INTEGRATION OF TOPOLOGICAL DATA ANALYSIS AND MACHINE LEARNING FOR SMALL MOLECULE PROPERTY PREDICTIONS

By

Kedi Wu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Mathematics – Doctor of Philosophy

2018

ABSTRACT

INTEGRATION OF TOPOLOGICAL DATA ANALYSIS AND MACHINE LEARNING FOR SMALL MOLECULE PROPERTY PREDICTIONS

By

Kedi Wu

Accurate prediction of small molecule properties is of paramount importance to drug design and discovery. A variety of quantitative properties of small molecules has been studied in this thesis. These properties include solvation free energy, partition coefficient, aqueous solubility, and toxicity endpoints.

The highlight of this thesis is to introduce an algebraic topology based method, called element specific persistent homology (ESPH), to predict small molecule properties. Essentially ESPH describes molecular properties in terms of multiscale and multicomponent topological invariants and is different from conventional chemical and physical representations. Based on ESPH and its modified version, element-specific topological descriptors (ESTDs) are constructed. The advantage of ESTDs is that they are systematical, comprehensive, and scalable with respect to molecular size and composition variations, and are readily suitable for machine learning methods, rendering topological learning algorithms. Due to the inherent correlation between different small molecule properties, multi-task frameworks are further employed to simultaneously predict related properties.

Deep neural networks, along with ensemble methods such as random forest and gradient boosting trees, are used to develop quantitative predictive models. Physical based molecular descriptors and auxiliary descriptors are also used in addition to ESTDs. As a result, we obtain state-of-the-art results for various benchmark data sets of small molecule properties.

We have also developed two online servers for predicting properties of small molecules, TopP-S and TopTox. TopP-S is a software for topological learning predictions of partition coefficient and aqueous solubility, and TopTox is a software for computing element-specific tological descriptors (ESTDs) for toxicity endpoint predictions. They are available at `http://weilab.math.msu.edu/TopP-S/` and `http://weilab.math.msu.edu/TopTox/`, respectively.

This thesis is dedicated to my parents, for their love.

ACKNOWLEDGEMENTS

First and foremost I would like to sincerely express my gratitude to my advisor, Dr. Guo-Wei Wei. It has been an honor to be his Ph.D. student. I am truly thankful for his unselfish contributions of time, ideas, and funding to make my Ph.D. study productive and enjoyable. The joy and enthusiasm he has for his research was motivational for me, and this work would have not come into existence without his supervision. He has shown me, by his own example, how a great scientist and person should be.

I also want to thank Dr. Chichia Chiu, Dr. Moxun Tang and Dr. Yiying Tong, for serving on my thesis committee and providing me with extensive guidance and helpful comments.

My time at Michigan State was made enjoyable in large part due to my group members, David Bramer, Zixuan Cang, Yin Cao, Duc Nguyen, Bao Wang, Menglun Wang, Kelin Xia, Zhixiong Zhao, that become a part of my life. My time at Michigan State was also enriched by fellow graduate students, Tianbo Chen, Wenzhao Chen, Anqi Chen and Weicong Zhou, just to name a few. I am grateful for the time spent with friends playing poker all night long, for the memorable conference trips to Boston and Columbus, and for many other precious memories.

Last but not the least, I would like to thank my family for all their love and encouragement. For my parents, Tenghua Wu and Lianxiang Pan, who unconditionally support me in all my pursuits throughout my life. For my grandma who raised me up as a little kid. And most of all for Yafei Gu whose faithful support during the final stages of this Ph.D. is so appreciated. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

PH Persistent homology

ML Machine learning

DL Deep learning

ESN Element-specific network

ESPH Element-specific persistent homology

ESTD Element-specific topological descriptor

STL Single-task learning

MTL Multi-task learning

NN Neural network

CNN Convolutional neural network

DNN Deep neural network

ST-DNN Single-task deep neural network

MT-DNN Multi-task deep neural network

log P Logarithm of partition coefficient

log S Logarithm of aqueous solubility

QSAR Quantitative structure-activity relationship

QSPR Quantitative structure-property relationship

ADME absorption, distribution, metabolism, and elimination

GBDT Gradient boosting decision tree

RF Random Forest

RMSE Root mean square error

MUE Mean unsigned error

GAFF General amber force field

MIBPB Matched interface and boundary - Poisson-Boltzmann

ESES Eulerian solvent excluded surface

SAMPL Statistical assessment of the modeling of proteins and ligands

TEST Toxicity estimation software tool

HPK Hybrid physical and knowledge

AMBER Assisted model building with energy refinement

CHAPTER 1

INTRODUCTION TO CHEMINFORMATICS

1.1 An overview of small molecule modeling

The understanding of small molecule properties is key to drug design and discovery. Properties can be physical or chemical – physical properties may include partition coefficient, aqueous solubility, boiling points, density, etc., while chemical properties may refer to toxicity, chemical stability or flammability. This work is mainly focused on quantitative predictions of solvation free energy, partition coefficient, aqueous solubility, and toxicity endpoints.

Solvation process is foundamental to sophisticated processes such as protein binding, protein DNA and RNA binding, protein-protein interaction, etc [7, 8, 9, 10]. Many studies have been conducted in modeling and analysis of the solvation processes over past few decades [8, 11, 9, 7, 12, 13] and accurate prediction of solvation free energies is one of the most popular and challenging topics[14] as it is intrinsically related to binding free energy.

Many approaches have been developed to predict small molecule solvation free energies. Categorically, there are physical models and knowledge-based models. The advantage of physical models mainly lies on physical intepretability. Physical models can be further classified as explicit and implicit models. For explict solvent models, molecular mechanics (MM) [15] and hybrid quantum mechanics/molecular mechanics (QM/MM) are commonly used [16]. On the other hand, implicit solvent models include the generalized Born (GB) model as well as many other variations [17, 18, 19, 20, 21, 22] such as GBSA [23] and SM.x [24, 25]. The most popular implicit solvent model is based on the Poisson-Boltzmann (PB) theory, which retains an atomistic description of the solute molecule, while treating the solvent and includes possible ions and cofactors as a dielectric continuum [26, 27, 28, 29, 30, 31]. More recently, Gaussian-based smooth dielectric functions have also shown success for computing solvation energy of both small molecules and proteins. [32, 33]. In classical implicit solvent models, solvation free energy is split into polar and nonpolar

contributions, where polar and nonpolar parts are computed separately. Aforementioned models can be used to calculate polar part. For nonpolar part, it is shown that the solvent-accessible surface area (SASA) can be very useful [34, 35] while there are still some drawbacks [36, 37, 38, 39]. More ecently, the coupling of polar and nonpolar components has been considered in several models [40, 41, 42]. One representative model for this coupling is based on differential geometry theory, variational approach and geometric measure theory. These mathematical apparatuses give rise to an elegant dynamical coupling of polar and nonpolar solvation components [41, 43, 42, 44]. By applying constrained optimization to nonpolar parameter selections, this model provides some of the best solvation free energy fitting and cross validation results for a large amount of solute molecules [45].

The partition coefficient, denoted P and defined to be the ratio of concentrations of a solute in a mixture of two immiscible solvents at equilibrium, is of great importance in pharmacology. It measures the drug-likeness of a compound as well as its hydrophobic effect on human body. The logarithm of this coefficient, i.e., $\log P$, has proved to be one of the key parameters in drug design and discovery. Optimal $\log P$ along with low molecular weight and low polar surface area plays an important role in governing kinetic and dynamic aspects of drug action. In particular, Hansch et al. [46] gave a detailed description of how lipophilicity impacted pharmacodynamics. This being said, surveys show that approximately half of the drug candidates fail to reach market due to unsatisfactory pharmacokinetic properties or toxicity [47], which indeed makes $\log P$ predictions even more important. The extent of existing reliable experimental $\log P$ data is negligible compared to tremendous compounds whose $\log P$ data are practically needed. Therefore, computational prediction of partition coefficient is an indispensable approach in modern drug design and discovery.

Since the pioneering work of Hansch *et al.* [48, 49, 50], a large variety of octanol-water partition coefficient predictors has been developed over the past few decades. Many methods are generally called as quantitative structure-activity relationship (QSAR) models. In general, these models can be categorized into atom-based additive methods, fragment/compound-based methods, and property based methods. One of atom-based additive methods, which was first proposed by Crippen

and his co-workers [51], is essentially purely additive and effectively a table look-up per atom. Later on, XLOGP3, a refined version of atom-based additive methods, was developed [5]. This approach considers various atom types, contributions from neighbors, as well as correction factors which help overcome known difficulties in purely atomistic additive methods. However additivity may fail in some cases, where unexpected contributions to $\log P$ occur, especially for complicated structures. Fragment/compound based predictors, instead of employing information from single atom, are built at compounds or fragments level. Compounds or fragments are then added up with correction factors. Popular fragment methods include KOWWIN [52, 53], CLOGP [54, 55], ACD/LOGP [56, 57], and KLOGP [58, 59]. A major challenge for fragment/compound based methods is the optimal classification of "building blocks". The number of fragments and corrections involved in current methods range from hundreds to thousands, which could be even larger if remote atoms are also taken into account. This fact may lead to technical problems in practice and may also cause overfitting in modeling. The third category is property-based. Basically property-based methods determine partition coefficient using properties, empirical approaches, three dimensional (3D) structures (e.g., implicit solvent models, molecule dynamics (MD) methods), and topological or electrostatic indices. Most of these methods are modeled using statistical tools such as associative neural network (ALOGPS) [60, 61]. It is worthy to mention that property-based methods are relatively computationally expensive, and depend largely on the choice of descriptors and accuracy of computations. This to some extent results in a preference of methods in the first two categories over those in the third.

Another closely related physical property is aqueous solubility, denoted by S, or its logarithm value $\log S$. In drug discovery and other related pharmaceutical fields, it is of great significance to identify molecules with undesirable water solubility on early stages as solubility affects absorption, distribution, metabolism, and elimination processes (ADME) [62, 63]. QSPR models, along with atom/group additive models [64, 3, 65, 2], have been developed to predict solubility. For example, QSPR models assume that aqueous solubility correlates with experimental properties such as aforementioned partition coefficient and melting point [66], or molecular descriptors such

as solvent accessible area. However, due to the difficulty of experimentally measuring solubility for certain compounds, the experimental data can contain errors up to 1.5 log units [67, 68] and no less than 0.6 log units [69]. Such a high variability brings challenge to solubility prediction.

Speaking of chemical properties, toxicity is among the most significant ones. Toxicity is a measure of the degree to which a chemical can adversely affect an organism. These adverse effects, which are called toxicity endpoints, can be either quantitatively or qualitatively measured by their effects on given targets. Qualitative toxicity classifies chemicals into toxic and nontoxic categories, while quantitative toxicity data set records the minimal amount of chemicals that can reach certain lethal effects. Most toxicity tests aim to protect human from harmful effects caused by chemical substances and are traditionally conducted in *in vivo* or *in vitro* manner. Nevertheless, such experiments are usually very time consuming and cost intensive, and even give rise to ethical concerns when it comes to animal tests for chemical properties. Therefore, computer-aided methods, or *in silico* methods, have been developed to improve prediction efficiency without sacrificing too much of accuracy.

