therapeutic class of drugs, and the model showed significantly improved prediction accuracies compared to those of previous works.

3.2 Proposed method

The objective of this study is to predict ATC code of drugs with a DL-based model, using two types of input string data: drugs' molecular strings and corresponding target amino acid sequences. The intuition behind this work is that with the structural features of a molecule and their binding target, the model can predict the therapeutic or MoA of drugs. We applied two types of a neural network: a 1DCNN and a BiLSTM for both string-typed data.

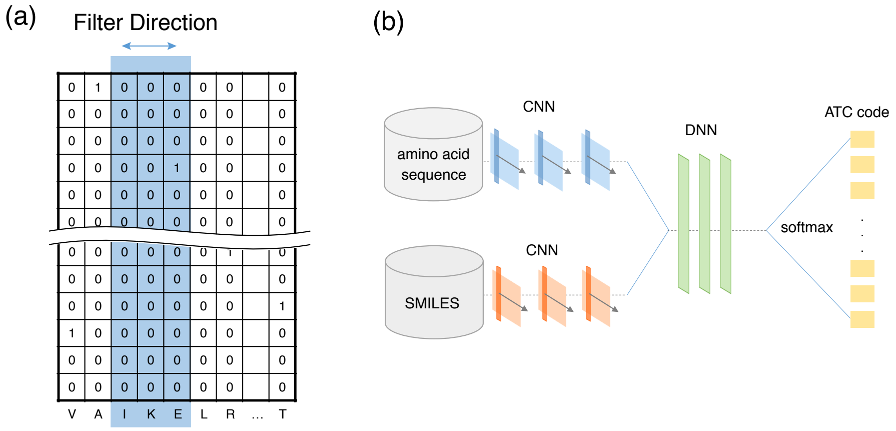


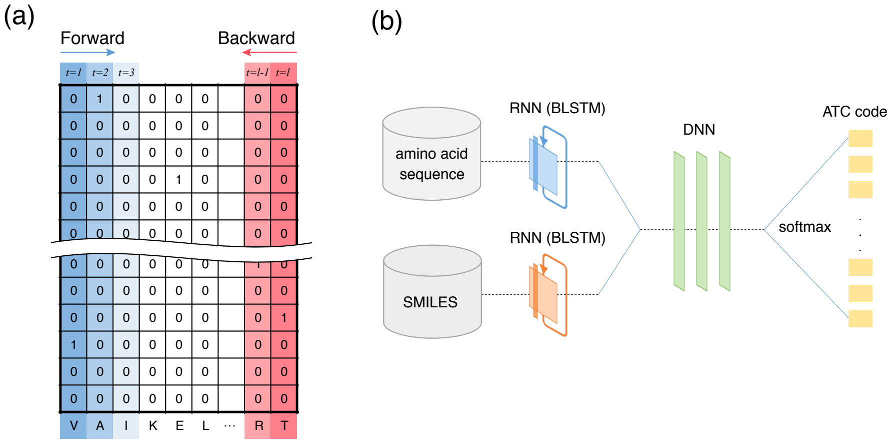
Figure 3.1 One-dimensional CNN-based architecture. (a) A one-hot encoded input example. (b) An architecture of the CNN-based model.

3.2.1 Preprocessing

We used one-hot encoding to embed SMILES strings and we fixed the length of input SMILES strings as L. Accordingly, N number of SMILES data was expressed by $N \times M \times L$ where M is the number of character types in the SMILES protocol. We also one-hot encoded target amino acid sequences, in the same manner with SMILES strings. We changed the notations for two-character chemical elements (e.g., Cl, Na, Pt) into arbitrary one character alphabets, not to overlap with another chemical element.

3.2.2 Model architecture

We applied two types of neural network for sequential data: a 1DCNN and a BiLSTM. For 1DCNN, we set the height of a convolutional filter to the same with M. The convolutional filters move only one direction. Then, each output from a convolution layer becomes a one-dimensional vector. In the latter part, we applied a fully connected (FC) layer to output predicted values. Details of two architectures are described in Figure 3.1 and 3.2, respectively.



t=1	*t=2*	*t=3*					*t=l-1*	*t=l*
0	1	0	0	0	0		0	0
0	0	0	0	0	0		0	0
0	0	0	0	0	0		0	0
0	0	0	0	1	0		0	0
0	0	0	0	0	0		0	0
0	0	0	0	0	0		0	0
0	0	0	0	0	0		1	0
0	0	0	0	0	0		0	0
0	0	0	0	0	0		0	1
1	0	0	0	0	0		0	0
0	0	0	0	0	0		0	0
0	0	0	0	0	0		0	0
V	A	I	K	E	L	···	R	T

Figure 3.2 One-dimensional BiLSTM-based architecture. (a) A one-hot encoded input example. (b) An architecture of the BiLSTM-based model.

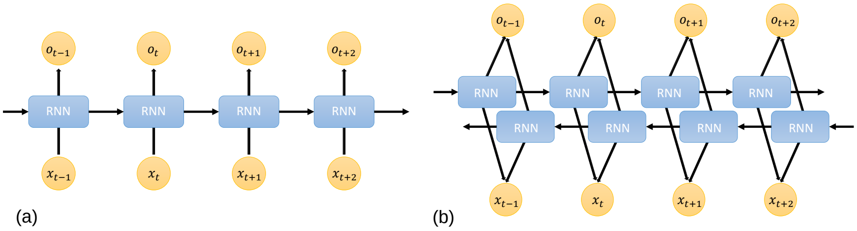


Figure 3.3 Two type of RNN architectures (a) standard RNN, and (b) bidirectional RNN.

The input shape of a CNN for a amino acid sequence is 100 × 20 where 100 is the length of an amino acid sequence and 20 is the dimension of one-hot encoded sequence representing 20 types of amino acids. The filter of the first layer has a shape of 3 × 20, so the filter moves in only one direction (Figure 3.1a). The other CNN for chemical structures has a 30×32 matrix as the input, 30 for the length of a SMILES data, and 32 for the number of discrete characters used in SMILES. Each output vector of the two CNN is a one-dimensional vector (Figure 3.2b). These are concatenated into a one one-dimensional vector. The deep neural network is made up of three layers and outputs with a 13 × 1 vector for ATC code classification via Softmax. A pair of a one-hot encoded amino acid sequence and a one-hot encoded SMILES from a drug is the base unit of both the training and the test dataset.

We adopted bidirectional Long Short-Term Memory (BiLSTM) (Graves et al. 2005) as an RNN cell, which is a variation of Bidirectional RNN (BRNN) (Graves and Jaitly 2014). In BRNN the neurons of a regular RNN into two directions, one for positive time direction, and another for negative time direction, shown in Figure 3.3. It often results with higher performance, as it can take into account both past and future sequence elements (Weininger 1988b). Another common RNN type is Long Short-Term Memory (LSTM), which has additional gates in an RNN cell (Chen et al. 2014). LSTM has been dealing with vanishing gradient problem successfully so it can capable of learning long time-dependencies contrary to vanilla RNN (Weininger 1988b). BiLSTM is the hybrid of BRNN and LSTM, and it showed better performance in classification and recognition task (Gurulingappa et al. 2009; Lumini and Nanni 2018).

3.2.3 Training and evaluation

We retrieved drug datasets and the corresponding peptide sequence of target proteins and SMILE sequence data in public database. We select drugs and the corresponding peptide sequence of target proteins and SMILE sequence data which are long enough for model to learn features.

All drug strings and corresponding amino acid sequences are retrieved from the Drugbank version 5.0.8 (Law et al. 2013). We selected the drugs that (1) have at least one known target gene, (2) have the SMILES string which is longer than 30 characters, (3) have only one ATC code, and (4) do not belong to class V. The class V of ATC classification system means "various types", so the drugs in this class are miscellaneous, sharing no similarity. Only 0.52% of the drug database were included in class V, so the number of excluded data was not critical. The number of ATC classes used in this work is 13.

Drugs can have a various numberof target genes. Some have only one target gene, but others may have more target genes. Genes produce proteins madeup of linearly concatenated amino-acids. We fixed the length for input amino acid sequence as 100, and those shorter than100 are excluded. In an amino acid sequence longer than 100, randomly chosen segments of 100 amino acids of each gene and each drug were used. This makes any two different input drugs share a target gene pair with different amino acid segments the same target

Table 3.2 Prediction accuracy (%) comparison with other methods

		All classes	**5 major**
Model comparison	CNN *(proposed)*	74.09	**89.12**
	RNN *(proposed)*	**77.19**	88.24
	SVM	53.86	66.70
	RF	53.86	66.70
	DT	51.02	63.08
Previous work	Aliper et al. 2016*	55	62
	Gurulingappa et al. 2009	77.12	-
	Xie et al. 2017	72.93	-

*12 categories from MeSH code was used as class labels.

gene. Hundred-length amino acid segments were then one-hot encoded for 20 dimensions.

In our experiments, the length of input SMILES strings L and the number of character types of SMILES M are fixed to 30 and 100, respectively.

We split the dataset 85% for train and 15% for test, respectively. The numbers of drugs of each ATC class are extremely imbalanced, so we constructed two experimental settings. First, we used every drug in 13 classes. Second, we removed the drugs from minor classes, and used drugs only from five major classes: A, C, J, L, and N. All other experimental settings remained the same in both of the experiments.

We measured classification accuracy on test split, and compared our result to previous studies and other ML-based models. For ML-based models, support vector machine (SVM), random forest (RF), and decision tree (DT) were used for comparison. We plotted the receiving operator characteristic (ROC) curve and estimated the area under the curve (AUC) for each experiment.

3.3 Experimental results

We compared the prediction accuracies of our model with those of other methods in Table 3.2. With the CNN-based model, the prediction accuracies

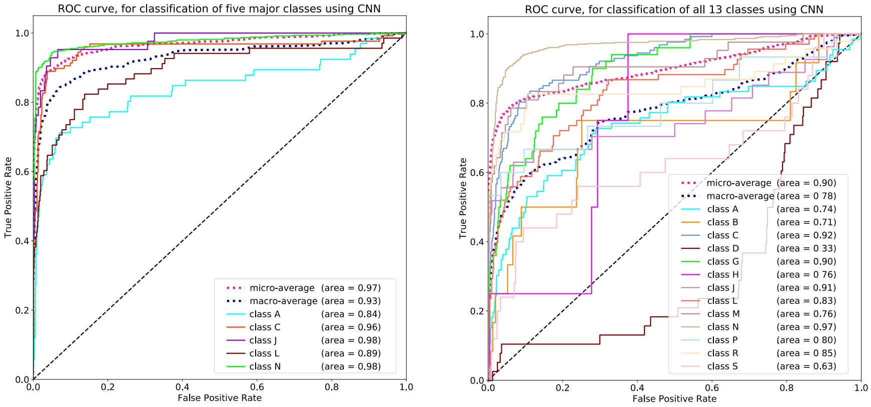


Figure 3.4 ROC curve in 1DCNN-based model. The classification accuracies (left) from all 13 classes and from five major classes (right).

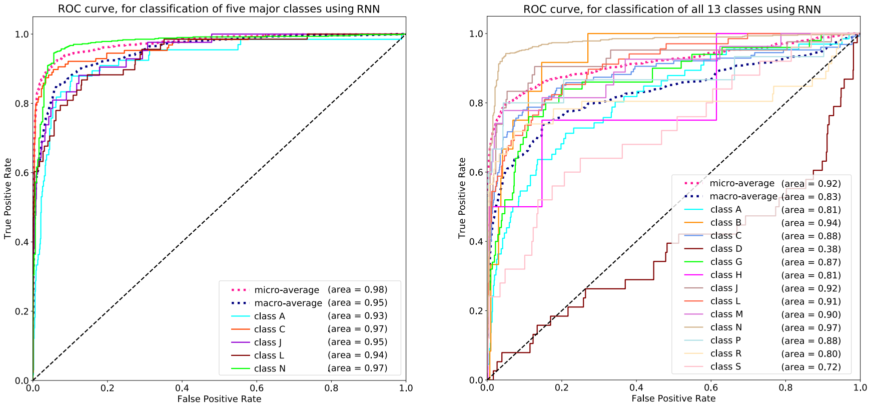


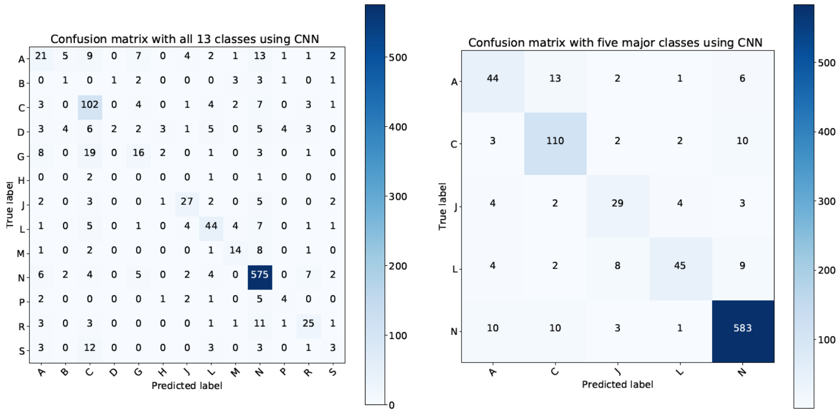
Figure 3.5 ROC curve in BiLSTM-based model. The classification accuracies (left) from all 13 classes and from five major classes (right).

were 74.09% and 89.12% from the experiments using all 13 classes and five major classes, respectively. When RNN was used, the prediction accuracies were 77.19% and 88.24% from the same experimental settings, respectively. In summary, in the case of using drugs from the five major classes, the CNN-based model performed slightly better. On the contrary, in the case of using from all 13 classes, the RNN-based model showed better model predictability.