1.2 An overview of topological modeling and topological learning

The key to successful predictions of small molecule properties lies on accurate representation of a given molecule. In fact, geometric representation of molecules, particularly macromolecules, often involves too much structural details and thus may become intractable for large and complex biomolecular data sets. On the contrary, topology offers the highest level of abstraction and truly metric free representations of molecules, although in most cases traditional topology incurs too much geometric reduction to be practically useful for molecules. Persistent homology bridges classical geometry and topology, offering a multiscale representation of molecular systems [70, 71]. In doing so, it creates a family of topologies via a filtration parameter, which leads to a one-dimensional topological invariants, i.e., barcodes of Betti numbers, and Betti-0, Betti-1 and Betti-2 numbers can be physically interpreted as the number of isolated components, circles, and cavities, respectively. Persistent homology has been successfully applied to the modeling and prediction

of nano particles, proteins and other biomolecules [72, 73, 74, 75, 76]. Nonetheless, it was found that primitive persistent homology has very limited predictive power in machine learning based classification of biomolecules [77], which motivates us to introduce a more sophisticated, element specific persistent homology (ESPH) to retain crucial biological information during the topological simplification of geometric complexity [78, 79, 80]. ESPH has found its success in the predictions of protein-ligand binding affinities [79, 80] and mutation induced protein stability changes [78, 80]. Thus topological tools certainly have the potential when it comes to small molecule modeling.

1.3 An overview of QSAR and machine learning

Quantitative structure activity relationship (QSAR) approach is one of the most popular and commonly used approaches in cheminformatics modeling. The basic QASR assumption is that similar molecules have similar activities. Therefore by studying the relationship between chemical structures and biological activities, it is possible to predict the activities of new molecules without actually conducting lab experiments. There are several types of algorithms to generate QSAR models: linear models based on linear regression and linear discriminant analysis [81]; nonlinear models including nearest neighbor [82, 83], support vector machine [81, 84, 85] and random forest [86]. These methods have advantages and disadvantages [87] due to their statistics natures. For instance, linear models overlook the relatedness between different features, while nearest neighbor method largely depends on the choice of descriptors. To overcome these difficulties, more refined and advanced machine learning methods have been introduced. Multi-task learning (MTL) [88] was proposed partially to deal with data sparsity problem, which is commonly encountered in QSAR applications. The idea of MTL is to learn the so-called "inductive bias" from related tasks to improve accuracy using the same representation. In other words, MT learning aims at learning a shared and generalized feature representation from multiple tasks. Indeed, MT learning strategies have brought new insights to bioinformatics since compounds from related assays may share features at various feature levels, which is extremely helpful if data set is small. Successful applications include splice-site and MHC-I binding prediction [89] in sequence biology, gene expression analysis, and

system biology [90]. MTL becomes more efficient when it is incorporated with deep learning (DL) [91, 92] strategies. Deep neural network (DNN), particularly convolutional neural network (CNN), has emerged as a powerful paradigm to render a wide range of state-of-the-art results in signal and information processing fields, such as speech recognition [93, 94] and natural language processing [95, 96], as well as toxicity prediction [97, 98, 99, 100, 101] and aqueous solubility prediction [102]. The major advancement of DNN models as compared to non-DNN models is that DNN models consist of a larger number of layers and neurons, making it possible to extract more abstract features.

1.4 Motivations and objectives

We would like to study the descriptive and predictive power of PH and ESPH for small molecules. The difficulty of small molecule modeling is that small molecules involve a wide range of chemical elements and their properties are very sensitive to their chemical constitutions, symmetry and stereochemistry. Therefore, it is not clear whether PH and ESPH are suitable descriptors for small molecules.

The objective of this thesis is to explore the representability and predictive power of ESPH for small molecules, using state-of-the-art machine learning and deep learning algorithms. We focus on the analysis and prediction of several different small molecule properties, including both physical and chemical properties. Specifically, we aim to predict partition coefficient, aqueous solubility and four different toxicity endpoints. Due to their relevance to drug design and discovery, relatively large data sets have been collected in the literature for these problems, which provides a way to validate the representability of proposed topological descriptors along with machine learning algorithms. Certainly, to overcome the difficulty of predicting datasets with small training set for certain problems, we construct topological learning by integrating ESPH and multitask deep learning. We show that ESPH provides a competitive description of relatively small drug-like molecules and MT-DNN architecture is capable of promoting model performances on related tasks.

CHAPTER 2

TOPOLOGICAL MODELING OF SMALL MOLECULES

In this chapter, we will focus on our approach to biomolecules modeling using topological tools. First we briefly review the background of persistent homology, then we will introduce the so-called element specific persistent homology (ESPH) and its modified version. Several concrete examples will be given to illustrate the motivation and predictive power of ESPH. Lastly the essential idea of element specific topological descriptors (ESTD) and how ESTDs are computed will be discussed.

2.1 Persistent homology

For atomic coordinates in a molecule, algebraic groups can be defined via simplicial complexes, which are constructed from simplices, i.e., generalizations of the geometric notion of nodes, edges, triangles, tetrahedrons, etc. Homology associates a sequence of algebraic objects to topological spaces, and characterizes the topological connectivity of geometric objects in terms of topological invariants, i.e., Betti numbers. Betti-0, Betti-1 and Betti-2, represent the number of isolated connected components, rings and cavities respectively. A filtration parameter, such as the radius of a ball, is used to continuously vary over an interval so as to generate a family of structures. Loosely speaking, the corresponding family of homology groups induced by the filtration is a persistent homology. The variation of the topological invariants, i.e., Betti numbers, gives rise to a unique characterization of physical objects such as protein complex and small molecules.

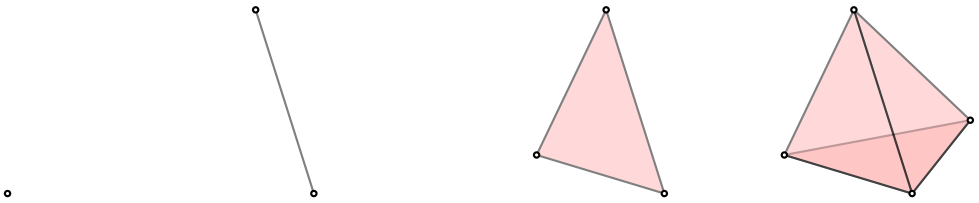
Simplex Let $u_0, u_1, \ldots, u_k$ be a set of points in $\mathbb{R}^d$. A point $x = \sum_{i=0}^{k} \lambda_i u_i$ is called an *affine combination* of the u_i if $\sum_{i=0}^{k} \lambda_i = 1$. The $k+1$ points are said to be *affinely independent*, if and only if $u_i - u_0$, $1 \leq i \leq k$ are linearly independent. We can find at most d linearly independent vectors and at most $d+1$ affinely independent points in $\mathbb{R}^d$.

An affine combination, $x = \sum_{i=0}^{k} \lambda_i u_i$ is a *convex combination* if λ_i are nonnegative. A *k-simplex*, which is defined to be the *convex hull* (the set of convex combinations) of $k+1$ affinely

independent points, can be formally represented as

$$\sigma = \left\{ \sum_{i=0}^{k} \lambda_i u_i | \sum \lambda_i = 1, \lambda_i \geq 0, i = 0, 1, ..., k \right\}, \tag{2.1}$$

where $\{u_0, u_1, ..., u_k\} \subset \mathbb{R}^d$ is a set of affinely independent points. Examples of k-simplex for the first few dimensions are shown in Figure 2.1. Essentially, a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, and a 3-simplex is a tetrahedron. A *face* τ of σ is the convex hull of



(a) 0-simplex (b) 1-simplex (c) 2-simplex (d) 3-simplex

Figure 2.1: Examples of simplex of different dimensions. (a), (b), (c) and (d) above represent 0-simplex, 1-simplex, 2-simplex, and 3-simplex, respectively.

a non-empty subset of u_i and is *proper* if the subset does not contain all $k + 1$ points. Equivalently, we can write as $\tau \leq \sigma$ if τ is a face or σ, or $\tau < \sigma$ if τ is proper. The *boundary* of σ, is defined to be the union of all proper faces of σ.

Simplicial complex A *simplicial complex* is a finite collection of simplices K such that $\sigma \in K$ and $\tau \leq \sigma$ implies $\tau \in K$, and $\sigma, \sigma_0 \in K$ implies $\sigma \cap \sigma_0$ is either empty or a face of both. The *dimension* of K is defined to be the maximum dimension of its simplices.

Chain complex Given a simplicial complex K and a constant p as dimension, a *p-chain* is a formal sum of p-simplices in K, denoted as $c = a_i \sigma_i$. Here σ_i are the p-simplices and the a_i are the coefficients, mostly defined as 0 or 1 (module 2 coefficients) for computational considerations. Specifically, p-chains can be added as polynomials. If $c_0 = \sum a_i \sigma_i$ and $c_1 = \sum b_i \sigma_i$, then $c_0 + c_1 = \sum (a_i + b_i)\sigma_i$, where the coefficients follow $\mathbb{Z}_2$ addition rules. The p-chains with the previously defined addition form an Abelian group and can be written as $(C_p, +)$. A *boundary*

operator of a p-simplex σ is defined as

$$\partial_p \sigma = \sum_{j=0}^{p} (-1)^j [u_0, u_1, ..., \widehat{u_j}, ..., u_p], \tag{2.2}$$

where $[u_0, u_1, ..., \widehat{u_j}, ..., u_p]$ means that vertex u_j is excluded in computation. Given a p-chain $c = a_i \sigma_i$, we have $\partial_p c = \sum a_i \partial_p \sigma_i$. Notice that ∂_p maps p-chain to $\{p-1\}$-chain and that boundary operation commutes with addition, a *b*oundary homomorphism $\partial_p : \sigma_p \to \sigma_{p-1}$ can be defined. The chain complex can be further defined using such boundary homomorphism as following:

$$\cdots \longrightarrow C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \xrightarrow{\partial_{p-1}} \cdots \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0. \tag{2.3}$$

Cycles and boundaries A p-cycle is defined to be a p-chain c with empty boundary ($\partial_p c = 0$), and the group of p-cycles of K is denoted as $Z_p = Z_p(K)$. In other words, Z_p is the kernel of the p-th boundary homomorphism, $Z_p = \ker \partial_p$. A p-boundary is a p-chain, say c, such that there exists $d \in C_{p+1}$ and $\partial_p d = c$, and the group of p-boundaries is written as $B_p = B_p(K)$. Similarly, we can rewrite B_p as $B_p = \mathrm{im}\partial_{p+1}$ since the group of p-boundaries is the image of the $(p+1)$-st boundary homomorphism.

Homology groups The fundamental lemma of homology says that the composition operator $\partial_p \circ \partial_{p+1}$ is a zero map [103]. With this lemma, we conclude that $\mathrm{im}\partial_{p+1}$ is a subgroup of $\ker \partial_p$. Then the *p-th homology group* of simplicial complex is defined as the p-th cycle group modulo the p-th boundary group,

$$H_p = Z_p / B_p \tag{2.4}$$

and the p-th Betti number is the rank of this group, $\beta_p = \mathrm{rank} H_p$. Geometrically, Betti numbers can be used to describe the connectivity of given simplicial complexes. Intuitively, β_0, β_1 and β_2 are numbers of connected components, tunnels, and cavities, respectively, for the first few Betti numbers.

Filtration and persistence A filtration of a simplicial complex K is a nested sequence of subcomplexes of K.

$$\varnothing = K_0 \subseteq K_1 \subseteq ... \subseteq K_n = K. \tag{2.5}$$

For each $i \leq j$, there exists an inclusion map from K_i to K_j and therefore an induced homomorphism $f_p^{i,j} : H_p(K_i) \rightarrow H_p(K_j)$ for each dimension p. The filtration defined in Equation (2.5) thus corresponds to a sequence of homology groups connected by homomorphisms.

$$0 = H_p(K_0) \rightarrow H_p(K_1) \rightarrow \cdots \rightarrow H_p(K_n) = H_p(K) \tag{2.6}$$

for each dimension p. The p-th persistent homology groups are defined as the images of the homomorphisms induced by inclusion,

$$H_p^{i,j} = \mathrm{im} f_p^{i,j} \tag{2.7}$$

where $0 \leq i \leq j \leq n$. In other words, $H_p^{i,j}$ contains the homology classes of K_i that are still alive at K_j for given dimension p and each pair i, j. We can reformulate the p-th persistent homology group as

$$H_p^{i,j} = Z_p(K_i) / \Big(B_p(K_j) \cap Z_p(K_i) \Big). \tag{2.8}$$

The corresponding p-th persistent Betti numbers are the ranks of these groups, $\beta_p^{i,j} = \mathrm{rank} H_p^{i,j}$. The birth, death and persistence of a Betti number carry important chemical and/or biological information, which is the basics of the present method.

2.2 Persistent homology for characterizing molecules

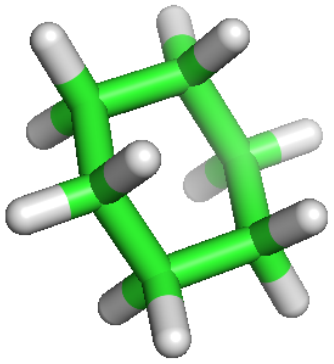
As introduced before, persistent homology indeed reveals long lasting properties of a given object and offers a practical method for computing topological invariants of a space, which captures the underlying features of the object directly from discrete point cloud data. An intuitive way to construct simplicial complex from point cloud data is to adopt Euclidean distance, or to use Vietoris-Rips complex with Euclidean distance. Vietoris-Rips complex is defined to be a simplicial complex whose k-simplices correspond to unordered $(k + 1)$-tuples of points which are pairwise within distance ϵ.

However, a particular radius ϵ is not sufficient since it is difficult to see if a hole is essential. Therefore, it is necessary to increase radius ϵ systematically, and see how the homology groups and Betti-numbers evolve. The persistence [103, 71] of each Betti number over the filtration can be recorded in barcodes [104, 105]. The persistence of topological invariants observed from barcodes offers an important characterization of small molecular structures. For instance, given the 3D coordinates of a small molecule, a short-lived Betti-0 bar may be the consequence of a strong covalent bond while a long-lived Betti-0 bar can indicate a weak covalent bond. Similarly, a long-lived Betti-1 bar may represent a chemical ring.

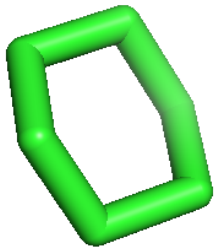
Such observations motivate us to design persistent homology based topological descriptors (TDs). However, it is important to note that the filtration radius is not a chemical bond and topological connectivity is not a physical relationship. In other words, persistent homology offers a representation of molecules that is entirely different from classical theories of chemical and/or physical bonds. Such a representation is systematical and comprehensive, and thus is able to unveil structure-activity relationships when it is coupled with advanced machine learning algorithms.

An example of PH Figure 2.2 is a detailed example of how persistent homology can be applied to a simple molecule – cyclohexane. An all-element representation of cyclohexane is given in Fig 2.2a, where carbon atoms are in green and hydrogen atoms are in white. As we can see from its barcodes in Fig. 2.2c, there are 18 Betti-0 bars that correspond to 18 atoms at the very beginning, 12 of which disappear when the filtration value increases to 1.08Å. It indicates that each carbon atom has merged with its closest 2 hydrogen atoms as the filtration value becomes larger than the length of C-H bond and these three atoms are regarded as one single connected component. When the filtration value further increases to 1.44Å, a Betti-1 bar emerges which means that a hexagonal carbon ring is captured and there is only one connected component left. As the filtration value eventually exceeds the radius of the hexagon, the ring structure disappears and the Betti-1 bar dies. The longest Betti-0 bar corresponds to the existence of the connected component. When only carbon atoms are selected, it is relatively straightforward to interpret the barcode plot. The cutoff

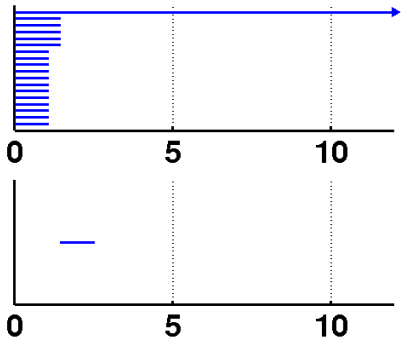
where 5 Betti-0 bars disappear corresponds to the C-C bond length and the Betti-1 bar represents the existence of the hexagonal carbon ring.



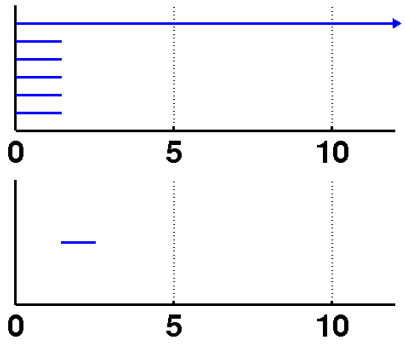
(a) Cyclohexane



(b) Cyclohexane with only carbon atoms



(c) Barcodes of cyclohexane



(d) Barcodes of cyclohexane with only carbon atoms

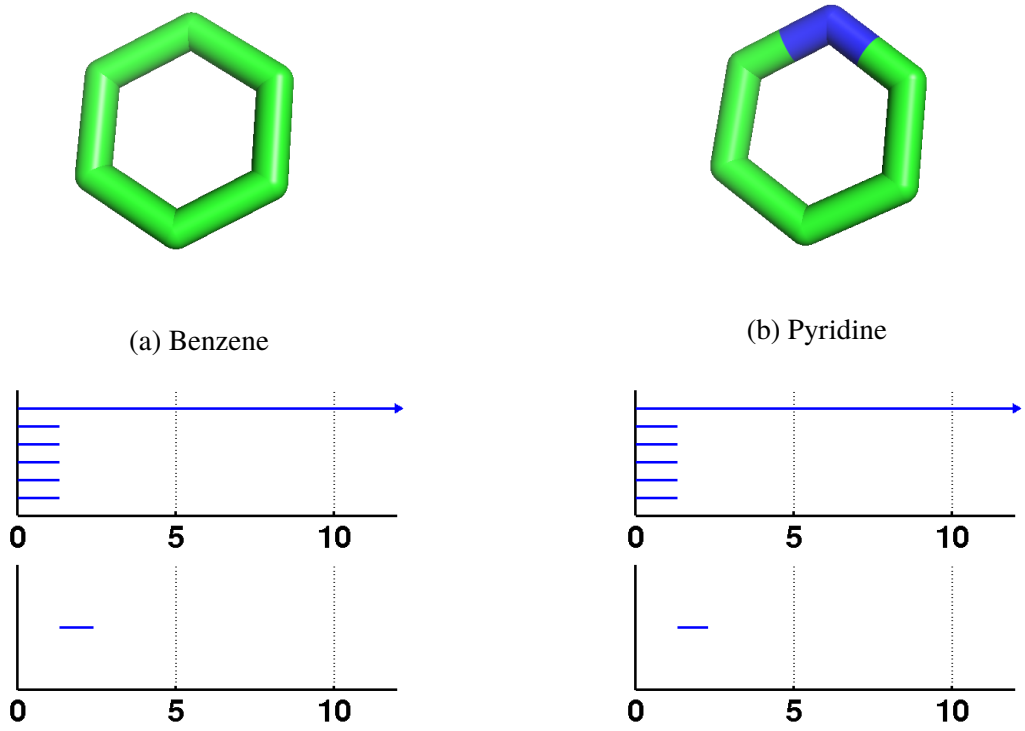
Figure 2.2: Different representations of cyclohexane and their persistent homology barcode plots. In subfigure(a) and (b), complete cyclohexane and cyclohexane with only carbon atoms being selected, respectively. In subfigure (c) and (d), from top to bottom, the results are for Betti-0 and Betti-1, respectively.

2.3 Element specific persistent homology (ESPH)

2.3.1 Limitations of persistent homology

Persistent homology, as discussed before, is efficient at characterizing covalent bonding or chemical structures of higher dimensions. Nevertheless, such information is not sufficient under most circumstances, especially for small molecules. For instance, it is not possible to distinguish a carbon-nitrogen ring from a all-carbon ring as primitive persistent homology can only capture the

persistence of a Betti-1 bar whereas there is no indication whether a nitrogen atom exists. Figure 2.3 shows why primitive persistent homology has limitations when dealing with some small molecules of similar structures.



(a) Benzene

(b) Pyridine

(c) Barcodes of benzene

(d) Barcodes of pyridine

Figure 2.3: Benzene, pyridine and their persistent homology barcode plots. In subfigure(a) and (b), benzene and pyridine are shown with hydrogen atoms being neglected, respectively. In subfigure (c) and (d), from top to bottom, the results are for Betti-0 and Betti-1, respectively, for benzene and pyrdine.

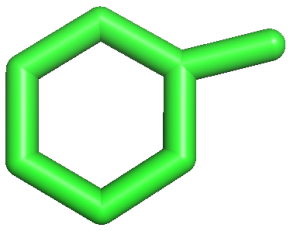
As shown in Fig 2.3a and Fig 2.3b, both benzene and pyridine have a hexagonal ring except that pyridine has a nitrogen atom. From primitive persistent homology point of view, there is no difference between these two molecules except a slight difference between the lengths of Betti bars – the Betti-1 bar of benzene has length 1.08 Åwhile pyridine's Betti-1 bar has length 0.96 Å– which is caused by the fact that the carbon-carbon bond is generally longer than carbon-nitrogen bond.