The performances from other ML-based models were worse than those from our model. Gurulingappa et al. 2009 used information extracted from scientific articles with a ML-based method. This model scored classification accuracy of 77.12% for classifying ATC codes. Xie et al. 2017 used transcriptome data, their softmax-regression model achieved a prediction accuracy of 72.93%. Other work using DL-based model (Chen et al. 2012), their accuracies were 55% and 62% for all drug classes and five major classes, respectively. Note that Aliper et al. 2016 used a different drug classification system (Medical Subject Headings, or MeSH, which consists of 12 categories) (Lipscomb 2000) for prediction labels.

The ROC curve plots are described in Figure 3.4 and Figure 3.5. The AUC values for five major classes are generally higher than those from the minor classes.

The detailed results are shown in confusion matrices in Figure 3.6 and 3.7. The number of samples in class N is extremely large, because the drugs in class N have generally have more target proteins than drugs in other classes, so the number of input pairs (SMILES and corresponding target amino sequence) was increased. In prediction of all 13 classes, both of the model predict more accurately in more larger classes (Figure 3.6 and Figure 3.7), respectively.



Confusion matrix with all 13 classes using CNN

True label \ Predicted label	A	B	C	D	G	H	J	L	M	N	P	R	S
A	21	5	9	0	7	0	4	2	1	13	1	1	2
B	0	1	0	1	2	0	0	0	3	3	1	0	1
C	3	0	102	0	4	0	1	4	2	7	0	3	1
D	3	4	6	2	2	3	1	5	0	5	4	3	0
G	8	0	19	0	16	2	0	1	0	3	0	1	0
H	0	0	2	0	0	0	0	1	0	1	0	0	0
J	2	0	3	0	0	1	27	2	0	5	0	0	2
L	1	0	5	0	1	0	4	44	4	7	0	1	1
M	1	0	2	0	0	0	0	1	14	8	0	1	0
N	6	2	4	0	5	0	2	4	0	575	0	7	2
P	2	0	0	0	0	1	2	1	0	5	4	0	0
R	3	0	3	0	0	0	0	1	1	11	1	25	1
S	3	0	12	0	0	0	0	3	0	3	0	1	3

Confusion matrix with five major classes using CNN

True label \ Predicted label	A	C	J	L	N
A	44	13	2	1	6
C	3	110	2	2	10
J	4	2	29	4	3
L	4	2	8	45	9
N	10	10	3	1	583

Figure 3.6 Confusion matrix of prediction accuracy from 1DCNN-based model. The classification results with all 13 classes (left) and five major classes (right).

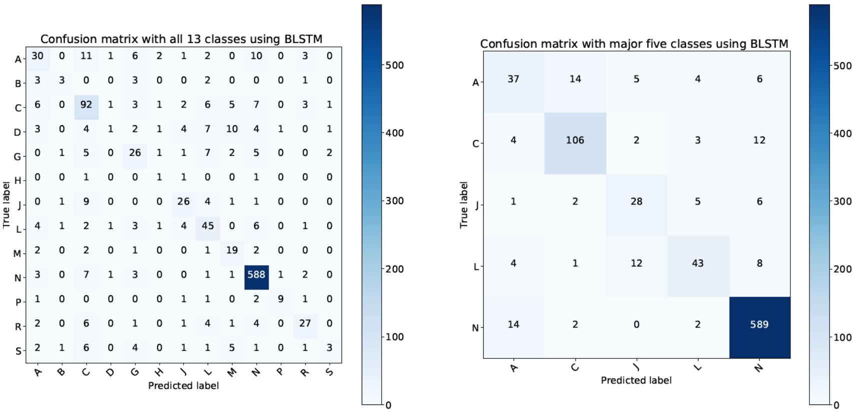


Figure 3.7 Confusion matrix of prediction accuracy from BiLSTM-based model. The classification results with all 13 classes (left) and five major classes (right).

3.4 Discussion

In summary, we constructed two kinds of deep neural networks for drug classification prediction using DrugBank database, and evaluated the model performances. In front part, we build two one-dimensional deep convolutional neural networks: one for amino acid sequence and one for chemical string data. We integrated one deep neural network to multi-class classification for drugs. The architecture of the other model is the same, except that it uses two one-dimensional BiLSTM cells rather than convolution layers. We measured the prediction accuracy rates of our models and compared our results to other machine learning classifiers and previous works based on learning algorithms. We concluded that our models outperformed the other classifiers in both experiments with different groups of classes. The model's better accuracy in the five major groups (A, C, J, L, and N), it is because the model could learn relatively more patterns in these groups than the smaller groups. Especially, the model with RNN outperforms slightly better the one with CNN in AUC scores. It may be caused by because an RNN cell can remember past (or next) patterns, so naturally more fit to string data whereas a convolution filter in CNN is localized and catches patterns spatially. So an RNN fits better for string data in general. The prediction performance should be higher in minor classes. It is because the deep learning could not learn enough for the classes for few samples. It could be solved if the more drug datasets are created in this classes.

In conclusion, we verified that the molecule representation and their binding targets are sufficient to predict therapeutic properties of corresponding drugs in DrugBank database. Our model showed remarkable performances compared with previous models or ML-based models, without extensive feature engineering. However, there are a few limitations of the research: Drugs in more than one category (multi-label drugs) were excluded. Second, there was a size limit for the input data, so our model was validated on only small-sized partial segments of sequences. Third, our model was not validated the real-world dataset. These problems have to be solved in the future.

4 Chapter 2. Biophysical property prediction with molecular string

4.1 Introduction

In this Chapter, we discuss more basic property prediction tasks in chemistry. Drug classification can be considered as a high-level property of chemicals. Drug metabolism is determined by the biochemical and physiological properties of the drug molecule. To improve the performance of a drug property prediction model, it is important to extract complex molecular dynamics from limited data. Especially, more advanced research, such as novel drug development or personalized medicine, low-level properties such as biophysical or physiological properties would be more useful. Toxicity, lipophilicity, and binding affinity are the examples of low-level properties of organic molecules. These properties are essential to predict efficacy of medicinal supplements in pharmacological research. Many public databases provide the related dataset for constructing of the ML or DL-based model.

From early days, SMILES strings have been widely adopted for various chemical tasks (Ertl et al. 2017a; Goh et al. 2017; Shin et al. 2019; Honda et al. 2019; Zheng et al. 2019). LSTM networks (Ertl et al. 2017a), CNNs (Öztürk et al. 2018), or both of two architectures (Paul et al. 2018) are used to predict physiological or pharmaceutical properties on SMILES strings. A self-attention method and Bidirectional gated recurrent unit were also attempted Shin et al. 2019; Zheng et al. 2019 and Lin et al. 2019, respectively.

Many previous studies (Atwood and Towsley 2016; Kearnes et al. 2016; Niepert et al. 2016; Altae-Tran et al. 2017; De Cao and Kipf 2018; Lee et al. 2018) have also employed graph neural networks (GNNs), to predict various molecular properties based on molecular structures. More recently, spatial-based graph convolutional neural networks (GCNNs), a type of spatial-based GNNs, showed better performances compared with those from other architectures (Kearnes et al. 2016; Shin et al. 2019; Atwood and Towsley 2016; Niepert et al. 2016; Altae-Tran et al. 2017; De Cao and Kipf 2018; Lee et al. 2018; Huang et al. 2019).

Among various types of deep architectures, message passing neural networks (MPNNs) (Gilmer et al. 2017) is the most widely used architecture nowadays. MPNN was originally proposed to predict quantum-property prediction tasks, and widely adopted following studies (Unke and Meuwly 2019; Lu et al. 2019; Anderson et al. 2019; Yang et al. 2019; Klicpera et al. 2020b; Withnall et al. 2020) because it is simple and correlated with chemical theory.

However, most of these approaches can handle the graph-formatted data only, which can not include auxiliary molecular attributes. As mentioned above, SMILES strings consist of valuable information about molecular structures such as bond types or aromaticity.

Motivated by the success of DL methods in natural language processing (NLP), we utilized architectures which were originally developed for language modeling to train SMILES strings. We assumed that a string format of a molecule could also be considered as a sentence, so NLP-based methods could also be effective for molecular tasks. In particular, we adopt the MP-based

Table 4.1 The details of benchmark datasets used in this paper.

Dataset	HIV	BACE	BBBP	Tox21	SIDER	ClinTox
The num. of tasks	1	1	1	12	27	2
The num. of molecules	41127	1513	2039	7831	1427	1478

Table 4.2 The examples of the byte-pair encoding tokenization of SMILES strings.

Dataset	tokens of byte-pair encoding
HIV	CC, 1=, Cl, ccc, CN, Br,)[, +](=, CCOC, ...
Tox21	CC, ccc, nc,)[, CCN, @@, nH, Mg, ...

attention network (MPAD) (Nikolentzos et al. 2019), which uses an MPNN framework for language tasks with word co-occurrence graphs. In addition, we also adopt other preprocessing methods for languages, such as byte-pair encoding (Sennrich et al. 2016), word2vec (Mikolov et al. 2013) and word co-occurrence graph Mihalcea and Tarau 2004 methods.

To the best of our knowledge, it is the first proposed study to learn chemical strings based on an MPNN framework. The model performances showed remarkable performances compared with previous studies.

4.2 Proposed method

We constructed co-occurrence graphs from SMILES strings using token embeddings and Word2Vec (Mikolov et al. 2013). We applied MPAD (Nikolentzos et al. 2019) to train those graphs for predicting biochemical properties. Details are described below.

4.2.1 Preprocessing

In the initial step, we converted string data to a co-occurrence graph (Mihalcea and Tarau 2004). We used two types of word tokenization methods. First, we tokenized all characters individually, which assumed that each character in the SMILES string corresponded to a word in a sentence. Next, we adopted byte-pair encoding (Sennrich et al. 2016). Byte-pair encoding is a kind of data compression method, which groups common consecutive bytes into a one byte. It is also the case for molecular formulas, because several atoms have two-character names such as Na, Mg, or Pb. We expected that by using byte-pair encoding, any atom type notation with two characters could automatically be grouped by one embedding. The examples of encodings are in Table 4.2.

We applied the word2vec method (Mikolov et al. 2013) to both types of molecular tokens obtained from input SMILES strings. Word2vec is a type of word embedding method, which embeds words into vectors based on their semantic associations in a large context. Note that we do not adopt any pretraining method which is widely used in NLP tasks, because we found that the protocols of SMILES strings were slightly different with databases.

Subsequently, we made co-occurrence graphs from the embedded SMILES strings. We built edges between all pairs of two consecutive tokens following the previous work MPAD (Nikolentzos et al. 2019). Consequently, each token is represented as individual nodes in a graph, and all nodes have two neighbours which are located right before and after the token in the string. Note that the edges do not necessarily accord with the original covalent bonds in a molecule. We do not utilize any bond information for graph representations.

4.2.2 model architecture

We utilized the previous model MPAD proposed in (Nikolentzos et al. 2019). MPAD is a type of MPNNs, which handles the word graph G from text documents. The document graph can be denoted as $G = (V, E)$, where V and

E are the vertices of words and edges of connections. The MPAD architecture consists of consecutive blocks, which include *message function* f and *update function* g:

$$M_{t+1} = f^{t+1}(D^{-1}AH^{t}) \tag{4.1}$$

$$H^{t+1} = g(H^{t}, M^{t+1}) \tag{4.2}$$

D is the diagonal in-degree matrix of the text graph G and A is the adjacency matrix of G. $D^{-1}AH$ assigns the neighbors of a H, then f^{t} produces the message of t-th timestep M^{t} for feed-forward propagation. Then g receives M^{t} and the node features of t-th timestep H^{t} and updates the node features of $(t+1)$-th timestep H^{t+1}. In MPAD, f and g are a multilayer perceptron (MLP) layer and a gated recurrent unit (GRU) layer (Cho et al. 2014), respectively.

The *readout function* is formalized by the final document representation step from the transpose of all word representations H^{T} with a self-attention α^{T}

$$y = \sum {\alpha_i}^{T} \hat{H_i}^{T} \tag{4.3}$$

MPAD introduced a master node, which represents the global node summarizing the whole document. The master node has a skip connection with the output from the readout function. The multi-readout is the concatenation of all H^{t} time steps (t=1,...,T) to form a $h_G \in \mathbb{R}^{T \times 2d}$, which is not described for simplicity. Overall architectures are described in Figure 4.1.

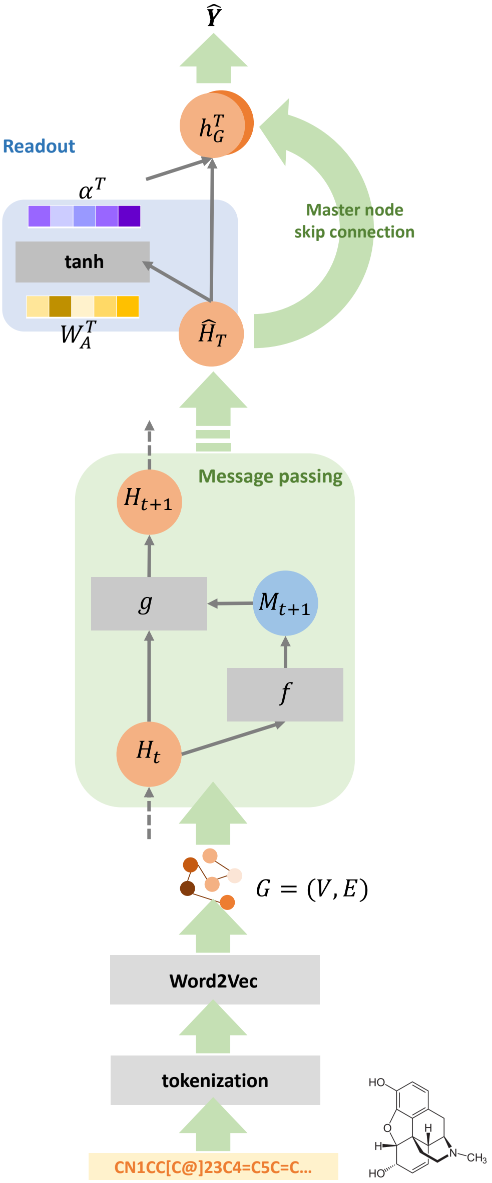


Figure 4.1 The model architecture of the MPAD. In the beginning, the SMILES strings are tokenized either by each character or by byte-pair encoding. The tokenized n words are vectorized by Word2Vec, and represented as a word co-occurrence network. The hidden representation $H \in \mathbb{R}^{n \times d}$ of a graph G comes into T an iterative *message passing* phase. The last hidden representation H^T goes to the *readout* stage, and combines with a learnable weight vector ${W_A}^T$. The final output comes out after global attention. Besides, *master node skip connection* is the process that the master node representation bypasses the attention mechanism and concatenated to the output from the attention step. ${h_G}^T \in \mathbb{R}^{2d}$ is the final result of the network on given SMILES strings.