However, Fig 2.3c and Fig 2.3d follow a very similar pattern. Both barcodes contain 6 Betti-0 bars and 1 Betti-1 bar, and it is nearly impossible for us to distinguish these two molecules purely

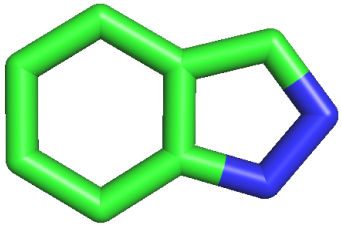
from barcodes calculated from their structures. Therefore it is necessary for us to introduce the idea of ESPH, where persistent homology is computed based on different combinations of specific element types.

2.3.2 An introduction to ESPH

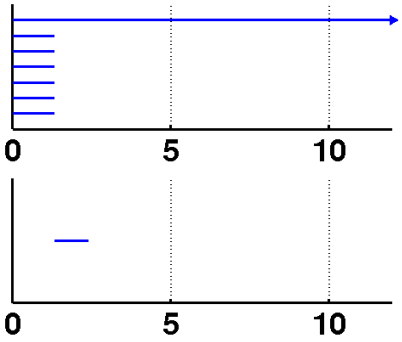
An example of ESPH representation Figure 2.4 depicts how ESPH modeling can be applied to small molecules. In the following case, indazole (PubChem id: 9221) is chosen. For simplicity, hydrogen atoms are neglected. Apparently there are two chemical rings within the indazole molecule - one hexagonal carbon ring and one pentagonal carbon-nitrogen ring.



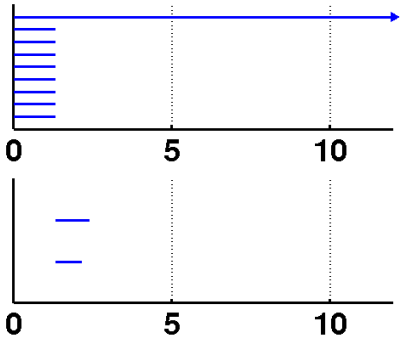
(a) C representation of indazole



(b) C-N representation of indazole



(c) Barcodes of the C representation



(d) Barcodes of the C-N representation

Figure 2.4: Indazole and its persistent homology barcodes. In subfigure(a) and (b), indazole is shown with carbon and carbon-nitrogen atoms selected, respectively. In subfigure (c) and (d), from top to bottom, the results are for Betti-0 and Betti-1, respectively

Mathematically if we ignore element types, there are two 1-simplices (loops) for indazole. We have to notice, however, that the properties of these two rings are dramatically different and

they have to be handled very carefully. Thus if only carbon atoms are selected, we can observe 7 Betti-0 bars and 1 Betti-1 bar, which correspond to 7 carbon atoms and the hexagonal carbon ring, respectively. While carbon and nitrogen atoms are selected, there are 9 Betti-0 bars, 1 short Betti-1 bar and 1 longer Betti-0 bar. By comparing these two barcodes, we may conclude the following:

- The difference between number of Betti-0 bars represents the number of nitrogen atoms (2 nitrogen atoms).
- There exists a carbon-nitrogen ring, and its size is smaller than the previous carbon ring. Notice that the carbon-nitrogen ring is not captured when only carbon atoms are selected, and its Betti-1 bar is shorter than the other Betti-1 bar that corresponds to the carbon ring.

If we further consider the length of Betti-0 and Betti-1 bars, it is possible for us to find out the length of chemical bonds. For example, in Fig 2.4a, the length of the 6 shorter Betti-0 bars is 1.32 Å and it indicates that the carbon-carbon bond has length 1.32 Å for this particular molecule, while generally the length of carbon-carbon bonds falls within the range 1.20 Å– 1.54 Å [106]. Thus it is reasonable to apply ESPH to a wider range of applications, especially when bond lengths and number of Betti bars are taken into account.

2.4 Modified ESPH for characterizing intra-molecular interactions

Another important component of ESPH is the filtration matrix that defines the distance in persistent homology analysis. Traditionally, the distance between atom i at (x_i, y_i, z_i) and atom j at (x_j, y_j, z_j) is defined to be the Euclidean distance between them:

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}. \tag{2.9}$$

Indeed, by using Euclidean distance defined in Eq. 2.9, persistent homology is able to capture the information such as covalent bonds between different atom types easily. However, it does not necessarily reflect intramolecular interactions such as hydrogen bonding and van der Waals interaction, which is not ideal for the purpose of small molecule modeling. In other words, the Betti-0 bar between two atoms with certain hydrogen bonding or van der Waals cannot be

captured since there already exists shorter Betti-0 bar between them (essentially covalent bonds). To circumvent such deficiencies, we redefine the distance between atom i and atom j to be:

$$M_{i,j} = \begin{cases} d_{i,j}, & \text{if} \quad d_{i,j} \geq r_i + r_j + |\Delta d| \\ d_\infty, & \text{otherwise,} \end{cases} \tag{2.10}$$

where r_i and r_j are the atomic radius of atom i and j, respectively. Here Δd is the bond length deviation in the data set and d_∞ is a large number which is set to be greater than the maximal filtration value. Since the distance between two atoms with covalent bonds can never exceed the preset maximum filtration value, we are able to use such modified ESPH to capture important intramolecular interactions, since covalent bonds can never be built.

CHAPTER 3

MACHINE LEARNING METHODS AND ALGORITHMS

In this chapter, we will give an overview of machine learning algorithms and multi-task deep learning architectures used in this study.

3.1 An overview of machine learning algorithms

The concept of machine learning was first proposed by Arthur Samuel [107]. Machine learning algorithms can learn from and make predictions on given data, and have the potential to overcome complicated computational problems when explicit solutions are difficult to determine.

Basically speaking, machine learning algorithms can be classified into three different categories – supervised learning, unsupervised learning and reinforcement learning. The features of each category can be summarized below:

- Supervised learning: Each sample data in training set consists of a target value (categorical for classification and continuous for regression) and a given set of (independent) descriptors. The purpose of supervised learning is to learn a function that map inputs to desired outputs. The training process continues until the model reaches a predefined level of accuracy on the training data.

- Unsupervised learning: The difference between unsupervised learning and supervised learning is that there is no target value for each training sample and there is no evaluation of the accuracy of output. The purpose of unsupervised learning is to perform clustering for population in different groups.

- Reinforcement learning: Reinforcement learning is trained to make specific decisions and is typically formulated as Markov decision process. During learning process, correct input/output pairs are never presented, nor sub-optimal actions are explicitly corrected. The machine is exposed to an environment where it trains itself continually using trial and error,

and eventually finds a balance between exploration (of uncharted territory) and exploitation (of current knowledge) [108].

In this thesis, small molecule properties to be predicted are all quantitative, therefore we will employ supervised learning algorithms to perform training and testing. More precisely, ensemble methods and supervised deep neural networks will be discussed in the next sections.

3.2 Ensemble methods

In this section, we will first review decision tree induction. Then we will also introduce several ensemble methods that are essentially based on decision tree algorithms, including both random forests and gradient boosting decision tree algorithms.

3.2.1 Decision tree induction

3.2.1.1 Basics of decision tree induction

For a decision tree model, there are three types of nodes: root node, internal nodes, and terminal nodes. The model solves problem by answering a series of questions at each node and returns a conclusion when a terminal node is reached.

Hunt's algorithm Hunt's algorithm [109] is the basis of many existing decision tree induction algorithms, where a decision tree is grown in a recursive fashion by partitioning the training records into successive subsets. A general procedure for Hunt's algorithm is described below:

1. If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
2. If D_t contains records that belong to more than one class, an **attribute test condition** is used to split records into smaller subset. The procedure continues until all the records in the subset belong to the same class.

Attribute test condition As mentioned in previous paragraph, there are different attribute test conditions depending on attribute types. For instance, binary attributes output two potential outcomes, while continuous attributes use comparison test to express the test condition. Thus it is of great significance to select proper measures to evaluate different splits at each node.

Measures for selecting best split The measures for selecting the best split are based on the degree of impurity of the child nodes. Generally for classification tasks, some typical measures include:

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t) \tag{3.1}$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 \tag{3.2}$$

$$\text{Information Gain}(t) = \text{Entropy}(t) - \text{weighted entropy of t's child nodes} \tag{3.3}$$

where t is a node of the decision tree, c is the number of classes, $p(i|t)$ is the probability, and weighted entropy of all child nodes of node t can be calculated as the weighted average of child node entropies where the weight of each child node is the number of records at that child node divided by the number of records at the parent node c. To select the best split, impurity measures for each candidate split are calculated and the best split can be chosen from all candidates.

For regression tasks, variance reduction can be used to split nodes. A commonly used measure for variance reduction is analogous to information gain for classification in Eq. 3.3, and can be written as:

$$\text{Variance Reduction}(t) = \text{Variance}(t) - \text{weighted variance of t's child nodes} \tag{3.4}$$

The weight is calculated in the same way as that of information gain.

Pruning of decision tree Decision trees, however, are often susceptible to overfitting, in the sense that trained decision trees are so closely fit to training data that models can result in substantial

errors for unseen data as training data usually has some degree of error or random noise within it. Pruning is a strategy to reduce the size of decision tree.

One way to perform pruning is to apply **early stopping rule** during tree-training process. Specifically, the node stops expanding to child nodes when the observed gain in impurity measure falls below a predefined threshold, which helps to avoid constructing overly complicated subtrees that may cause overfitting issues [109]. However, the threshold is difficult to determine [109].

The other strategy is to perform **post-training pruning**. This can be done by replacing a subtree with 1) a new child node whose class is determined by majority class of records associated with the subtree, or 2) the most frequently used branch of the subtree. The pruning process stops until no further improvement can be observed.

3.2.2 Random forest

Random forest is an ensemble machine learning algorithm that can be used for regression and classification. It learns training data by constructing a multitude of decision trees, and returns prediction by averaging the outputs of individual trees (regression) or by taking the majority vote of individual trees (classification). Typically random forest does not overfit the training data and it is capable of reducing variance while maintaining the same level of bias [110] when compared with traditional decision tree models.

3.2.2.1 Bootstrap aggregating

Random forest takes advantage of the bootstrap aggregating techniques when building individual tree learners. Given a training set $\{(x_i, y_i)\}_{i=1}^{n}$ and the number of trees N, a bootstrap aggregating process is to repeatedly select a random sample $S^{(j)}$ with replacement from the training set and fit each sample with individual tree f_j ($j = 1, \ldots, N$). Eventually for regression problems, the prediction $\hat{f}$ for any unknown data x' shall be given by:

$$\hat{f} = \frac{1}{N} \sum_{j=1}^{N} f_j(x') \tag{3.5}$$

For classification problems, the predicted class $\hat{f}$ is the class that the majority of trees vote for.

3.2.2.2 Feature bagging

Random forest algorithm differs from the aforementioned bootstrap aggregating technique except that a random feature subset is used for tree splitting in random forest, or in other words, feature bagging is used to determine best splits.

3.2.3 Gradient boosting decision tree

3.2.3.1 Gradient boosting algorithm

Gradient boosting algorithm was first observed by Leo Breiman [111], and was subsequently developed by Friedman [112]. Gradient boosting algorithm can also be viewed as a iterative functional gradient descent algorithm [113, 114]. The idea of this algorithm is to iteratively find a series of weighted weak learners and eventually form a strong learner which can be expressed as the summation of weak learners.

3.2.3.2 Gradient tree boosting

Gradient tree boosting is a combination of aforementioned gradient boosting and decision trees of fixed size. Specifically decision trees are used as base learners to fit pseudo-residuals $h_m(x)$ at each iteration step m.

3.3 Multi-task learning and deep neural network

3.3.1 Single-task deep neural network (ST-DNN)

A neural network acts as a transformation that maps an input feature vector to an output vector. It essentially models the way a biological brain solves problems with numerous neuron units connected by axons. A typical shallow neural network consists of a few layers with neurons and uses back propogation to update weights on each layer. However, it is not able to construct hierarchical

features and thus falls short in revealing more abstract properties, which makes it difficult to model complex non linear relationships.

A single-task deep learning algorithm, compared to shallow networks, has a wider and deeper architecture – it consists of more layers and more neurons in each layer and reveals the facets of input features at different levels. Single-task deep learning algorithm is defined for each individual prediction task and only learns data from the specific task. A representation of such single task deep neural network (ST-DNN) can be found in Figure 3.1, where n represents the number of layers of a given ST-DNN, k_i and N_{k_i} $(i = 1, \ldots, n)$ is the number of neurons and node of i-th hidden layer, respectively.

Generally speaking, the objective of such a ST-DNN is to minimize a given loss function, which is essentially based on problems that one is trying to solve – such a loss function can be defined as cross-entropy loss function for a multi-class classification problem, or mean squared error function for a regression problem.

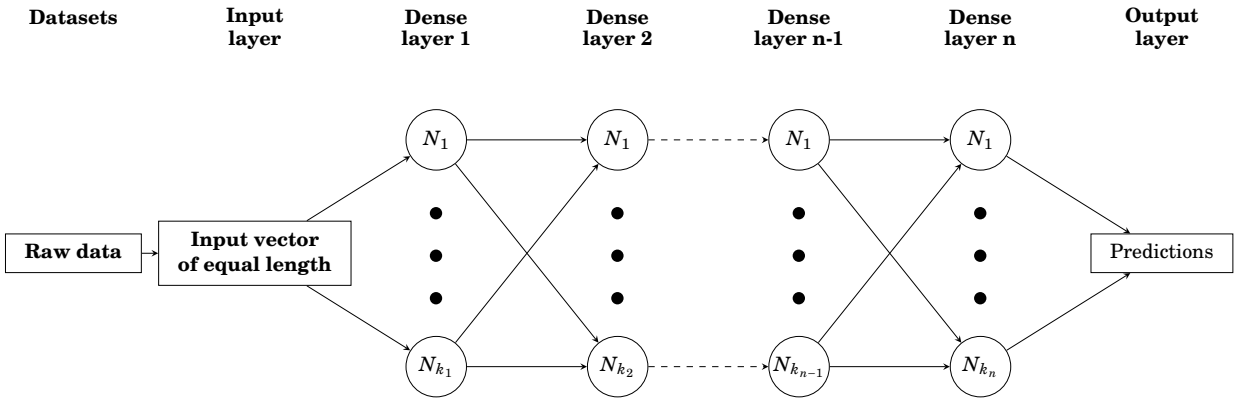


Figure 3.1: An illustration of ST-DNN architecture.

3.3.1.1 Multi-task deep neural network

Multi-task learning is a machine learning technique which has shown success in various fields. The main advantage of MT learning is to learn multiple tasks simultaneously and exploit commonalities as well as differences across different tasks. Another advantage of MT learning is that a small data

set with incomplete statistical distribution to establish an accurate predictive model can often be significantly benefited from relatively large data sets with more complete statistical distributions.

Suppose we have a total of T tasks and the training data for the t-th task are denoted as $(\mathbf{x}_i^t, y_i^t)_{i=1}^{N_t}$, where $t = 1, \ldots, T, i = 1, \ldots, N_t$, N_t is the number of samples of the t-th tasks, with $\mathbf{x}_i^t$ and y_i^t being the topological descriptor vector that consists of ESTDs and the target value of the i-th molecule in t-th task, respectively. The goal of MTL is to minimize the following loss function for all tasks simultaneously:

$$\operatorname{argmin} \sum_{i=1}^{N_t} L(y_i^t, f^t(\mathbf{x}_i^t; \{\mathbf{W}^t, \mathbf{b}^t\})) \tag{3.6}$$

where f^t is a functional of the topological descriptor vector $\mathbf{x_i^t}$ parametrized by a weight vector $\mathbf{W}^t$ and bias term $\mathbf{b}^t$, and L is the loss function. A typical cost function for quantitative regression is the mean squared error, thus the loss of the t-th task can be defined as:

$$\text{Loss of Task } t = \frac{1}{2} \sum_{i=1}^{N_t} L(\mathbf{x}_i^t, y_i^t) = \frac{1}{2} \sum_{i=1}^{N_t} (y_i^t - f^t(\mathbf{x}_i^t; \{\mathbf{W}^t, \mathbf{b}^t\})^2 \tag{3.7}$$

To avoid overfitting problem, it is usually beneficial to customize above loss function (3.7) by adding a regularization term on weight vectors, giving us an improved loss function for t-th task:

$$\text{Loss of Task } t = \frac{1}{2} \sum_{i=1}^{N_t} (y_i^t - f^t(\mathbf{x}_i^t; \{\mathbf{W}^t, \mathbf{b}^t\})^2 + \beta ||\mathbf{W}^t||_2^2 \tag{3.8}$$

where $|| \cdot ||$ denotes the L_2 norm and β represents a penalty constant.

The goal of topology based MTL is to learn different small molecule properties jointly, and to potentially improve the overall performances of multiple models simultaneously. More concretely, it is reasonable to assume that different small molecules comprise distinct physical/chemical features, while descriptors such as the occurrence of certain chemical structure, can result in similar physical/chemical properties. A simple representation of multitask deep neural network (MT-DNN) for our study is shown in Figure 3.2, where k_i $(i = 1, \ldots, n)$ represents the number of neurons on the i-th hidden layer, N_{k_i} are neurons on i-th layer, and $\text{Prd}_1, \ldots, \text{Prd}_t$ represent predicted values for t different tasks.

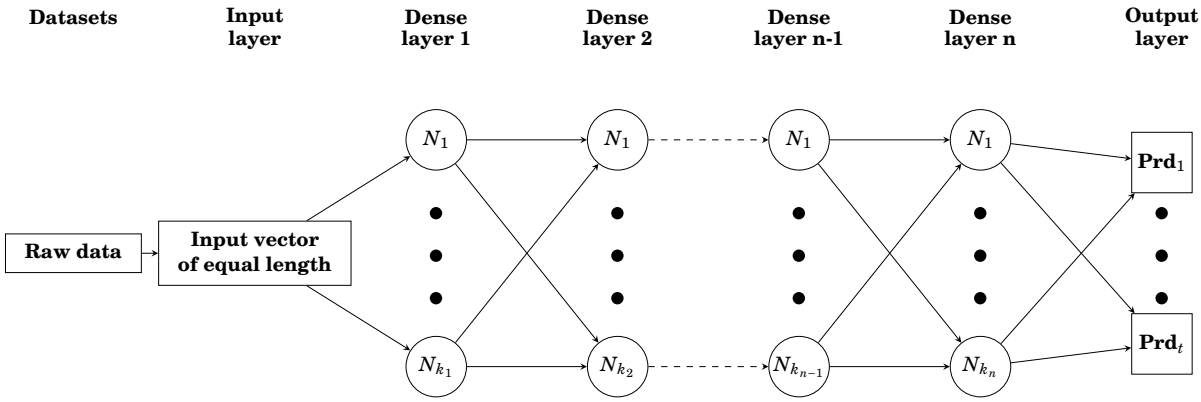


Figure 3.2: An illustration of MT-DNN architecture.

3.3.1.2 Network parameters and training

The hyperparameters tuning of DNN is known to be very complicated. In order to come up with a reasonable set of hyperparameters, we perform a grid search of each hyperparameter within a wide range. Hyperparameters in Table 3.1 are chosen so that we can have a reasonable training speed and accuracy. It turns out that adding dropout or L^2 decay does not necessarily increase the accuracy and as a consequence we omit these two techniques. The underlying reason may be that the ensemble results of different DNN models is essentially capable of reducing bias from individual predictions. A list of hyperparameters used to train all models can be found in Table 3.1

Table 3.1: Proposed hyperparameters for MT-DNN

Number of epochs	1000
Number of hidden layers	7
Number of neurons on each layer	1000 for first 3 layers, and 100 for the next 4 layers
Optimizer	ADAM
Learning rate	0.001

In each training epoch, molecules in each training set are randomly shuffled and then divided into mini-batches of size 200, which are then used to update parameters. When all mini-batches are traversed, an training “epoch" is done. All the training processes were done using Keras wrapper [115] with Theano (v0.8.2) [116] as the backend. All training were run on Nvidia Tesla K80 GPU and the approximate training time for a total of 1000 epochs is about 80 minutes.

3.4 Pipelines for predicting properties of small molecules using MT-DNN

In this section, we provide graphical pipelines for predicting properties of small molecules to help readers understand how our MT-DNN architecture works along with ESPH. In Fig. 3.3, a general procedure for predicting log P and log S simultaneously is presented. Given any molecule, we use ESPH to extract information such as intra-molecular interactions and geometrical connections from the molecule and construct ESTDs accordingly. Notice that the numbers of ESTDs for each molecule are the same, thus they can be directly fed into a MT-DNN architecture, where joint tasks can be learned and predicted simultaneously.