4.2.3 Training and evaluation

Data collection

We utilized the public benchmark tasks in MoleculeNet (Wu et al. 2018). MoleculeNet provides various types of chemical subtasks, which are classified into one of four categories: quantum mechanics, physical chemistry, biophysics, or physiology. The SMILES strings are provided in all subtasks. We selected classification subtasks in biophysics and physiology categories. These subtasks are as follows: HIV, BACE, BBBP, Tox21, ClinTox, and SIDER (Table 4.1). Note that all classification tasks are binary, which have labels of positive and negative.

All subtasks used in this study are imbalanced. For example, in 12 labels of Tox21 tasks, only 2.4-12.0% of the labels were positive. For training, we adopt three loss functions for classifications, binary cross entropy (BCE), weighted binary cross entropy (WBCE), and focal loss (FL) (Lin et al. 2017) to tackle these imbalances. The formula of three functions are described below.

Loss function

$$BCE\,loss = -y\log p - (1-y)\log(1-p) \tag{4.4}$$

$$WBCE\,loss = -w_0 y\log p - w_1(1-y)\log(1-p) \tag{4.5}$$

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{4.6}$$

BCE function (4.4) is the standard loss function for binary classification tasks. WBCE function (4.5) is more suitable when the label is seriously imbalanced. In Particular, we used 0.9 and 0.1 for positive and negative labeled data, respectively. Focal loss (4.6) was originally developed for object detection tasks, especially when the foreground and background pixel classes are extremely imbalanced. We assumed that this imbalance is analogous to other imbalanced binary tasks, so we adopt FL to our classification tasks. We compared three types of loss functions in our experiments.

Evaluation

We employed the receiver operating characteristic-area under curve (ROC-AUC) following the MoleculeNet guidelines (Wu et al. 2018). We also followed the train, validation, and test splits of DeepChem guidelines (Ramsundar et al. 2019). We compared our results to previous works and two types of baseline models. The baseline models were 1D-CNN and BiGRU, respectively. In multiple classification tasks (Tox21, SIDER, and ClinTox), we trained and evaluated the model performances in each task, individually. The overall performances were obtained by averaging.

Experimental setting

For preprocessing, we used deepchem package 2.3.0 (Ramsundar et al. 2019), and sentencepiece 0.1.85 (Sennrich et al. 2016) for byte-pair encoding, and gensim 3.8.1 (Řehůřek and Sojka 2010) for Word2Vec embedding. For training, PyTorch 1.4.0 (Paszke et al. 2019) was used. The train and evaluation processes were built based on the original code (Nikolentzos et al. 2019). For ablation studies, we executed the experiments of two types of tokenization, and three loss function types. We also used three types of word embedding dimensions, 256, 128, and 64. The number of training epochs was up to 64. Adam optimizer (Kingma and Ba 2014) was used with the learning rate of 0.00025. The hidden layer dimension was set to 256. The α and γ of FL were set at 0.2 and 2, respectively. All other hyperparameters followed those used in the original MPAD.

Table 4.3 The model performances over all datasets compared with the previous studies and baseline models.

Dataset	HIV	BACE	BBBP	Tox21	SIDER	ClinTox
ChemixNet	0.9678	-	-	0.8526	-	-
ST	0.683	0.719	0.900	0.706	0.559	0.963
SMILES2vec	0.80	-	-	0.81	-	-
SA-BiLSTM	0.80	-	-	0.842	-	-
Smi2Vec	0.9457	**0.844**	0.8539	0.7806	0.6071	0.9779
1D-CNN	0.9706	0.7737	**0.9422**	0.7483	0.5806	0.9687
BiGRU	0.9704	0.7362	0.8903	0.7082	0.6078	0.9625
MPAD (Ours)	**0.9707**	0.8137	**0.9422**	**0.8551**	**0.6368**	**0.9866**

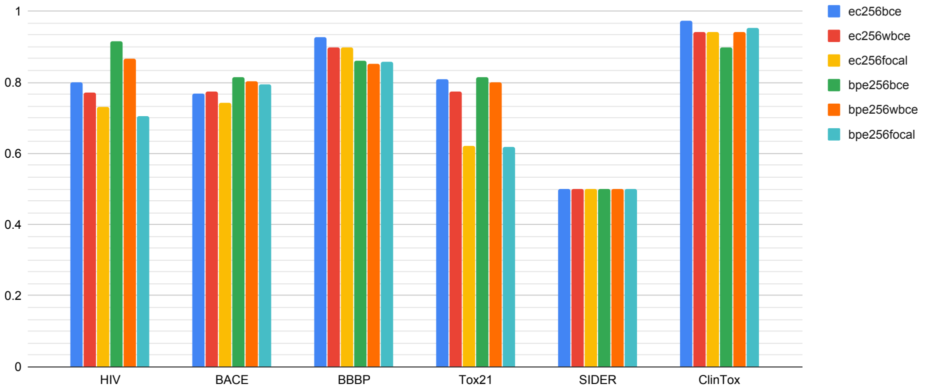


Figure 4.2 The performance of each experimental setting from five datasets. We fixed the token embedding dimension as 256.

4.3 Experimental results

We compared our ROC-AUC values with five previous studies and two baseline models. Note that the previous models may use different data splits from those of ours. The overall results are provided in Table 4.3. Our model achieved the best performance from five of the six subtasks. On the HIV task, the AUC score of our model was 0.9707, which is the best score among other models. The ChemixNet (Paul et al. 2018), Smi2Vec (Lin et al. 2019), BiGRU and 1DCNN baselines also achieved high scores as those of ours. On the BACE

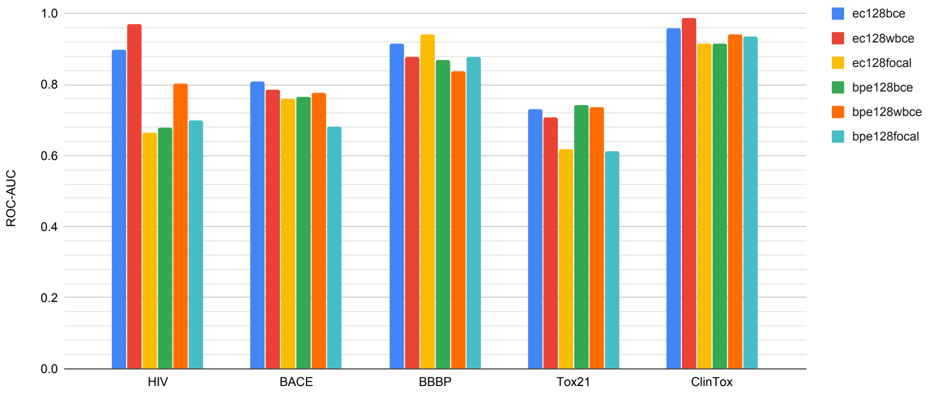


Figure 4.3 The performance of each experimental setting from five datasets. We fixed the token embedding dimension as 128.

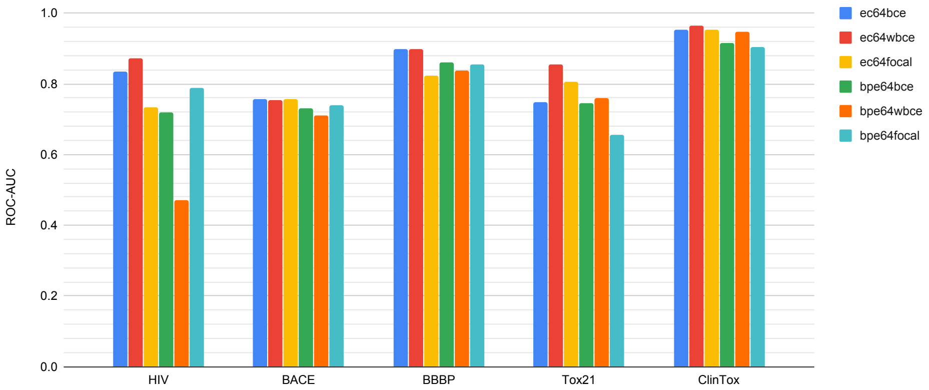


Figure 4.4 The performance of each experimental setting from five datasets. We fixed the token embedding dimension as 64.

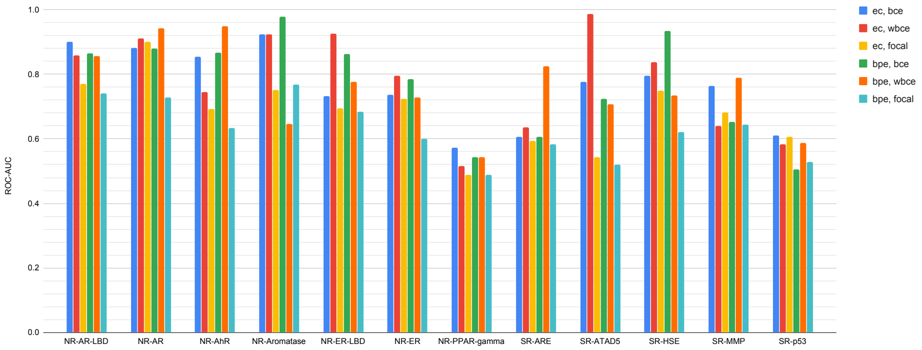


Figure 4.5 The detailed performances of each subtask in Tox21 dataset.

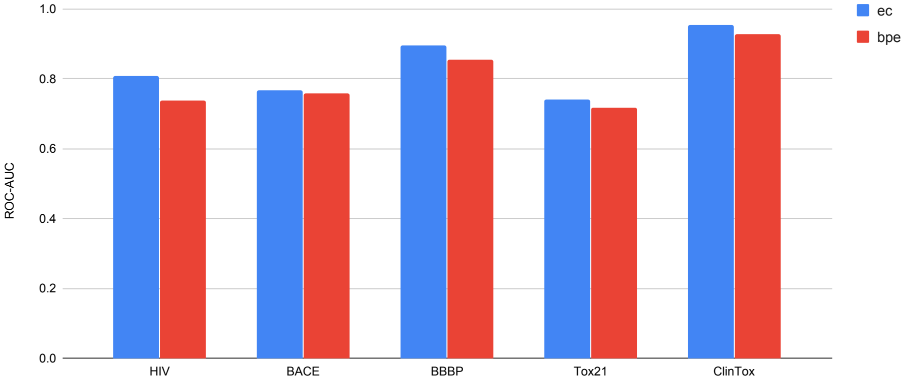


Figure 4.6 The average performance by the tokenizing methods from five datasets.

task, Smi2Vec achieved the highest score of 0.844, followed by 0.8137 from our model. On the BBBP task, our model and 1DCNN baselines achieved the same best score, 0.9422. Other three tasks Tox21, SIDER, and ClinTox, our method outperformed all other methods.

Results from ablation studies are provided in Figure 4.2, 4.3 and 4.4. Figure 4.2 presents the results from two tokenizations and three loss functions, with fixed embedding dimension to 256. Figure 4.3 and 4.4 are with fixed embedding dimensions to 128 and 64, respectively. We denoted two character embedding types as *ec* for 'each character' and *bpe* for byte-pair encoding. The following numbers mean embedding dimensions of their experimental settings. The three

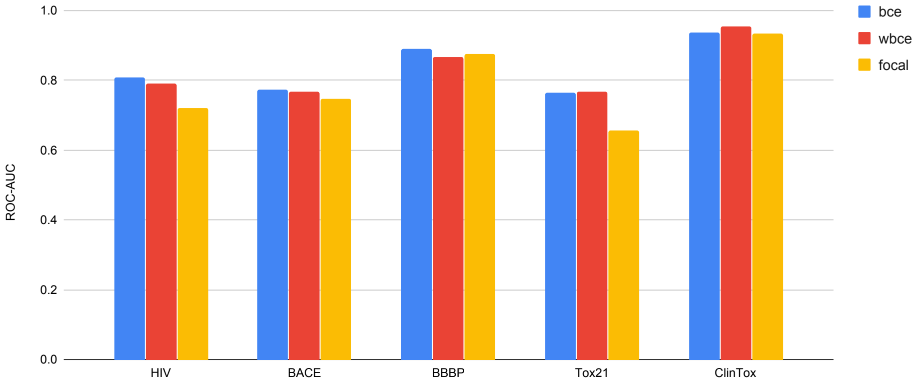


Figure 4.7 The average performance by the loss functions from five datasets.

types of loss functions are denoted as *bce*, *wbce*, and *focal* for binary cross-entropy, weighted binary cross-entropy, and focal loss, respectively.