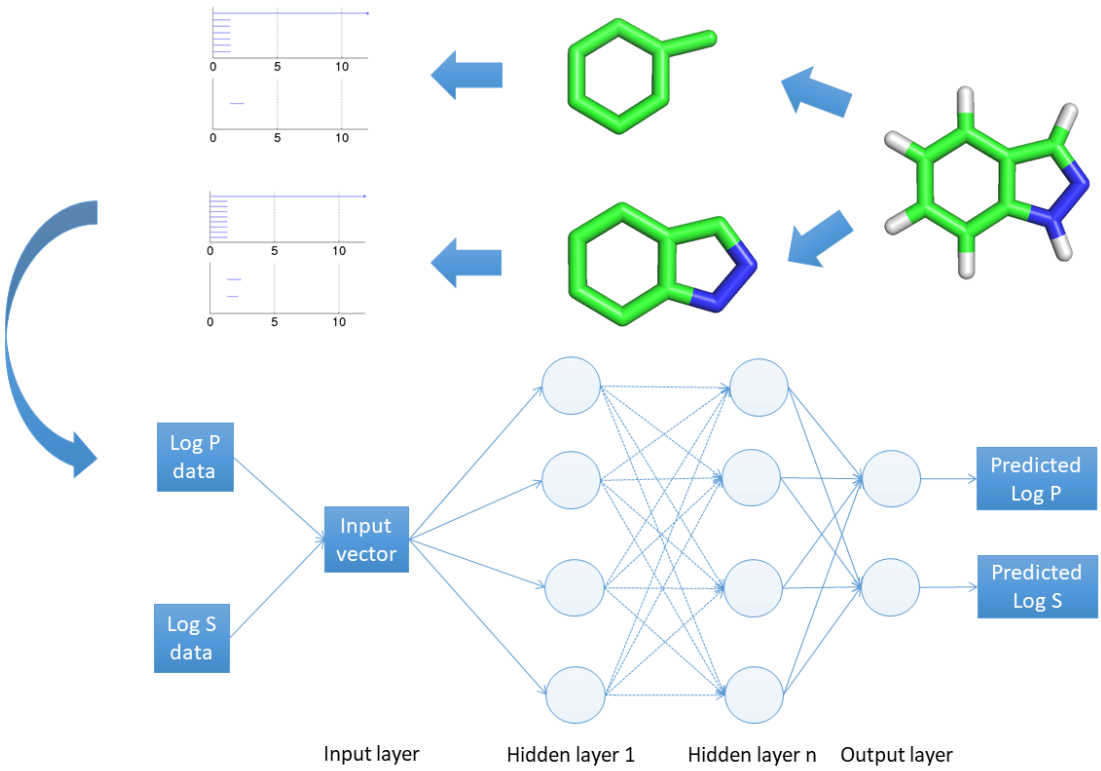


Figure 3.3: Graphical pipeline for simultaneous prediction of partition coefficient and aqueous solubility

3.5 Ranking, query construction and prediction using nearest neighbors

The algorithms in this section are specifically used for solvation free energy, as a continuation work of Ref [4]. A generally procedure for solvation free energy prediction can be summarized below:

1. Molecules in training set are divided into different queries based on element types, and molecules are ranked within each query.

2. Given any molecule in test set, first determine the query that the molecule belongs to, then use trained models to determine the order of the molecule within that specific query.

3. Finally, a predefined number of nearest neighbors (molecules) based on the ranking are selected and used for solvation free energy prediction of the target molecule.

3.5.1 Ranking algorithms

The essentially idea of ranking algorithms is to train a list of data points with some partial orders (either numerical or ordinal scores) so that the learners are able to predict the order of an unseen item with respect to the training data. Apparently if we use solvation free energy as a numerical score for each small molecule, the ranking algorithm can be directly applied to solvation free energy prediction. In this study, GBDT is used to rank a list of molecules.

3.5.2 Query construction

Query construction is an essential step for accurate prediction of solvation free energy. We follow the same principle in Ref [117]. Basically, seven groups of molecules are constructed according to element types: i) H, C; ii) H, C, O; iii) H, C, N/H, C, N, O; iv) H, C, Cl; v) H, C, O, Cl; vi) H, S; and vii) anything else, respectively. Detailed information of different queries can be found in Appendix.

3.5.3 Prediction using nearest neighbors

After a given molecule is fed into trained models and the order of the molecule within the query is returned, we select a number of nearest neighbors to predict its solvation free energy. Let m be the number of nearest neighbors. The purpose of local linear regression is to determine a weight vector $\mathbf{w} := (w_1, \ldots, w_n)^T$ and bias b, such that the training error on nearest neighbors can be minimized.

Mathematically, the problem can be formulated in matrix multiplication form:

$$\begin{pmatrix} \Delta G_1 \\ \Delta G_2 \\ \vdots \\ \Delta G_m \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} + \begin{pmatrix} b \\ b \\ \vdots \\ b \end{pmatrix}. \tag{3.9}$$

or equivalently

$$\Delta \mathbf{G} = \mathbf{X}\mathbf{w} + b\mathbf{1}, \tag{3.10}$$

where $\Delta \mathbf{G} = (\Delta G_1, \Delta G_2, \cdots, \Delta G_m)^T$ are experimental solvation free energy for m nearest neighbors, $\mathbf{w} = (w_1, w_2, \cdots, w_n)^T$, $\mathbf{1}$ is a column unit vector of length m, and matrix $\mathbf{X}$ contain descriptor vector for these m nearest neighbors and is written as:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}.$$

Since m is relatively small comparing to n, such local linear model is likely to overfit. Thus a L_2 penalty term can be added to training error, and thus Eq. (3.10) can be viewed as the following optimization problem

$$\arg\min_{\mathbf{w},b} \left(||\Delta \mathbf{G} - \mathbf{X}\mathbf{w} - b\mathbf{1}||_2^2 + \lambda ||\mathbf{w}||_2^2 \right) \tag{3.11}$$

where λ is the regularization parameter, which is set to 1000 in this work, $|| * ||_2$ denotes the L_2 norm of the quantity $*$.

Let $\frac{\partial \mathbf{F}}{\partial \mathbf{w}} = 0$, a direct computation returns:

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X} + \ \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\Delta\mathbf{G} - \mathbf{X}^T(b\mathbf{1})\right), \tag{3.12}$$

where I is $m \times m$ identity matrix.

Similarly, if we relax $b\mathbf{1}$ to arbitrary vector $\mathbf{b} = (b_1, b_2, \cdots, b_m)^T$ and let $\frac{\partial \mathbf{F}}{\partial \mathbf{b}} = 0$, we have

$$\mathbf{b} = \Delta\mathbf{G} - \mathbf{X}\mathbf{w}. \tag{3.13}$$

Thus the unbiased estimation of b can be written as

$$b = \frac{\sum_{i=1}^{m}(\Delta\mathbf{G} - \mathbf{X}\mathbf{w})_i}{m}, \tag{3.14}$$

where $(\Delta\mathbf{G} - \mathbf{X}\mathbf{w})_i$ is the ith component of the vector $\Delta\mathbf{G} - \mathbf{X}\mathbf{w}$.

With Eq. 3.12 and Eq. 3.14, we may solve the optimization problem 3.11 in an iterative manner. The iteration continues until the solution converges.

Let $\mathbf{x}' = (x'_1, \ldots, x'_n)$ be the descriptor vector of the target molecule and $\Delta\hat{G}$ be its predicted solvation free energy, we can now compute the solvation free energy of target molecule as below in Eq. 3.15, using $\mathbf{w}$ and b calculated from previous steps.

$$\Delta\hat{G} = \mathbf{x}'\mathbf{w} + b \tag{3.15}$$

CHAPTER 4

DATA SETS, PREPROCESSING AND DESCRIPTORS

In this chapter, we first introduce the data sets used to train and test quantitative models. Second, detailed descriptions of data preprocessing techniques for small molecules are discussed. Finally, we propose a variety of molecular descriptors, including element specific topological descriptors (ESTDs), physical descriptors and auxiliary descriptors. A detailed description of how they are calculated will also be provided.

4.1 Data sets

4.1.1 Solvation free energy

In an earlier work [4], a data set that contains a total of 668 molecules was proposed and it is the largest for solavtion free energies to the best of our knowledge. The data set contains both mono-functional group and polyfunctional group molecules. Experimental solvation free energies are collected from the literature [118, 119, 120]. All the structures of this dataset are downloaded from the PubChem project (`https://pubchem.ncbi.nlm.nih.gov/`). More detailed description of the dataset can be found in our earlier work [4].

SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) is a set of community-wide blind challenges aimed to advance computational techniques as standard predictive tools in rational drug design [121, 122, 123, 124, 125]. One major subject of SAMPL challenge for bind prediction is to predict solvation free energy for small molecules. In order to check how our models perform comparing to others', we make sure that all molecules of SAMPL0, SAMPL1, SAMPL2, SAMPL3 and SAMPL4, except for 5-iodouracil in SAMPL2, are properly processed and predicted. By excluding the SAMPLx molecules from the training set, we can use the remaining molecules from the 668 training set to independently predict the solvation free energy for SAMPLx molecules.

Moreover, we are also interested in knowing how our descriptors and algorithms can be extended

to other datasets. An external dataset [1] is used for validation purpose. Wang et al. [1] first introduced this dataset to evaluated their weighted solvent accessible surface areas based models. In Model III, the authors further divided 387 molecules (which exclude ions) into a training set (293) and a test set (94). In our study, a slightly smaller training set of 289 molecules is used because 4 molecules in the training set have ambiguous chemical names in the PubChem database, while all molecules in the test set are included in our prediction which enables us to compare to theirs on equal footing.

It should be noted that there exist discrepancies in experimental solvation free energies for some molecules in Ref [1] and the 668 set [4]. When such discrepancies occur, the experimental values reported by Wang et al [1] are used for training and testing in the above comparison. It should be noted that most experimental value differences are within a very small range. Only 23 differences are greater than 0.2 kcal/mol and 5 out of these 23 molecules are in the test set. These 5 compounds and their experimental values corresponding to Table 3 of Wang et al [1] are listed in Table 4.1. Additionally, 4 molecules listed in the training set of Ref. [1] are excluded in our training due to their absence of structures in PubChem. These 4 molecules have compound ID of 363, 364, 385 and 388 in Table 3 of Ref. [1].

Table 4.1: Molecules in the test sets with large discrepancies in their experimental solvation free energies. Here "ID" refers to the ID of Table 3 of Ref. [1]

ID	Exp1 [1]	Exp2 [4]
46	0.29	0.01
67	-3.15	-3.4
97	-0.78	-1.73
103	-0.64	-1.4
352	-4.71	-5.22

Moreover, we have also noticed that there are 11 duplicates in the training set and the test set of Ref. [1]. Their compound IDs and duplicated IDs (Dup-IDs) in the Table 3 of Ref. [1] are listed in Table 4.2. Molecules that are in the test set are marked with a superscript "*b*" to be consistent with the notation of Ref. [1]. Finally, we provide information about all datasets used for solvation free energy prediction in Table 4.3

Table 4.2: Duplicated molecules in Ref. [1]

ID	Dup-ID	ID	Dup-ID
104^b	119	333^b	335
334^b	336	384^b	389
161^b	202	82^b	84
140^b	142	184^b	194
97^b	116	58^b	59
196^b	203		

Table 4.3: Statistics of solvation free energy data sets. The numbers within parenthesis represent the actual numbers of molecules used in this study.

	Number of molecules		Number of molecules
SAMPL0 set	17	Wang's[1] train set	293 (289)
SAMPL1 set	63	Wang's[1] test set	94
SAMPL2 set	30 (20)		
SAMPL3 set	36		
SAMPL4 set	47		

4.1.2 Partition coefficient and aqueous solubility

The primary work of this thesis is to explore the proposed topology based multi-task methods for learning related tasks. Thus data sets can naturally be divided into two parts – one for partition coefficient prediction and the other for aqueous solubility prediction.

Partition coefficient data sets The training set used for partition coefficient prediction was originally compiled by *C*heng et al. [5] and consists of 8199 compounds, which is based on Hansch *e*t al.'s compilation [126]. These compounds are considered to have reliable experimental $\log P$ values by Hansch (marked with * or checkmark). In addition, three sets were chosen as test sets. The first test set, which is completely independent from the training set, contains 406 small-molecule organic drugs approved by the Food and Drug Administration (FDA) of the United States and represents a variety of organic compounds of pharmaceutical interests. This set was also compiled by *C*heng et al. [5]. The remaining two test sets, Star set and Non-star set, were publicly available and originated from a monograph of Avdeef [127]. Star set comprises 223 compounds that are part of BioByte Star set and have been widely used to develop $\log P$ prediction method.

The Non-star set contains 43 compounds that represent relatively new chemical structures and properties. The compound list and corresponding partition coefficient is available for download at `http://ochem.eu/article/17434`. We also made an attempt to expand our training set by searching the NIH database as other software packages use a large number of molecules for supervised learning. In this way, more than 3000 additional molecules were added to the training set.

Aqueous solubility data sets In order to develop and validate prediction models for aqueous solubility, several well-defined aqueous solubility datasets were used. Firstly, a diverse data set of 1708 molecules proposed by Wang *et al.* [128] was used to verify the predictive power of descriptors. Both leave-one-out and 10-fold cross-validation were carried out on this set. Furthermore, we also tested our models on a relatively small set with independent test sets [3]. As Hou [3] suggested, we also removed some molecules from the training set to ensure that training set and test set have no overlapping molecules.

In addition, two more widely used, publicly available solubility data sets are also used to train and evaluate our models. The first set is the 'small' Delaney data set [129] that contains 1144 molecules and their measured aqueous solubility (log mol/L at 25 degree Celcius. The second set was originally built by Huuksonen [130] from AQUASOL database [131] and PHYSPROP database [132]. It consists of 1026 organic molecules with their aqueous solubility in log mol/L at 20-25 degree Celcius.

A summary of data sets used for the proposed models is given in Table 4.4.

Table 4.4: Summary of log P and log S data sets used

logP data	Number of molecules	logS data	Number of molecules
logP train set	8199	logS train set 1 [128]	1708
FDA test set	406	logS train set 2 [3]	1290 (1207 for test set 2)
Star test set	223	logS test set 1 [3]	21
Nonstar test set	43	logS test set 2 [3]	120
		Huuskonen logS set	1033 (1030)
		Small delaney logS set	1144 (1135)

4.1.3 Quantitative toxicity endpoints

Four different quantitative toxicity datasets, anmely, 96 hour fathead minnow LC_{50} data set (LC_{50} set), 48 hour Daphnia magna LC_{50} data set (LC_{50}-DM set), 40 hour Tetrahymena pyriformis IGC_{50} data set (IGC_{50} set), and oral rat LD_{50} data set (LD_{50} set), are studied in this work. Among them, LC_{50} set reports at the concentration of test chemicals in water in mg/L that causes 50% of fathead minnow to die after 96 hours. Similarly, LC_{50}-DM set records the concentration of test chemicals in water in mg/L that causes 50% Daphnia maga to die after 48 hours. Both sets were originally downloadable from the ECOTOX aquatic toxicity database via web site `http://cfpub.epa.gov/ecotox/` and were preprocessed using filter criterion including media type, test location, etc [133]. The third set, IGC_{50} set, measures the 50% growth inhibitory concentration of Tetrahymena pyriformis organism after 40 hours. It was obtained from Schultz and coworkers [134, 135]. The endpoint LD_{50} represents the amount of chemicals that can kill half of rates when orally ingested. The LD_{50} was constructed from ChemIDplus databse (`http://chem.sis.nlm.nih.gov/chemidplus/chemidheavy.jsp`) and then filtered according to several criteria [133].

The data sets used in this work are identical to those that were preprocessed and used to develop the Toxicity Estimation Software Tool (T.E.S.T.) at `https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test` [133]. TEST was developed to estimate chemical toxicity using various QSAR methodologies and is very convenient to use as it does not require any external programs. It follows the general QSAR workflow — it first calculates 797 2D molecular descriptors and then predicts the toxicity of a given target by utilizing these precalculated molecular descriptors.

All molecular structures and their toxicity endpoints are available on the T.E.S.T. website (`https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test`) . It should be noted that we are particularly interested in predicting quantitative toxicity endpoints so other data sets that contain qualitative endpoints or physical properties were not used. Moreover, different toxicity endpoints have different units. The units of LC_{50}, LC_{50}-DM, IGC_{50} endpoints are $-\log_{10}(\mathrm{T\,mol/L})$, where T represents corresponding endpoint. For LD_{50} set, the units are

$-\log_{10}(\mathrm{LD}_{50}\,\mathrm{mol/kg})$. Although the units are not exactly the same, it should be pointed out that no additional attempt was made to rescale the values since endpoints are of the same magnitude order. These four data sets also differ in their sizes, ranging from hundreds to thousands, which essentially challenges the robustness of our methods. A detailed statistics of our datasets is presented in Table 4.5.

Table 4.5: Statistics of quantitative toxicity data sets

	Total # of mols	Train set size	Test set size	Max value	Min value
LC_{50} set	823	659	164	9.261	0.037
LC_{50}-DM set	353	283	70	10.064	0.117
IGC_{50} set	1792	1434	358	6.36	0.334
LD_{50} set	7413 (7403)	5931 (5924)	1482 (1479)	7.201	0.291

4.2 Data Preprocessing

A major step to successful small molecule modeling is to properly preprocess input data. All small molecules are available in Tripos Mol2, SDF (Structure-Data File) or SMILES (Simplified Molecular Input Line Entry Specification) format. Essentially these formats contain chemical structure information, and conversions can be done between different formats.

In order to consistently get well prepared structures for descriptor generations, LigPrep utility in Schrödinger suites (`https://www.schrodinger.com/`) are used to get optimized 3D structures, which are readily suitable for ESTD and auxiliary physical descriptors computations. Additionally, SMILES files are also essential for preparing auxiliary molecular descriptors by ChemoPy[136].

4.3 Molecular Descriptors

In this section, we will introduce molecular descriptors used in this work. These molecular descriptors can be divided into several categories: element specific topological descriptors (ESTDs), physical model based descriptors, and auxiliary molecular descriptors calculated by ChemoPy [136]. The motivation and details of our molecular descriptors construction process will also be presented.

4.3.1 Element specific topological descriptors

Element specific networks The key to accurate prediction is to engineer ESTDs from corresponding element specific networks (ESNs) based on which persistent homology is computed. It is therefore necessary to choose different element combinations in order to characterize the properties of a given molecule.

Filtration matrix The essential idea of element specific persistent homology (ESPH) and its modified version have been discussed in Section 2.3 and 2.4. Throughout this study, the definition of distance between atom i and atom j will be consistent with Eq. (2.10). The underlying reason has already been discussed in Section 2.4. Simply speaking, for small molecules, intramolecular effects such as van der Waals interaction play a much more significant role than covalent bonding effects. This is essentially different from large biomolecules such as proteins.

Topological dimension We also need to consider the dimensions of topological invariants. For large molecules such as proteins, it is important to compute the persistent homology of first three dimensions (Betti-0, Betti-1 and Betti-2). The underlying reason is that proteins generally consists of thousands of atoms, and Betti-1 and Betti-2 bars usually contain very rich geometric information such as internal loops and cavities. However, small molecules are geometrically simple and their barcodes of high dimensions are usually very sparse. Additionally, small molecules are chemically complex due to their involvement of many element types and oxidation states. As such, high dimensional barcodes of element specific networks carry little information. Therefore, we only consider Betti-0 bars for small molecule modeling.

4.3.1.1 ESNs for partition coefficient and aqueous solubility prediction

Inspired by classic atom-additive models for partition coefficient prediction, we utilize a total of 61 basic element types calculated by antechamber [137, 138] using general amber force field (GAFF)

[137]. Atoms of given atom type and their appropriate combinations are selected to construct Vietoris-Rips complex and ESTDs are subsequently calculated.

Table 4.6: ESNs for partition coefficient and aqueous solubility prediction

Network type	Element specific networks
Single-element	$\{a_i\}$, where $a_i \in \mathcal{A}$, $\mathcal{A} = \{\text{GAFF}_{61}\}$
Two-element	$\{b_i, b_j\}$, where $b_i, b_j \in \mathcal{B}$, $i, j \in \{1 \dots 3\}$, and $i < j$. Here $\mathcal{B}$={C, N, O}.

Remark For two-element type ESNs, we choose C, N, O for two considerations. The first reason is that rare elements are already included in single element group. The second reason is based on a statistical point of view. Specifically, we perform statistical analysis of the dataset and the occurrences of different element types are shown in Table 4.7.

Table 4.7: Statistics of element occurrences for partition coefficient training set

C	8198	S	1360
H	8172	F	672
O	6612	Br	308
N	6212	P	188
Cl	1361	I	115

4.3.1.2 ESNs for toxicity endpoint prediction

The ESTD construction for toxicity endpoint prediction is very similar. It may not make sense if we continue to use atom types generated by GAFF force field. Instead, we focus on intra-molecular interactions on a wider range of element types. It is reasonable to assume that rare elements (such as Br or I) are more capable of indicating a higher level of toxicity as compared to more frequently appeared elements (such as C or N). Thus these elements are considered when calculating ESTDs.

Different combinations of ESNs are tested, and we propose the following ESTDs for toxicity endpoint prediction, although our search may not be exhaustive. A list of ESNs used can be found in Table 4.8.

Table 4.8: ESNs for toxicity endpoint prediction

Network type	Element specific networks
Single-element	$\{a_i\}$, where $a_i \in \mathcal{A}$, $\mathcal{A}$={H, C, N, O}
Two-element	$\{b_i, c_j\}$, where $b_i \in \mathcal{B}$, $c_j \in C$, $i \in \{1 \ldots 3\}$, $j \in \{1 \ldots 9\}$, and $i < j$. Here $\mathcal{B}$={C, N, O} and C={C, N, O, F, P, S, Cl, Br, I}.

4.3.1.3 A general workflow for computing ESTDs from ESNs

A general workflow process for computing ESTDs from ESNs can be summarized as follows.

1. Given an ESN, 3D coordinates of atoms in the ESN are selected, and their Vietoris-Rips complexes are constructed. Note that distance defined in Eq. (2.10) is used for persistent homology barcodes generation in this study.

2. The maximum filtration size is set to a large number (12 Å for small molecules). After barcodes are obtained, barcodes are divided into several subintervals so that intra-molecular interactions of different strengths can be captured. For instance, ESTDs can be calculated based on the barcodes of the first 10 small subintervals $\text{Int}_i = [0.5\text{i}, 0.5(\text{i} + 1)]$, $i = 0, \ldots, 9$.

    - Within each Int_i, search Betti-0 bars whose birth time falls within this interval and Betti-0 bars that dies within Int_i, respectively, and denote these two sets of Betti-0 bars as $\text{S}_{\text{birth}_i}$ and $\text{S}_{\text{death}_i}$.
    - Count the number of Betti-0 bars within $\text{S}_{\text{birth}_i}$ and $\text{S}_{\text{death}_i}$, and these two counts yield 2 ESTDs for the interval Int_i.