We could not find notable effects of embedding dimensions from five tasks. The results from SIDER are not described in Figures 4.2-4.7, because the results obtained for 27 labels varied significantly, so we did not average on these values. On the HIV dataset, the best ROC-AUC was 0.9707 from *ec* tokenization with embedding dimension of 128, and trained on *wbce*. On the BACE dataset, the highest performance was 0.8137, obtained from *bpe* tokenizing with embedding dimension of 256, and BCE loss. On the BBBP task, *ec* tokenizing was always better than *bpe* tokenizing. On the Tox21, it is obvious that *focal* loss always yielded lower performances compared to other loss types, whereas it was not the case of other tasks. The best performance was 0.8551, yielded from *ec* tokenizing with 64 dimensions, and *wbce* loss. On the ClinTox, *ec* tokenizing of 128 embedding dimensions with *wbce* loss yielded the best performance of 0.9866.

We presented more detailed results on Tox21 dataset in Figure 4.5. The performances differed from each other. In all 12 subtasks, the highest averaged score was 0.8745 from the NR-AR subtask. The lowest score was 0.5251 from NR-PPAR-gamma subtask. These differences may have been due to the degree of label imbalances. In NR-PPAR-gamma subtask, the rates of the positive labels in the train and test are only 2.38% and 2.42%, respectively. These are the lowest rates of positive labels in those of the Tox21 subtasks. Similar results were already reported in the previous study (Lin et al. 2019).

In Figure 4.6, the averaged performances by both tokenizing methods are described. In all subtasks, the model performed better when SMILES strings were tokenized by *ec* than when *bpe* was used. The performance gaps between using *ec* and *bpe* are 0.07, 0.01, 0.04, 0.02, and 0.03 from the HIV, BACE, BBBP, Tox21, and ClinTox, respectively.

We also provide the averaged performances by three types of loss functions, in Figure 4.7. In all tasks except the BBBP, the models trained on *bce* or *wbce* loss outperformed the models trained on *focal* loss. The performance gaps between *bce* and *wbce* were not significantly different from five subtasks.

4.4 Discussion

We applied the MPNN framework model for text (Nikolentzos et al. 2019) on the public chemical classification benchmarks with SMILES string data. Our model showed outstanding performances in most of the subtasks, compared to previous models and baseline models. To the best of our knowledge, our model is the first MPNN for molecular string data. We did extensive experiments on six benchmark datasets for binary classification. These datasets are included in either biochemical or physiological subfields.

In details, we used two types of SMILES tokenization, three types of embeddings, and three types of loss functions. We found the best combination of these experimental settings. The benchmarks used in this work are all critical in pharmacological studies, to design a new drug or to repurpose existing drugs. So we expect our results to be the novel guide for drug development or pharmacological researches. We think that there is still a possibility to improve the prediction performances with optimizing the message-passing architecture of the networks. The regression and generation tasks are also remained as future works.

Our model suggests that message passing-based methods are also effective on string data, not only on graph data. We also expect our results to be the novel guide for drug development or pharmacological researches. It will be possible to improve the prediction performances with more optimizations or additional datasets. We expect these methods can be utilized in other tasks, such as regression, in the future. Furthermore, the validation of real dataset is also required.

5 Chapter 3. Quantum-mechanical property prediction with molecular graph

5.1 Introduction

Among various branches of chemistry, quantum chemistry covers a fundamental level of study in atoms and their mechanics. Atomic energies and polarities are examples of quantum chemical properties. Therefore, predicting quantum chemical properties is essential in most applied chemistry fields.

Conventionally, density functional theory (DFT) has been successfully adopted to investigate quantum-mechanical properties for years. However, DFT-based analyses require combinatorial search space, so tremendous computations are inevitable. In addition, DFT cannot filter out unnecessary features, which may decrease prediction accuracy. For these reasons, DL-based approaches have arisen to relieve those computational and efficacy issues. Early DL-based works built a simple neural network (Behler and Parrinello 2007; Bartók et al. 2010) which consists of perceptrons with the same number of target molecule atoms. However, these approaches could not be extended to variable sizes of molecules.

A message passing neural network (MPNN) (Gilmer et al. 2017), aforementioned in the previous section, has been proposed to overcome all of the problems of early studies. An MPNN framework does a graph convolution on

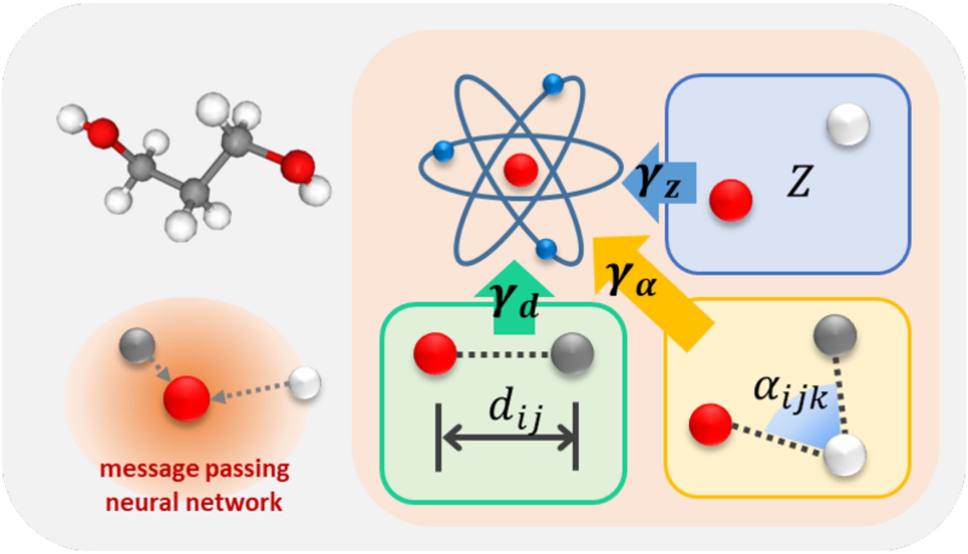


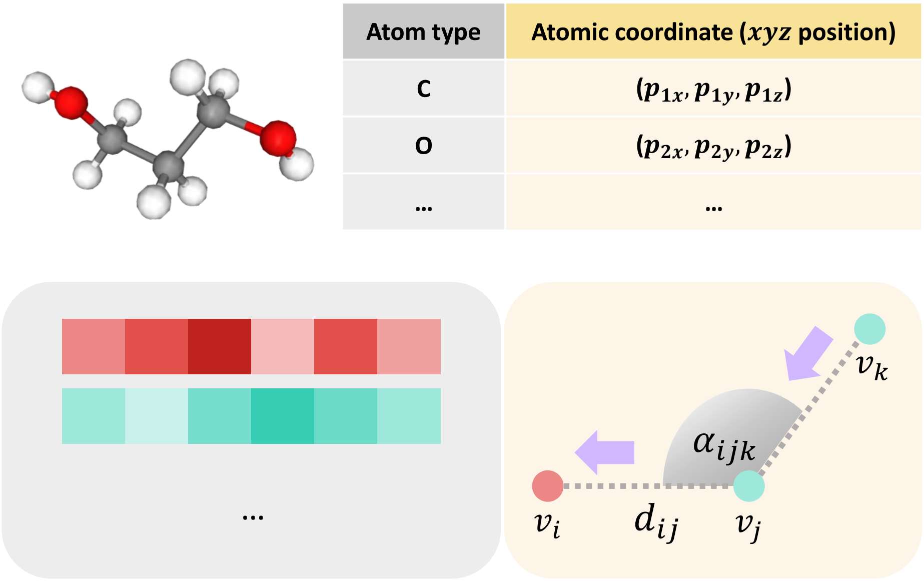
Figure 5.1 Basic concept of the message passing-based pathway of our model. We used atom charge Z, distances of atom pairs d_{ij}, and angles from three atoms α_{ijk} as inputs. We introduced trainable γ values, to flexibily weight more significant features, according to the target.

each atom and its neighbors, and updates the aggregated *message* to learn atom-wise features. Most recent studies (Schütt et al. 2017, 2018; Chmiela et al. 2017; Kondor 2018; Unke and Meuwly 2019; Chen et al. 2019; Klicpera et al. 2020b; Anderson et al. 2019; Kondor et al. 2018; Fuchs et al. 2020) have adopted a message passing framework for molecular property prediction at atom-wise level. In particular, several works (Kondor 2018; Kondor et al. 2018; Anderson et al. 2019; Fuchs et al. 2020; Smidt 2020) adopted roto-translation group equivariance, to meet the criteria of molecular mechanics. However, there are still several limitations in this approach. MPNNs may suffer from an over-smoothing problem (Li et al. 2018) because of repeatedly mixing local features. Furthermore, the MPNNs cannot differentiate over-mixed heterogeneous features or regions, so it is difficult to differentiate important features in MPNNs.

To tackle these issues, we developed a novel message-passing based neural network. Our proposed network has two branches, one for message-passing and the other for atom-wise. We also introduced trainable scalars, to automatically amplify or reduce more important features by each target label. Following previous MPNNs for molecular graphs, we used two types of features: charges and positions of atoms. Broad concepts are described in Figure 5.1. Our experiments verified that various molecular features contribute differently according to target molecular properties.

5.2 Proposed method

Our proposed network is dual-branched, consisting of two different types of paths. One is the MPNN, a type of spatial-based GCNs, and the other is the MLP-based network. Trainable scalars are attached to each block.



Atom type	Atomic coordinate (xyz position)
C	(p_{1x}, p_{1y}, p_{1z})
O	(p_{2x}, p_{2y}, p_{2z})
…	…

Figure 5.2 Data preprocessing. An atom type Z and atomic coordinate $\mathbf{p}$ were used in the model. We created trainable embeddings X from Z, and calculated distances d_{ij} and angles α_{ijk} from the coordinates $\mathbf{p_i}$, $\mathbf{p_j}$, and $\mathbf{p_k}$.

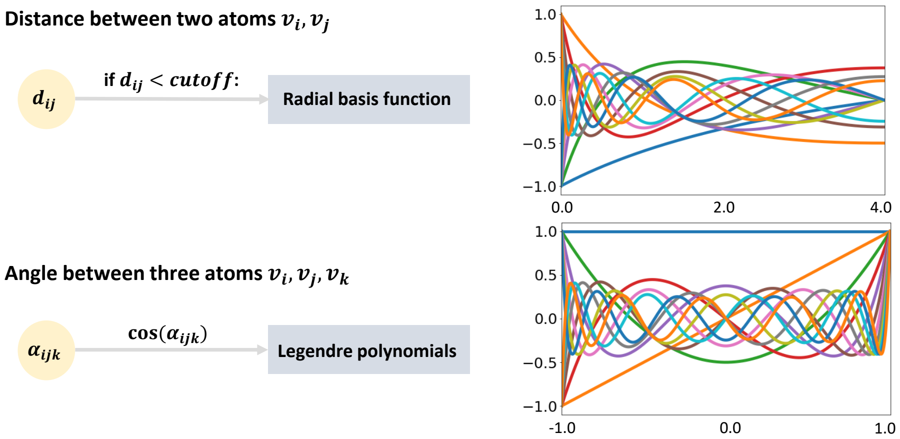


Figure 5.3 Basis functions for data preprocessing. For any pair of two atoms located closer than a given $cutoff$, we created an edge representation between two atoms regardless of molecular bond information. A scalar-valued distance is expanded as n-dimensional vector by radial basis functions. For radial basis functions, we used Legendre rational polynomials. Angle representations are depicted by any two edges sharing one atom. A cosine-valued angle is represented as the degree$=1, 2, ..., m$-th Legendre polynomials (top right).

5.2.1 Preprocessing

Data embedding

We briefly introduce the notations. We describe a molecule as a group of different atoms $v_i = \{z_i, \mathbf{p}_\mathrm{i}\}$ where $\mathbf{z}$, $\mathbf{p}$, and i denote a scalar-valued atom charge, a vector-valued atom position in the Euclidean space, and the index of atom in a molecule. Note that the index i is arbitrary, because the order of index does not affect the training. Therefore, we denote a group of atom charges and positions as sets: $\mathrm{z} = \{\mathrm{z_i}\}$ and $\mathbf{p_i} = (p_{ix}, p_{iy}, p_{iz})$. We denote $v_{j,j \in N(i)}$ as a set of neighboring atoms of v_i.

Data preprocessing

We used charges of atoms and their spatial arrangements as input features. We did not use atom positions directly; instead, we calculated distances d_{ij} and angles α_{ijk} from combinations of atom coordinates $\mathbf{p_i}$, $\mathbf{p_j}$, and $\mathbf{p_k}$, followed by previous works (Klicpera et al. 2020b). We also embedded scalar-valued atom charges Z as vectors X. More details are described in Figure 5.2.

Distance representation

We used radial basis functions to represent a scalar-valued d_{ij} to a vector. We adopted sequences of Legendre rational polynomials, one of the continuous and bounded orthogonal polynomials. The Legendre rational function $R_n(x)$ and Legendre rational polynomial $P_n(x)$ are defined as:

$$R_n(x) = \frac{\sqrt{2}}{x+1} P_n(\frac{x-1}{x+1}) \tag{5.1}$$

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n \tag{5.2}$$

where n denotes the degree. The *n-th order* Legendre rational polynomials are recursive relations as follows:

$$R_{n+1}(x) = \frac{2n+1}{n+1} \frac{x-1}{x+1} R_n(x) - \frac{n}{n+1} R_{n-1}(x) \text{ for } n \geq 1. \tag{5.3}$$

We selected the first *n-th order* polynomials $R_1(x), R_2(x), \ldots, R_n(x)$, such that $d_{ij} \in \mathbb{R}$ can be embedded in an n-dimensional vector $\mathbf{d} \in \mathbb{R}^n$. The plot of functions degree of 1,2.,,,12 are described in Figure 5.3. We modified the equations to be bounded at 1.0, to make their distributions be similar to those of atom charge embeddings z.

Recall that we did not use any molecular bond information. Following previous research (Schütt et al. 2017, 2018; Unke and Meuwly 2019; Klicpera et al. 2020b), we set a single scalar cutoff c and assumed that any atom pair v_i and v_j located closer than the cutoff can interact with each other, and vice versa, before training. In other words, we made edges $e_{ij} = (\mathrm{z_i},\ \mathrm{z_j},\ R_1(d_{ij})|R_2(d_{ij})|...|R_{12}(d_{ij}))$ between any two atoms close to each other. The cutoff value is analogous to kernel length in the standard convolution layer, because it determines the size of the receptive field of local representations.

Angle representation

In a similar way of distance representation, we adopted other sequences of orthogonal polynomials to represent a scalar-valued angle α_{ijk} as vectors: Legendre polynomials of the first kind. Legendre polynomials of the first kind are the orthogonal polynomials that exhibit several useful properties. They are the solutions to the Legendre differential equation (Anli and Gungor 2007). They are usually used to represent many types of physical systems. Their formula is:

$$Q_n(x) = 2^n \sum_{k=0}^{n} x^k \binom{n}{k} \binom{\frac{(n+k-1)}{2}}{n} \tag{5.4}$$

The polynomials of degree n are orthogonal each other, such that:

$$\int_{-1}^{1} Q_m(x) Q_n(x)\, dx = 0 \quad \text{if } n \neq m \tag{5.5}$$

We calculated angles α_{ijk} between any pair of two edges sharing one node, e_{ij} and e_{jk}. We selected the first *m-th order* polynomials $Q_1(x), Q_2(x), \ldots, Q_m(x)$, such that a scalar-valued angle $\alpha_{ijk} \in \mathbb{R}$ is embedded as a m-dimensional vector $\alpha \in \mathbb{R}^m$. The scheme is presented in Figure 5.3. We calculated cosine values of each angle, and embedded them with Legendre polynomials.

Comparison of distances and angle representations

We briefly introduce the difference between two types of Legendre polynomials. First, the sequences of Legendre rational functions $\{P_k(x)\}$ are adopted for representing $\{d_{ij}\}$, which have different distributions over the distances. The standard deviations of the distributions of $\{P_k(x)\}$ are higher at shorter distance values. Therefore, they can approximate the interatomic potentials, on the assumption that the atom pair in shorter distances would have stronger relationships. In other words, $\{P_k(x)\}$ can make richer representations for closely located atom pairs.

Next, all Legendre polynomials of the first kind $\{Q_k(x)\}$ are defined within the range of [-1, 1]. These functions are symmetric (even functions) or antisymmetric (odd functions) at the zero point. Because the cosine function is a type of even function and ranges from -1 to 1, $\{Q_k(x)\}$ can cover the cosine angle of $[0, 2\pi]$, and is symmetric at the π. In the Euclidean spaces, the angle of θ $(0 \leq \theta \leq \pi)$ between two vectors is equal to the angle of $(\pi - \theta)$. Therefore, Legendre polynomials of the first kind $\{Q_k(x)\}$ can represent desired properties for representing the cosine values of angles.

5.2.2 Model architecture

The overall model architecture is described in Figure 5.4. There are two separate branches in the network, which include the MLP-based branch for atom-wise representations and MPNN-based branch for locally mixed representations. First, an atom type Z is embedded by the embedding block. The distances are calculated from atomic positions p, and if a distance d_{ij} is shorter than the c, then d_{ij} is represented as an edge of a molecular graph. The scalar d_{ij} is expanded to a vector d_{ij} in the radial basis function $\{P_k(x)\}$. The angles

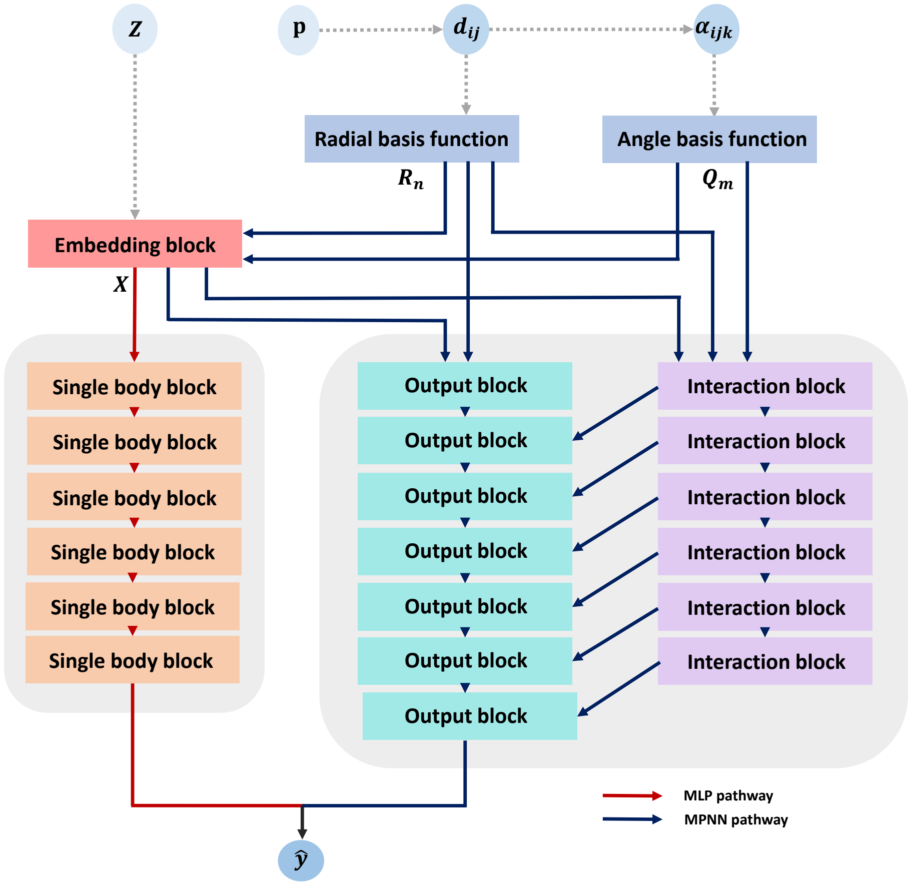


Figure 5.4 Model architecture. The network comprises two separate branches, MLP-based (red arrow) and MPNN-based (blue arrow) pathways. Atom types Z are embedded as trainable matrices. Distances d_{ij} and angles α_{ijk} are calculated from atomic coordinates. We calculated $R_n \in \mathbb{R}^{12}$ from d_{ij}. We also calculated $Q_m \in \mathbb{R}^{12}$ from α_{ijk}. All blocks except the embedding block are stacked multiple times (not sharing weights). In this model, we stacked six, seven, and six blocks for the single body blocks, output blocks, and interaction blocks, respectively. For simplicity, skip connections are not shown.

α_{ijk} are calculated from the edges. Then α_{ijk} is encoded to an m-dimensional vector before training via Legendre polynomials of the first kind $\{Q_k(x)\}$. In this study, we set $m = 12$.

In our model, the message passing and readout steps of our model can be summarized as:

$$\begin{aligned}
m_i^{t+1} &= f_m^t(h_i^t, h_j^t, \overrightarrow{e_{ij}}, \alpha_{ijk}) \quad (t = 1, ..., T) \qquad \text{(message passing)} \\
h_i^{t+1} &= f_u^t(h_i^t, m_i^{t+1}) \qquad \text{(update)} \qquad (5.6) \\
\hat{y} &= f_r(v_i \in V \mid h^T) \qquad \text{(readout)} \qquad (5.7) \\
& \qquad (5.8)
\end{aligned}$$

where m, h, e, and $alpha$ are the message of atom v_i and its neighbors, single-atom representation of atom v_i, interaction between two neighboring atoms (v_i, v_j) and angle value between three atoms (v_i, v_j, v_k), respectively. t is the time step or layer order in the model, and $yhat$ represents the predicted target value. Both f_m^t and f_u^t represent the graph convolution, and f_r is the sum operation.

Our model has four types of blocks: the embedding block, output block, single body block, and interaction block. All blocks except the embedding block are consecutively stacked. Detailed descriptions are presented in the next section and Figure 5.5.

Block details

We introduce four types of blocks used in our model. The embedding block E, the output blocks O_t, the single body blocks S_t, and the interaction blocks I_t, where $t = 0, \ldots, (T-1)$ is the time step of the multiple sequential blocks.

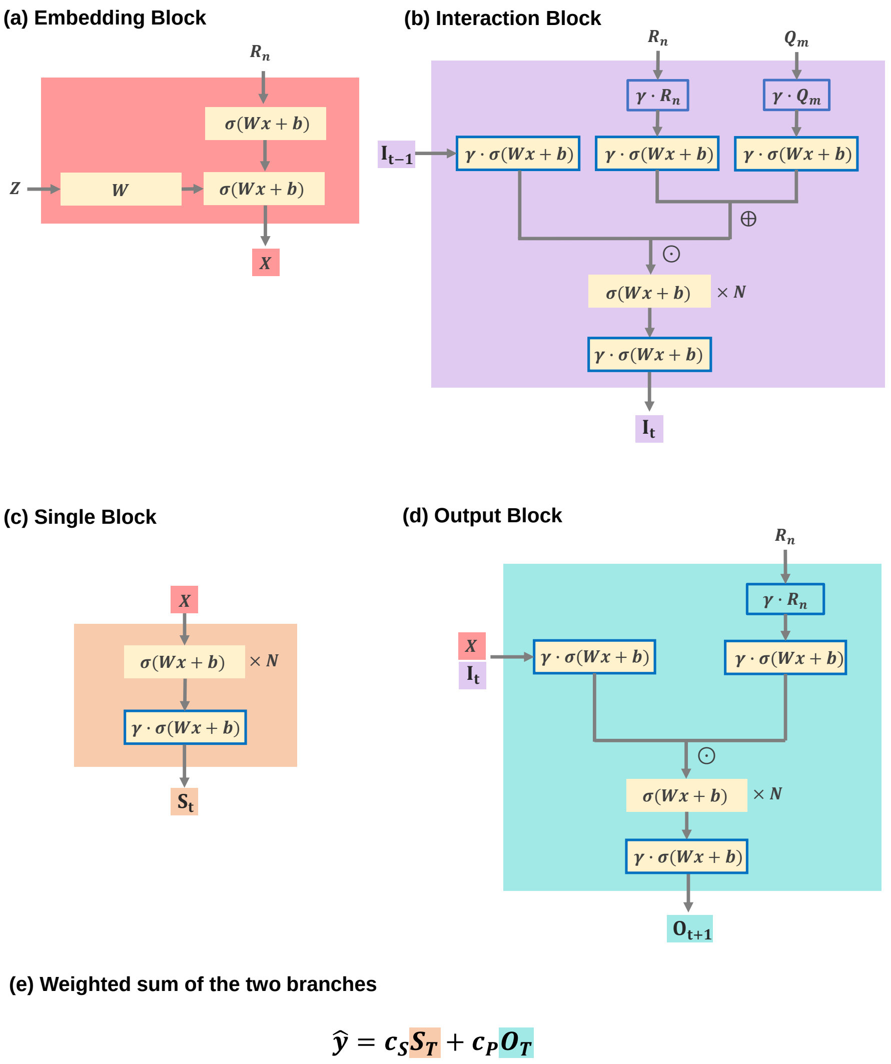


Figure 5.5 Block structures. (a) embedding block (left top). (b) interaction block (right top). (c) single body block (left bottom). (d) output block (right bottom). We denote the dense layer with the activation as $\sigma(Wx+b)$. We also denote the trainable scalars as γ. All blue boxed items, including dense layers and features are γ-*trainable*. Subscripts are omitted for simplicity. (e) weighted aggregation formula of two final outputs from the single block and output block pathways. c_S and c_P denote the trainable scalar-valued coefficients.

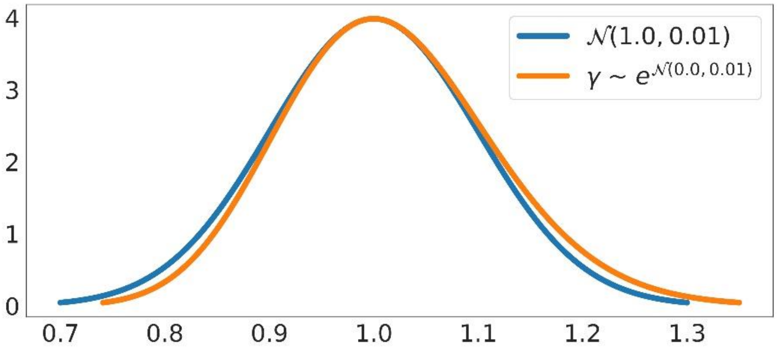


Figure 5.6 Gaussian distribution and the initialization of γ distribution. Both distributions attain their maximum values at $x = 1.0$

Atom type embeddings Z enter the embedding block E at the first step, and the embedded X moves to other three blocks at time step $t = 0$. Distance embeddings d_{ij} are used for all block types, except the single body blocks S_t. Angle embeddings α_{ijk} are used only in I_t. All blocks except the E are sequentially stacked by time steps $t = 0, \ldots, (T - 1)$ including several skip connections. The length of time steps T of each block can differ from each other.

In detail, the embedding block E embeds atom representations, and the output blocks O_t update the messages from the interaction blocks I_t. The interaction blocks I_t make messages of each atom v_i with X_i, d_{ij}, and α_{ijk}. The single body blocks are separated from the message-passing path consisting of I_t and O_t, and train atom-wise representations with MLP layers.

Trainable scalars

As mentioned above, we do not consider any additional layer or attention, preventing the model from the over-smoothing problem and becoming cost-intensive. Instead, we introduce a more simple solution. We attached trainable

Table 5.1 Description of the 12 targets of the QM9 dataset (Ramakrishnan et al. 2014)

Property	Unit	Description
μ	D	Dipole moment
α	a_0^3	Isotropic polarizability
ϵ_{HOMO}	Ha	Energy of HOMO
ϵ_{LUMO}	Ha	Energy of LUMO
ϵ_{gap}	Ha	Gap(ϵ_{LUMO}-ϵ_{HOMO})
¡R^2¿	a_0^2	Electronic spatial extent
zpve	Ha	Zero point vibrational energy
U_0	Ha	Internal energy at 0 K
U	Ha	Internal energy at 298.15 K
H	Ha	Enthalpy at 298.15 K
G	Ha	Free energy at 298.15 K
C_v	$\frac{cal}{molK}$	Heat capacity at 298.15 K

scalars γ to each layer. This makes the intensity of signals from heterogenous factors flexible, according to each target. Mathematical formulas are described as:

$$\begin{aligned} h^{t+1} &= \sigma(Wh^t + b) \quad \text{(conventional dense layer)} \\ h^{t+1} &= \gamma * \sigma(Wh^t + b) \quad \text{(with trainable scalars)} \end{aligned} \tag{5.9}$$

We introduced γ to the layers of O_t, I_t, and S_t. All γ values are trained independently to each other (subscripts were omitted for simplicity). We initialized γ values from the exponential function of random normal distribution N, namely, . The distribution of γ initialization is presented in Figure 5.6.

5.2.3 Training and evaluation

Data collection

We used QM9 (Ramakrishnan et al. 2014) and molecular dynamics (MD) simulation datasets (Chmiela et al. 2017, 2018) for the experiments. QM9 is the most popular quantum mechanics database, consisting of 134 small organic molecules made up of five atom types (carbon, oxygen, hydrogen, nitrogen, and fluorine). Each molecule has twelve numerical target properties, so the entire task is a set of twelve multiple regressions. The details of targets is presented in Table 5.1.

The MD simulation dataset (Chmiela et al. 2017, 2018) was proposed for the energy prediction task from molecular conformations. The simulation data are given as trajectories. The subtasks are divided according to each molecule type. The energy values and forces of each atom are given as continuously defined scalars (kcal/mol) and three-dimensional vectors, respectively. The energies can be predicted from molecular geometry and forces. We used the most recently introduced subtasks (Chmiela et al. 2018) of which the properties were created from CCSD(T) calculations. The CCSD(T) is known as the more reliable method than the conventional DFT method.

Implementation

For QM9, we trained the model at least 300 epochs for each target with an early stopping method. For MD simulation, we modified the loss function. In particular, we add the loss term of forces, as suggested in the previous works (Unke and Meuwly 2019; Klicpera et al. 2020b). All data splits were set as the guidelines of the datasets (Ramakrishnan et al. 2014). The initial learning rate and decay rate were set to 10-3 and 0.96 per 20 epochs, respectively. Adam optimizer (Kingma and Ba 2014) was used, and the batch size was set to 24, both of the datasets.

Evaluation

The evaluation metric is a mean absolute error (MAE), also following the original guideline. We compared our results with previous methods (Schütt

Table 5.2 Mean absolute error on QM9 compared with previous works

Target	Unit	SchNet	Cormorant	LieConv	DimeNet	DimeNet++	DL-MPNN
μ	D	0.033	0.038	0.032	0.0286	0.0297	**0.0256**
α	$bohr^3$	0.235	0.085	0.084	0.0469	**0.0435**	0.0444
ϵ_{HOMO}	eV	0.041	0.034	0.030	0.0278	0.0246	**0.0223**
ϵ_{LUMO}	eV	0.034	0.038	0.025	0.0197	0.0195	**0.0169**
Δ_ϵ	eV	0.063	0.061	0.049	0.0348	**0.0326**	0.0391
¡R^2¿	$bohr^2$	**0.073**	0.961	0.800	0.331	0.331	0.414
$zpve$	meV	1.7	2.027	2.280	1.29	**1.2**	**1.2**
U_0	eV	0.014	0.022	0.019	0.00802	0.0063	0.0074
U	eV	0.019	0.021	0.019	0.00789	**0.0063**	0.0074
H	eV	0.014	0.021	0.024	0.00811	0.0065	0.0076
G	eV	0.014	0.020	0.022	0.00898	**0.0076**	**0.0076**
C_v	$\frac{cal}{molK}$	0.033	0.026	0.038	0.0249	0.0249	**0.023**

Table 5.3 Mean absolute error on MD simulation compared with previous works

Target	Train method	sGDML	SchNet	DimeNet	DL-MPNN
Aspirin	Forces	0.68	1.35	**0.499**	0.590
Benzene	Forces	0.06	0.31	0.187	**0.053**
Ethanol	Forces	0.33	0.39	0.230	**0.10**
Malonaldehyde	Forces	0.41	0.66	0.383	**0.225**
Toluene	Forces	0.14	0.57	0.216	**0.200**

et al. 2017; Anderson et al. 2019; Finzi et al. 2020; Klicpera et al. 2020b,a), which used the same features (atom types and positions) based on the same framework (message passing) with those of ours.

5.3 Experimental results

We evaluated our model performances with those from the previous models mentioned above. The MAEs for QM9 and the MD simulation are described in Table 5.2 and Table 5.3, respectively. For QM9, our model achieved advanced performances in six of the twelve targets. For MD simulation, our model exhibited the best performances among other models in four of the five targets.

In QM9, our model showed better performances on the targets, which are more related to molecular interactions such as dipole moment (μ), molecular orbitals (ϵ_{HOMO} and ϵ_{LUMO}), Gibbs free energy (G), and others. Otherwise, the prediction results for the targets of electronic spatial extent ($< R^2 >$) and internal energies ($U0$, U, H) were not the best. We found that this may be related to the distribution of each target value. We found that when the standard validation of the target is smaller, the prediction accuracy would be higher in general. The similar issue was already reported in the previous study (Miller et al. 2020). In the future, we will analyse the effect of the target distribution on the model convergence.

5.4 Discussion

We briefly discuss the relation between molecular graph density and cutoff values. As mentioned above, the cutoff c value determines which atom pair has an edge representation in molecular graphs. If a large cutoff value is used, then the molecular graph representation would be *dense*. Otherwise, the graph would be *sparse*.

The dense graphs generally have higher possibilities in capturing the interactions between nodes than the sparse graphs do. However, dense structures are exposed to higher risks of gathering excessive features even when there is less important information. Furthermore, the messages of every node from its neighbours are mixed over and over during the training process. If a graph is excessively dense, most messages become indistinguishable. This may increase the risk of over-smoothing problem (Li et al. 2018), especially in deeper message-passing architecture.

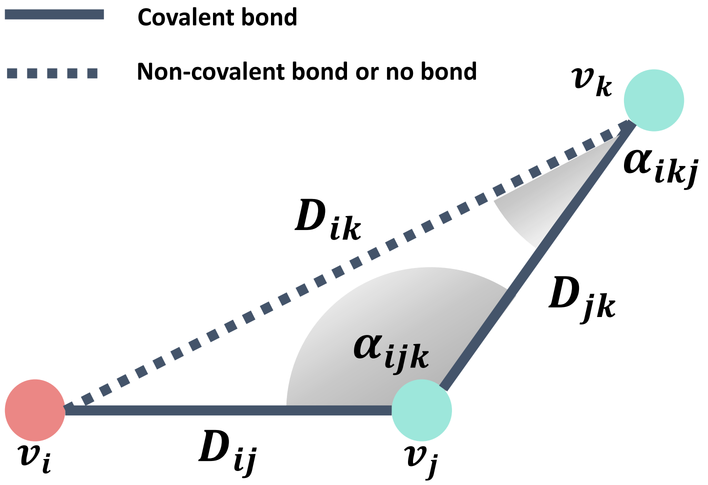


Figure 5.7 Example of bond and angle representation. All lengths of the covalent bond between two atoms of C, H, O, N, and F are less than 2.0, so any two-hop neighbor via covalent bond v_k from an atom v_i will always be located inside the cutoff = 4.0. Therefore, an atom v_i can always capture two successive covalent bonds (v_i, v_j) and (v_j, v_k).

We observed that the average distance between any atom pair in QM9 molecules is 3.27 angstrom (Å). The atom pairs within 4.0 angstrom (Å) which is the cutoff value used in this work account for 72% of all the pairs in the QM9 dataset. Previous models SchNet, MegNet, and DimeNet/DimeNet++ adopted 10.0, 5.0, and 5.0 as their cutoff values, respectively. We found that 99.99% and 89% for cutoff values of 10.0 and 5.0 of the distances are represented in molecular graphs, respectively.

We compared the densities according to different cutoff values. If the edges are within a cutoff value of 4.0, then the number of edge representations are half of the overall possible number of edges. If the cutoff value increases to 5.0, then the rate of the edge representations is increased to 80%.

We demonstrate that the cutoff value of 4.0 is sufficient in describing small organic molecules. The lengths of the covalent bonds between any two atmos among C, H, O, N, and F are always less than 2.0 (Allen et al. 1987). Because the maximum length of any two consecutive covalent bondings is bounded at 4.0, the model can capture any two-hop neighboring atoms, with a cutoff value of 4.0. This property makes it possible to identify all the angles between two covalent bonds, by triangular inequality. The idea is briefly described in Figure 5.7. From these observations, we argue that the 4.0 is the optimal value for the cutoff in chemical property prediction.

In summary, we developed a novel dual-branched network, message passing framework for atom-atom interactions, and fully connected layers for atom-wise representations. We introduced trainable scalars, to make heterogeneous flexibly contribute according to target types. In addition, we adopted a smaller cutoff value than those from the previous MPNN models, keeping the receptive field of a message at least two-hop neighbors. Our model showed comparable or better results in two public benchmark tasks, QM9 and MD simulation. Future works will be focused on other molecular property predictions for more complex molecules, and validation on the real datasets.

6 Conclusion

We studied several molecular property prediction tasks with molecular structure, based on deep learning architectures. We utilized two types of molecular representations and four types of model architectures. We adopt several public benchmark tasks. First, we used a SMILES string data with 1DCNN and BiLSTMS, on drug category classification prediction using DrugBank. Next, we referred to the previous study of the MPNN model for document classification. We build co-occurrence graphs from SMILES strings, and train on biochemical property prediction tasks on MoleculeNet. Finally, we investigated a more fundamental task, the prediction of quantum mechanical properties. We assumed that in diverse quantum mechanical properties prediction tasks including QM9 and MD simulation, the heterogeneous features may contribute differently according to target types. Therefore, we developed a novel dual-branched network, and introduced trainable scalars to endow flexibility with our model.

To summarize our studies, our contributions are as follows: We explored various computational public benchmark tasks in each subfield in chemistry, and validated that molecular structures are enough to predict various propertion labels on each dataset. Our proposed models showed competitive performances compared to most recent studies, and validated our intuitions behind the specific problems of each benchmark task. Our studies suggest that developing

DL-based models with molecular structures are prospective approaches, from fundamental research to high-level applications.

However, there were several limitations in our study. Our methods were developed based on small organic molecules only. The generalizability in large molecules has not been validated yet. Our methods depend on local feature maps in most of the cases, because they use a convolution in common. It may prevent the model from capturing long-range interactions, which is not desirable in molecular mechanics analyses. Finally, our methods were validated only on public benchmark datasets, not real-world dataset.

To address these limitations, we will explore other approaches rather than local convolutions, using other benchmarks of large molecules, or real-word datasets as the future works.

Bibliography

Agatonovic-Kustrin, S, and Rosemary Beresford, 2000: Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, **22**, 717–727.

Aliper, Alexander, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, and Alex Zhavoronkov, 2016: Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular Pharmaceutics*, **13**. PMID: 27200455, 2524–2530. DOI: `10.1021/acs.molpharmaceut.6b00248`. eprint: `http://dx.doi.org/10.1021/acs.molpharmaceut.6b00248`. URL: `http://dx.doi.org/10.1021/acs.molpharmaceut.6b00248`.

Allen, Frank H, Olga Kennard, David G Watson, Lee Brammer, A Guy Orpen, and Robin Taylor, 1987: Tables of bond lengths determined by x-ray and neutron diffraction. part 1. bond lengths in organic compounds. *Journal of the Chemical Society, Perkin Transactions 2*, S1–S19.

Altae-Tran, Han, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande, 2017: Low data drug discovery with one-shot learning. *ACS central science*, **3**, 283–293.

Anderson, Brandon, Truong Son Hy, and Risi Kondor, 2019: Cormorant: covariant molecular neural networks. *Advances in Neural Information Processing Systems*, 14510–14519.

Anderson, Eric, Gilman D Veith, and David Weininger, 1987: *SMILES, a line notation and computerized interpreter for chemical structures.* US Environmental Protection Agency, Environmental Research Laboratory.

Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle, 2016: Deep learning for computational biology. *Molecular systems biology*, **12**, 878.

Anli, Fikret, and Süleyman Gungor, 2007: Some useful properties of legendre polynomials and its applications to neutron transport equation in slab geometry. *Applied mathematical modelling*, **31**, 727–733.

Atwood, James, and Don Towsley, 2016: Diffusion-convolutional neural networks. *Advances in neural information processing systems*, 1993–2001.

Bartók, Albert P, Mike C Payne, Risi Kondor, and Gábor Csányi, 2010: Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Physical review letters*, **104**, 136403.

Behler, Jörg, and Michele Parrinello, 2007: Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, **98**, 146401.

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

Benning, Leo, Andreas Peintner, Günter Finkenzeller, and Lukas Peintner, 2020: Automated spheroid generation, drug application and efficacy screening using a deep learning classification: a feasibility study. *Scientific Reports*, **10**, 1–11.

Butler, Keith T, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh, 2018: Machine learning for molecular and materials science. *Nature*, **559**, 547–555.

Chang, Yoosup, Hyejin Park, Hyun-Jin Yang, Seungju Lee, Kwee-Yum Lee, Tae Soon Kim, Jongsun Jung, and Jae-Min Shin, 2018: Cancer drug response profile scan (cdrscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Scientific reports*, **8**, 1–11.

Chen, Chi, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong, 2019: Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, **31**, 3564–3572.

Chen, Fan-Shu, and Zhen-Ran Jiang, 2015: Prediction of drug's anatomical therapeutic chemical (atc) code by integrating drug–domain network. *Journal of Biomedical Informatics*, **58**, 80 –88. ISSN: 1532-0464. DOI: `https://doi.org/10.1016/j.jbi.2015.09.016`. URL: `http://www.sciencedirect.com/science/article/pii/S1532046415002105`.

Chen, Hongming, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke, 2018: The rise of deep learning in drug discovery. *Drug discovery today*, **23**, 1241–1250.

Chen, Jiarui, Hong-Hin Cheong, and Shirley Weng In Siu, 2020: Bestox: a convolutional neural network regression model based on binary-encoded smiles for acute oral toxicity prediction of chemical compounds. *International Conference on Algorithms for Computational Biology*. Springer, 155–166.

Chen, Lei, Jing Lu, Ning Zhang, Tao Huang, and Yu-Dong Cai, 2014: A hybrid method for prediction and repositioning of drug anatomical therapeutic chemical classes. *Mol. BioSyst.*, **10** (4), 868–877. DOI: 10.1039/C3MB70490D. URL: http://dx.doi.org/10.1039/C3MB70490D.

Chen, Lei, Wei-Ming Zeng, Yu-Dong Cai, Kai-Yan Feng, and Kuo-Chen Chou, Apr. 2012: Predicting anatomical therapeutic chemical (atc) classification of drugs by integrating chemical-chemical interactions and similarities. *PLOS ONE*, **7**, 1–7. DOI: 10.1371/journal.pone.0035254. URL: https://doi.org/10.1371/journal.pone.0035254.

Cherkasov, Artem, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al., 2014: Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, **57**, 4977–5010.

Chmiela, Stefan, Huziel E Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko, 2018: Towards exact molecular dynamics simulations with machine-learned force fields. *Nature communications*, **9**, 1–10.

Chmiela, Stefan, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller, 2017: Machine learning of accurate energy-conserving molecular force fields. *Science advances*, **3**, e1603015.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, 2014: Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cohen, N Claude, 1996: *Guidebook on molecular modeling in drug design*. Gulf Professional Publishing.

Coley, Connor W, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen, 2019: A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, **10**, 370–377.

Cova, Tânia FGG, and Alberto ACC Pais, 2019: Deep learning for deep chemistry: optimizing the prediction of chemical patterns. *Frontiers in chemistry*, **7**, 809.

De Cao, Nicola, and Thomas Kipf, 2018: Molgan: an implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*.

DeFever, Ryan S, Colin Targonski, Steven W Hall, Melissa C Smith, and Sapna Sarupria, 2019: A generalized deep learning approach for local structure identification in molecular simulations. *Chemical science*, **10**, 7503–7515.

Dimiduk, Dennis M, Elizabeth A Holm, and Stephen R Niezgoda, 2018: Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering. *Integrating Materials and Manufacturing Innovation*, **7**, 157–172.

Doerr, Stefan, Maciej Majewski, Adrià Pérez, Andreas Kramer, Cecilia Clementi, Frank Noe, Toni Giorgino, and Gianni De Fabritiis, 2021: Torchmd: a deep learning framework for molecular simulations. *Journal of chemical theory and computation*, **17**, 2355–2363.

Duvenaud, David K, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams, 2015: Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 2224–2232.

Ekins, Sean, Ana C Puhl, Kimberley M Zorn, Thomas R Lane, Daniel P Russo, Jennifer J Klein, Anthony J Hickey, and Alex M Clark, 2019: Exploiting machine learning for end-to-end drug discovery and development. *Nature materials*, **18**, 435–441.

Ertl, Peter, Richard Lewis, Eric Martin, and Valery Polyakov, 2017a: In silico generation of novel, drug-like chemical matter using the lstm neural network. *arXiv preprint arXiv:1712.07449*.

Ertl, Peter, Richard Lewis, Eric J. Martin, and Valery Polyakov, 2017b: In silico generation of novel, drug-like chemical matter using the LSTM neural network. *CoRR*, **abs/1712.07449**. arXiv: `1712.07449`. URL: `http://arxiv.org/abs/1712.07449`.

Ferreira, Leonardo G, Ricardo N Dos Santos, Glaucius Oliva, and Adriano D Andricopulo, 2015: Molecular docking and structure-based drug design strategies. *Molecules*, **20**, 13384–13421.

Finzi, Marc, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson, 2020: Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. *International Conference on Machine Learning*. PMLR, 3165–3176.

Fuchs, Fabian B, Daniel E Worrall, Volker Fischer, and Max Welling, 2020: Se (3)-transformers: 3d roto-translation equivariant attention networks. *arXiv preprint arXiv:2006.10503*.

Gauss, Jürgen, 1995: Accurate calculation of nmr chemical shifts. *Berichte der Bunsengesellschaft für physikalische Chemie*, **99**, 1001–1008.

Gilmer, Justin, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl, 2017: Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1263–1272.

Glen, Robert C, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith, 2006: Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme. *IDrugs*, **9**, 199.

Gobbi, Alberto, and Dieter Poppinger, 1998: Genetic optimization of combinatorial libraries. *Biotechnology and bioengineering*, **61**, 47–54.

Goh, Garrett B, Nathan O Hodas, Charles Siegel, and Abhinav Vishnu, 2017: Smiles2vec: an interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034*.

Graves, Alex, Santiago Fernández, and Jürgen Schmidhuber, 2005: Bidirectional lstm networks for improved phoneme classification and recognition. *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*. Edited by Włodzisław Duch, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrożny. Berlin, Heidelberg: Springer Berlin Heidelberg, 799–804. ISBN: 978-3-540-28756-8.

Graves, Alex, and Navdeep Jaitly, 2014: Towards end-to-end speech recognition with recurrent neural networks. *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, II–1764–II–1772. URL: `http://dl.acm.org/citation.cfm?id=3044805.3045089`.

Graves, Alex, and Jürgen Schmidhuber, 2005: Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, **18**. IJCNN 2005, 602 –610. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2005.06.042. URL: http://www.sciencedirect.com/science/article/pii/S0893608005001206.

Gupta, Anvita, Alex T. Müller, Berend J.H. Huisman, Jens A. Fuchs, Petra Schneider, and Gisbert Schneider, 2018-01: Generative recurrent networks for de novo drug design. en. Molecular Informatics **37**, 1700111. ISSN: 1868-1743. DOI: 10.3929/ethz-b-000246172.

Gurulingappa, Harsha, Corinna Kolářik, Martin Hofmann-Apitius, and Juliane Fluck, 2009: Concept-based semi-automatic classification of drugs. *Journal of Chemical Information and Modeling*, **49**. PMID: 19663460, 1986–1992. DOI: 10.1021/ci9000844. eprint: https://doi.org/10.1021/ci9000844. URL: https://doi.org/10.1021/ci9000844.

Heller, Stephen, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev, 2013: Inchi-the worldwide chemical structure identifier standard. *Journal of cheminformatics*, **5**, 1–9.

Hirohara, Maya, Yutaka Saito, Yuki Koda, Kengo Sato, and Yasubumi Sakakibara, 2018: Convolutional neural network based on smiles representation of compounds for detecting chemical motif. *BMC bioinformatics*, **19**, 83–94.

Hochreiter, S, and J Schmidhuber, 1997: Long short-term memory. *Neural computation*, **9**, 1735—1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Honda, Shion, Shoi Shi, and Hiroki R Ueda, 2019: Smiles transformer: pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738.*

Hu, ShanShan, Peng Chen, Pengying Gu, and Bing Wang, 2020: A deep learning-based chemical system for qsar prediction. *IEEE journal of biomedical and health informatics*, **24**, 3020–3028.

Huang, Jingjia, Zhangheng Li, Nannan Li, Shan Liu, and Ge Li, 2019: Attpool: towards hierarchical feature representation in graph convolutional networks via attention mechanism. *Proceedings of the IEEE International Conference on Computer Vision*, 6480–6489.

Huang, Kexin, Cao Xiao, Trong Hoang, Lucas Glass, and Jimeng Sun, 2020: Caster: predicting drug interactions with chemical substructure representation. *Proceedings of the AAAI Conference on Artificial Intelligence.* Volume 34. 01, 702–709.

Huang, Ruili, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A Shahane, Anna Rossoshek, and Anton Simeonov, 2016: Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, **3**, 85.

Issa, Naiem T, Vasileios Stathias, Stephan Schürer, and Sivanesan Dakshanamurthy, 2021: Machine and deep learning approaches for cancer drug repurposing. *Seminars in cancer biology.* Volume 68. Elsevier, 132–142.

Karim, Abdul, Jaspreet Singh, Avinash Mishra, Abdollah Dehzangi, MA Hakim Newton, and Abdul Sattar, 2019: Toxicity prediction by multimodal deep learning. *Pacific Rim Knowledge Acquisition Workshop.* Springer, 142–152.

Katritzky, Alan R, Minati Kuanar, Svetoslav Slavov, C Dennis Hall, Mati Karelson, Iiris Kahn, and Dimitar A Dobchev, 2010: Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chemical reviews*, **110**, 5714–5789.

Kearnes, Steven, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley, 2016: Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, **30**, 595–608.

Kingma, Diederik P, and Jimmy Ba, 2014: Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Klicpera, Johannes, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann, 2020a: Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *NeurIPS-W.*

Klicpera, Johannes, Janek Groß, and Stephan Günnemann, 2020b: Directional message passing for molecular graphs. *International Conference on Learning Representations (ICLR).*

Kondor, Risi, 2018: N-body networks: a covariant hierarchical neural network architecture for learning atomic potentials. *arXiv preprint arXiv:1803.01588.*

Kondor, Risi, Zhen Lin, and Shubhendu Trivedi, 2018: Clebsch-gordan nets: a fully fourier space spherical convolutional neural network. *arXiv preprint arXiv:1806.09231.*

Kozma, Robert, Elaine Chin, Joel Russell, and Nancy Marx, 2000: The roles of representations and tools in the chemistry laboratory and their implications for chemistry learning. *The Journal of the Learning Sciences*, **9**, 105–143.

Kurotani, Atsushi, Toshifumi Kakiuchi, and Jun Kikuchi, 2021: Solubility prediction from molecular properties and analytical data using an in-phase deep neural network (ip-dnn). *ACS omega.*

Landrum, G, 2019: Rdkit: open-source cheminformatics, v. 2019. *GitHub (https://github. com/rdkit/rdkit).*

Lavecchia, Antonio, 2019: Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug discovery today*, **24**, 2017–2032.

Law, Vivian, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al., 2013: Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, **42**, D1091–D1097.

Lee, Geonhee, Chihyun Park, and Jaegyoon Ahn, 2019a: Novel deep learning model for more accurate prediction of drug-drug interaction effects. *BMC bioinformatics*, **20**, 1–8.

Lee, Ingoo, Jongsoo Keum, and Hojung Nam, 2019b: Deepconv-dti: prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, **15**, e1007129.

Lee, John Boaz, Ryan Rossi, and Xiangnan Kong, 2018: Graph classification using structural attention. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1666–1674.

Li, Chunyu, Eric Coons, and Alejandro Strachan, 2014: Material property prediction of thermoset polymers by molecular dynamics simulations. *Acta Mechanica*, **225**, 1187–1196.

Li, Qimai, Zhichao Han, and Xiao-Ming Wu, 2018: Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 32. 1.

Li, Xinhao, and Denis Fourches, 2021: Smiles pair encoding: a data-driven substructure tokenization algorithm for deep learning. *Journal of Chemical Information and Modeling*, **61**, 1560–1569.

Liao, Renjie, Zhizhen Zhao, Raquel Urtasun, and Richard S Zemel, 2019: Lanczosnet: multi-scale deep graph convolutional networks. *arXiv preprint arXiv:1901.01484*.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, 2017: Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, Xuan, Zhe Quan, Zhi-Jie Wang, Huang Huang, and Xiangxiang Zeng, 2019: A novel molecular representation with bigru neural networks for learning atom. *Briefings in bioinformatics*.

Lipscomb, Carolyn E, 2000: Medical subject headings (mesh). *Bulletin of the Medical Library Association*, **88**, 265.

Lipton, Zachary Chase, 2015: A critical review of recurrent neural networks for sequence learning. *CoRR*, **abs/1506.00019**. arXiv: 1506.00019. URL: http://arxiv.org/abs/1506.00019.

Liu, Zhongyang, Feifei Guo, Jiangyong Gu, Yong Wang, Yang Li, Dan Wang, Liang Lu, Dong Li, and Fuchu He, 2015: Similarity-based prediction for anatomical therapeutic chemical classification of drugs by integrating multiple data sources. *Bioinformatics*, **31**, 1788–1795. DOI: 10.1093/bioinformatics/btv055. eprint: /oup/backfile/content_public/journal/bioinformatics/31/11/10.1093_bioinformatics_btv055/2/btv055.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btv055.

Lorenz, Sönke, Axel Groß, and Matthias Scheffler, 2004: Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chemical Physics Letters*, **395**, 210–215.

Lu, Chengqiang, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He, 2019: Molecular property prediction: a multilevel quantum interactions modeling perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 33, 1052–1060.

Lumini, Alessandra, and Loris Nanni, 2018: Convolutional neural networks for atc classification. *Current pharmaceutical design*, **24**, 4007–4012.

Lusci, Alessandro, Gianluca Pollastri, and Pierre Baldi, 2013: Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of Chemical Information and Modeling*, **53**. PMID: 23795551, 1563–1575. DOI: `10.1021/ci400187y`. eprint: `https://doi.org/10.1021/ci400187y`. URL: `https://doi.org/10.1021/ci400187y`.

Maffucci, Irene, and Alessandro Contini, 2015: Tuning the solvation term in the mm-pbsa/gbsa binding affinity predictions. *Frontiers in Computational Chemistry*. Elsevier, 82–120.

Manica, Matteo, Ali Oskooei, Jannis Born, Vigneshwari Subramanian, Julio Sáez-Rodríguez, and Maria Rodriguez Martinez, 2019: Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular pharmaceutics*, **16**, 4797–4806.

Manly, Charles J, Shirley Louise-May, and Jack D Hammer, 2001: The impact of informatics and computational chemistry on synthesis and screening. *Drug discovery today*, **6**, 1101–1110.

Mayr, Andreas, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter, 2016: Deeptox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, **3**, 80.

McCarthy, Michael, and Kin Long Kelvin Lee, 2020: Molecule identification with rotational spectroscopy and probabilistic deep learning. *The Journal of Physical Chemistry A*, **124**, 3002–3017.

Michel, Julien, Nicolas Foloppe, and Jonathan W Essex, 2010: Rigorous free energy calculations in structure-based drug design. *Molecular informatics*, **29**, 570–578.

Mieres-Perez, Joel, and Elsa Sanchez-Garcia, 2020: Quantum mechanics/molecular mechanics multiscale modeling of biomolecules. *Advances in Physical Organic Chemistry*, **54**, 143.

Mihalcea, Rada, and Paul Tarau, 2004: Textrank: bringing order into text. *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404–411.

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

Miller, Benjamin Kurt, Mario Geiger, Tess E Smidt, and Frank Noé, 2020: Relevance of rotationally equivariant convolutions for predicting molecular properties. *arXiv preprint arXiv:2008.08461.*

Min, Seonwoo, Byunghan Lee, and Sungroh Yoon, 2017: Deep learning in bioinformatics. *Briefings in bioinformatics*, **18**, 851–869.

Montavon, Grégoire, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld, 2013: Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, **15**, 095003.

Monteiro, Nelson RC, Bernardete Ribeiro, and Joel Arrais, 2020: Drug-target interaction prediction: end-to-end deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics.*

Moridi, Mahroo, Marzieh Ghadirinia, Ali Sharifi-Zarchi, and Fatemeh Zare-Mirakabad, 2019: The assessment of efficient representation of drug features using deep learning for drug repositioning. *BMC bioinformatics*, **20**, 1–11.

Najafabadi, Maryam M, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic, 2015: Deep learning applications and challenges in big data analytics. *Journal of big data*, **2**, 1–21.

Niepert, Mathias, Mohamed Ahmed, and Konstantin Kutzkov, 2016: Learning convolutional neural networks for graphs. *International conference on machine learning*, 2014–2023.

Nikolentzos, Giannis, Antoine J-P Tixier, and Michalis Vazirgiannis, 2019: Message passing attention networks for document understanding. *arXiv preprint arXiv:1908.06267.*

Noé, Frank, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi, 2020: Machine learning for molecular simulation. *Annual review of physical chemistry*, **71**, 361–390.

Olah, Christopher, 2015: Understanding lstm networks.

Öztürk, Hakime, Arzucan Özgür, and Elif Ozkirimli, 2018: Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, **34**, i821–i829.

Öztürk, Hakime, Elif Ozkirimli, and Arzucan Özgür, 2016: A comparative study of smiles-based compound similarity functions for drug-target interaction prediction. *BMC bioinformatics*, **17**, 1–11.

O'Boyle, Noel M, 2012: Towards a universal smiles representation-a standard method to generate canonical smiles based on the inchi. *Journal of cheminformatics*, **4**, 1–14.

Paisitkriangkrai, Sakrapee, Jamie Sherrah, Pranam Janney, and Anton Van-Den Hengel, 2015: Effective semantic pixel labelling with convolutional networks and conditional random fields. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.*

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., 2019: Pytorch: an imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703.*

Paul, Arindam, Dipendra Jha, Reda Al-Bahrani, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal, 2018: Chemixnet: mixed dnn architectures for predicting chemical properties using multiple molecular representations. *arXiv preprint arXiv:1811.08283.*

Peterson, Kirk A, David Feller, and David A Dixon, 2012: Chemical accuracy in ab initio thermochemistry and spectroscopy: current strategies and future challenges. *Theoretical Chemistry Accounts*, **131**, 1–20.

Ramakrishnan, Raghunathan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld, 2014: Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, **1**, 1–7.

Ramsundar, Bharath, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu, 2019: *Deep Learning for the Life Sciences.* `https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837`. O'Reilly Media.

Řehůřek, Radim, and Petr Sojka, May 2010: Software Framework for Topic Modelling with Large Corpora. English. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* `http://is.muni.cz/publication/884893/en`. Valletta, Malta: ELRA, 45–50.

Reichenbächer, Manfred, and Jürgen Popp, 2012: *Challenges in molecular structure determination.* Springer Science & Business Media.

Rifaioglu, Ahmet Sureyya, Heval Atas, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Doğan, 2019: Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in bioinformatics*, **20**, 1878–1912.

Rogers, David, and Mathew Hahn, 2010: Extended-connectivity fingerprints. *Journal of chemical information and modeling*, **50**, 742–754.

Ruddigkeit, Lars, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond, 2012: Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, **52**, 2864–2875.

Ryu, Seongok, Jaechang Lim, Seung Hwan Hong, and Woo Youn Kim, 2018: Deeply learning molecular structure-property relationships using attention- and gate-augmented graph convolutional network. *arXiv preprint arXiv:1805.10988*.

Schneider, Gisbert, and Karl-Heinz Baringhaus, 2008: *Molecular design: concepts and applications*. John Wiley & Sons.

Schuster, M., and K. K. Paliwal, 1997: Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, **45**, 2673–2681. ISSN: 1053-587X. DOI: `10.1109/78.650093`.

Schütt, Kristof T, Pieter-Jan Kindermans, Huziel E Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller, 2017: Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. *arXiv preprint arXiv:1706.08566*.

Schütt, Kristof T, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller, 2018: Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, **148**, 241722.

Sennrich, Rico, Barry Haddow, and Alexandra Birch, Aug. 2016: Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 1715–1725. DOI: `10.18653/v1/P16-1162`. URL: `https://www.aclweb.org/anthology/P16-1162`.

Shin, Bonggun, Sungsoo Park, Keunsoo Kang, and Joyce C Ho, 2019: Self-attention based molecule representation for predicting drug-target interaction. *Machine Learning for Healthcare Conference*. PMLR, 230–248.

Sliwoski, Gregory, Sandeepkumar Kothiwale, Jens Meiler, and Edward W Lowe, 2014: Computational methods in drug discovery. *Pharmacological reviews*, **66**, 334–395.

Smidt, Tess E, 2020: Euclidean symmetry and equivariance in machine learning. *Trends in Chemistry.*

Smith, Justin S, Olexandr Isayev, and Adrian E Roitberg, 2017: Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, **8**, 3192–3203.

Staker, Joshua, Gabriel Marques, and J Dakka, 2020: Representation learning in chemistry. *Machine Learning in Chemistry*, **17**, 372.

Tang, Bowen, Skyler T Kramer, Meijuan Fang, Yingkun Qiu, Zhen Wu, and Dong Xu, 2020: A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *Journal of Cheminformatics*, **12**, 1–9.

Thomas, Nathaniel, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley, 2018: Tensor field networks: rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219.*

Tropsha, Alexander, 2010: Best practices for qsar model development, validation, and exploitation. *Molecular informatics*, **29**, 476–488.

Ungerer, Philippe, Carlos Nieto-Draghi, Bernard Rousseau, Göktug Ahunbay, and Véronique Lachet, 2007: Molecular simulation of the thermophysical properties of fluids: from understanding toward quantitative predictions. *Journal of Molecular Liquids*, **134**, 71–89.

Unke, Oliver T, and Markus Meuwly, 2019: Physnet: a neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation*, **15**, 3678–3693.

Vinyals, Oriol, Samy Bengio, and Manjunath Kudlur, 2015: Order matters: sequence to sequence for sets. *arXiv preprint arXiv:1511.06391.*

Wang, Haibo, Angel Cruz Roa, Ajay N Basavanhally, Hannah L Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi, 2014: Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, **1**, 034003.

Wang, Junmei, and Tingjun Hou, 2011: Application of molecular dynamics simulations in molecular property prediction. 1. density and heat of vaporization. *Journal of chemical theory and computation*, **7**, 2151–2165.

Wang, Sheng, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang, 2019a: Smiles-bert: large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 429–436.

Wang, Xiaofeng, Zhen Li, Mingjian Jiang, Shuang Wang, Shugang Zhang, and Zhiqiang Wei, 2019b: Molecule property prediction based on spatial graph embedding. *Journal of chemical information and modeling*, **59**, 3817–3828.

Wang, Yan-Bin, Zhu-Hong You, Shan Yang, Hai-Cheng Yi, Zhan-Heng Chen, and Kai Zheng, 2020: A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC medical informatics and decision making*, **20**, 1–9.

Wang, Yong-Cui, Shi-Long Chen, Nai-Yang Deng, and Yong Wang, 2013: Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics*, **29**, 1317–1324. DOI: `10.1093/bioinformatics/btt158`. eprint: `/oup/backfile/content\_public/journal/bioinformatics/29/10/10.1093\_bioinformatics\_btt158/2/btt158.pdf`. URL: `http://dx.doi.org/10.1093/bioinformatics/btt158`.

Ward, Logan, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton, 2016: A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, **2**, 1–7.

Weininger, David, 1988a: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, **28**, 31–36.

Weininger, David, Feb. 1988b: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36. ISSN: 0095-2338. DOI: `10.1021/ci00057a005`. URL: `http://dx.doi.org/10.1021/ci00057a005`.

WHO Collaborating Centre for Drug Statistics Methodology. `https://www.whocc.no/atc/structure_and_principles/`. Accessed 21 Sep 2018.

Withnall, M, E Lindelöf, O Engkvist, and H Chen, 2020: Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *Journal of Cheminformatics*, **12**, 1.

Wu, Zhenqin, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande, 2018: Moleculenet: a benchmark for molecular machine learning. *Chemical science*, **9**, 513–530.

Xie, L., S. He, Y. Wen, X. Bo, and Z. Zhang, Aug. 2017: Discovery of novel therapeutic properties of drugs from transcriptional responses based on multi-label classification. *Scientific Reports*, **7**, 7136, 7136. DOI: `10.1038/s41598-017-07705-8`.

Yang, Kevin, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al., 2019: Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, **59**, 3370–3388.

Zeng, Xiangxiang, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, and Feixiong Cheng, 2019: Deepdr: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, **35**, 5191–5198.

Zhang, Linfeng, Jiequn Han, Han Wang, Roberto Car, and E Weinan, 2018: Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters*, **120**, 143001.

Zheng, Shuangjia, Xin Yan, Yuedong Yang, and Jun Xu, 2019: Identifying structure–property relationships through smiles syntax analysis with self-attention mechanism. *Journal of chemical information and modeling*, **59**, 914–923.

초 록

딥러닝 방법론은 이미지 및 언어 처리 분야를 포함하여, 공학 및 자연과학을 포함한 여러 분야에서 진보하였다. 최근에는 특히 계산 화학 분야에서 딥러닝 기반으로 연구된 우수한 성과들이 여럿 보고되었다. 그러나 화학적인 계 내에서는 많은 종류의 요소들과 상호작용들이 복잡하게 얽혀있다. 따라서 이러한 요소들을 이용하여 화학 특성을 예측하는 것은 쉽지 않은 일이다. 결과적으로, 전통적인 방법들은 주로 상당한 비용과 시간이 소요되는 엄청난 계산량을 기반으로 하였다.

이러한 한계점을 해결하기 위하여, 본 연구는 딥러닝을 활용한 화학에서의 계산 문제를 연구하였다. 본 연구에서는 특히 분자 구조 표현 데이터를 이용, 분자의 특성을 예측하는 문제들에 집중하였다. 분자 구조는 다양한 원자들이 특정한 배열을 이루고 있는 복합체이며, 분자 특성은 이러한 원자 및 그들의 상호관계들에 의하여 결정 된다. 따라서, 분자 구조는 화학적 특성을 예측하는 문제에 있어서 필수적인 요소이다. 본 연구에서는 약학, 유기 화학, 양자 화학 등 다양한 분야에서의 화학 특성 예측연구들을 진행하였다. 분자 구조는 시퀀스 혹은 그래프 형태로 표현할 수 있고, 본 연구에서는 두 가지 형태를 모두 활용하여서 진행하였다. 본 연구는 분자 표현이 화학 분야 내의 여러 가지 태스크에 활용될 수 있으며, 분자 표현에 따른 적절한 딥러닝 모델의 선택이 모델 성능을 크게 높일 수 있음을 보였다.

주요어: 딥러닝, 분자 특성 예측, graph neural networks

학 번: 2016-30127