3. In addition to interval-wise descriptors, we also consider global ESTDs for the entire barcodes. All Betti-0 bars' birth times and death times are collected and added into S_{birth} and S_{death}, respectively. The maximum, minimum, mean and sum of each set of values are then computed as ESTDs. This step gives 8 more ESTDs.

4.3.1.4 The essence of ESTDs

To summarize, we would like to emphasize the essential ideas of our choice of ESTDs. In Step 2 of the ESTD generation process in Section 4.3.1.3, we collect all birth and death time of Betti-0 bars in order to capture intra-molecular effects such as hydrogen bonding and van der Waals interactions. These intra-molecular interactions are captured by eliminating the topological connectivity of covalent bonds. For instance, the birth position can signal the formation of hydrogen bonding, and the death position represents the disappearance of such effects, which in turn reflects the strength of these effects. In step 3 of the above process, we consider all potential element-specific intra-molecular effects together and use statistics of these effects as global descriptors for a given molecule. This would help us to better characterize small molecules.

4.3.1.5 ESTDs for partition coefficient and aqueous solubility prediction

Notice that the first 10 subintervals with length 0.5 Å of Betti-0 barcodes are used for ESTD computation. Mathematically, these subintervals can be denoted as $\text{Int}_i = [0.5\text{i}, 0.5(\text{i} + 1)]$, $i = 0, \ldots, 9$. We further denote sets that contain all birth or death values of Betti-0 bars as S_{birth} and S_{death}, respectively. The word "Statistics" in Table 4.9 refer to maximum, minimum, mean and sum of values in S_{birth} and S_{death}. As Table 4.9 suggests, we have 1 ESTDs from each single-element

Table 4.9: ESTDs for partition coefficient and aqueous solubility prediction

Element specific networks	ESTDs
$\{a_i\}$, where $a_i \in \{\text{GAFF}_{61}\}$	Counts of Betti-0 bars for each of the 61 atom types
$\{b_i, b_j\}$, where $b_i, b_j \in \{\text{C, N, O}\}$, and $b_i \neq b_j$.	1. Counts of Betti-0 bars with birth or death values falling within each Int_i, $i = 0, \ldots, 9$.
	2. Statistics of birth or death values for all Betti-0 bars (consider all birth and death values, i.e., S_{birth} and S_{death})

ESN (count of Betti-0 bars). Meanwhile, 28 ESTDs are also generated for each two-element ESN. Specifically, we have 2 ESTDs (birth and death count) for each Int_i, and 4 global ESTDs for S_{birth} and S_{death}. Thus in total we have 145 (61 + 28*3) ESTDs for each molecule.

Remark To see how different bin sizes affect prediction accuracy, we also choose a slightly larger interval size to evaluate ESTDs' predictive power. Simply put, an interval size of 1 Å is used to construct ESTDs, which results in 5 subintervals. We also want to include the effects of Betti-1 bars. Such construction results in 121 ESTDs (61 + 2*2*5*3) for each molecule as we consider both birth and death values of barcodes with 3 ESNs, 5 subintervals and 2 different topological dimensions. It turns out this set of ESTDs perform very well for some specific aqueous solubility datasets (Small Delaney set and Huuskonen set).

4.3.1.6 ESTDs for toxicity endpoint prediction

ESTDs for toxicity endpoint prediction are constructed in a very similar way. The only difference is that for single-element ESN, we no longer consider GAFF atom types. Instead, we calculate the same set of ESTDs in the way discussed in section 4.3.1.5, which results in 28 ESTDs for each ESN. For two-element type, we also consider more combinations comparing to solubility prediction. The reason is that it is necessary to include more element types as they have yet been included in single-element. Since we have 25 ESNs in table 4.8, we have a total of 700 (25*28) ESTDs for each molecule.

4.3.2 Physical model based descriptors

In addition to ESTDs discussed above, we are also interested in constructing a set of microscopic features based on physical models to describe molecular toxicity. This set of features should be convenient to use in different machine learning approaches, including deep learning and non deep learning, and single-task and multi-task ones. To make our feature generation feasible and robust to all compounds, we consider three types of basic physical information, i.e., atomic charges computed from quantum mechanics or molecular force fields, atomic surface areas calculated for solvent excluded surface definition, and atomic electrostatic solvation free energies estimated from the Poisson model. Optimized 3D structure obtained from section 4.2 are used to compute atomic properties, which can be summarized in several steps:

1. **Charge** Optimized 3D structures are fed in antechamber [138], using parametrization: AM1-BCC charge, Amber mbondi2 radii and general Amber force field (GAFF) [137]. This step leads to pqr files with corresponding charge assignments.

2. **Surface** ESES online server at `http://weilab.math.msu.edu/ESES/` [139] is used to compute atomic surface area of each molecule, using pqr files from the previous step. This step also results in molecular solvent excluded surface information.

3. **Energy** MIBPB online server at `http://weilab.math.msu.edu/MIBPB/` [140] is used to calculate the atomic electrostatic solvation free energy of each molecule, using surface and pqr files from previous steps.

Specifically, physical descriptors come from Step 1, Step 2 and Step 3 above. To make our method scalable and applicable to all kinds of molecules, we manually construct element-specific molecular descriptors so that it does not depend on atomic positions or the number of atoms. The essential idea of such construction is to derive atomic properties of the each element type, which is very similar to the idea of ESPH.

We consider 10 different commonly occurring element types, i.e., H, C, N, O, F, P, S, Cl, Br, and I and three different types of descriptors – charge, surface area and electrostatic solvation free energies. Given an element type and a descriptor type, we compute the statistics of the quantities obtained from the aforementioned physical model calculation, i.e., summation, maximum, minimum, mean and variance, giving rise to 5 physical descriptors. To capture absolute strengths of each element descriptor, we further generate 5 more physical descriptors after taking absolute values of the same quantities. Consequently, we have a total of 10 physical descriptors for each given element type and descriptor type. Thus 300 (10 descriptor $\times$ 10 element types $\times$ 3 descriptor type) molecular descriptors can be generated at element type level.

Additionally when all atoms are included for computation, 10 more physical descriptors can be constructed in a similar way (5 statistical quantities of original values, and another 5 for absolute values) for each element descriptor type (charge, surface area and electrostatic solvation free

energies). This step yields another 30 molecular descriptors. As a result, we organize all of the above information into a 1D descriptor vector with 330 components, which can be directly fed into machine learning algorithms.

4.3.3 Auxiliary molecular descriptors for partition coefficient and aqueous solubility prediction

ChemoPy[136] is a popular software for computing 2D and 3D molecular descriptors. However, preliminary results have shown that 3D descriptors by ChemoPy do not perform well for our prediction tasks due to inaccurate generation of 3D structures. Thus in this study, we only incorporate 2D molecule ChemoPy descriptors on top of our self-designed molecular descriptors. ChemoPy descriptors can be categorized as following - 30 molecular constitutional descriptors, 35 topological descriptors, 44 molecular connectivity indices, 7 Kappa shape descriptors, 64 Burden descriptors, 245 E-state indices, 21 Basak information indices, 96 autocorrelation descriptors, 6 molecular property descriptors, 25 charge descriptors, and 60 MOE-type descriptors. A more detailed description of descriptor and ChemoPy software is available on line at `https://code.google.com/archive/p/pychem/downloads`. Also notice that ChemoPy software only requires 2D SMILES as input.

Specifically for partition coefficient and aqueous solubility prediction, we also combine these features with ESTDs to create ESTD$^+$ in order to improve the overall performance. For consistency reasons, only molecules whose descriptors can be calculated by both our ESTD software and ChemoPy software are used for training purpose. It is worth to mention that our ESTD approaches are applicable to all molecules whereas ChemoPy has difficulty in dealing with some molecules. Separate results and discussions for different sets of descriptors will also be conducted in later chapters.

CHAPTER 5

RESULTS

In this chapter, we will first introduce evaluation metrics used to evaluate model performances. Next, we will present our results along with those in literature to see if the proposed models can truly improve over others' models. Specifically, we would like to predict solvation free energy, partition coefficient, aqueous solubility, and various toxicity endpoints. Both results of different machine learning methods using the same set of molecular descriptors, and results of different molecular descriptors using the same machine learning algorithm will be provided, with emphasis on both machine learning algorithms and predictive power of molecular descriptors. Notice we may use different evaluation metrics/descriptors for different tasks.

Since solvation free energy is isolated from the other small molecular properties mentioned above, a separate discussion will be provided for solvation free energy. On the other hand, partition coefficient and aqueous solubility are closely related as they all measure how chemicals dissolve in a given solvent, while different toxicity endpoints may share similarities. Thus we will learn these properties jointly using multi-task learning framework, and the same descriptors shall be used for related tasks.

5.1 Evaluation criteria

5.1.1 Commonly used evaluation metrics

Commonly used evaluation metrics used in quantitative predictions are Pearson correlation coefficient (R), root mean squared error (RMSE), and mean unsigned error (MUE). Mathematically, they are defined as following:

$$R = \frac{\sum_{i=1}^{N}(X_i^{\text{Pred}} - \overline{X^{\text{Pred}}})(X_i^{\text{Expl}} - \overline{X_i^{\text{Expl}}})}{\sqrt{\sum_{i=1}^{N}(X_i^{\text{Pred}} - \overline{X^{\text{Pred}}})^2}\sqrt{\sum_{i=1}^{N}(X_i^{\text{Expl}} - \overline{X^{\text{Expl}}})^2}}, \tag{5.1}$$

$$\mathrm{R}MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i^{\mathrm{Pred}} - X_i^{\mathrm{Expl}}\right)^2} \tag{5.2}$$

and

$$\mathrm{M}UE = \frac{1}{N}\sum_{i=1}^{N}\left|X_i^{\mathrm{Pred}} - X_i^{\mathrm{Expl}}\right|, \tag{5.3}$$

where N is the total number of molecules in the test set, X_i^{Expl} and X_i^{Pred} stand for the experimental and predicted value for the ith molecule, respectively, $\overline{X^{\mathrm{Pred}}}$ and $\overline{X^{\mathrm{Expl}}}$ is the average of predicted and experimental value for the entire test set, respectively.

5.1.2 Additional evaluation metrics for partition coefficient and aqueous solubility prediction

In Tetko's review paper [6] for partition coefficient prediction on Star and Non-star set, an additional metric based on the difference between experimental and predicted $\log P$ ($\Delta \log P$) was proposed. Specifically, the percentage within various error ranges was considered.:

- If $|\Delta \log P| < 0.5$, prediction is considered to be "acceptable";
- If $0.5 \leq |\Delta \log P| < 1.0$, prediction is considered to be "disputable";
- If $|\Delta \log P| \geq 1.0$, prediction is considered to be "unacceptable".

This metric is exclusively used for Star and Non-star set proposed by Tetko [6].

5.1.3 Additional evaluation metrics for toxicity endpoint prediction

In T.E.S.T. software, developers referred to Golbraikh *e*t al. [141]'s protocol to determine if a QSAR model has a predictive power.

$$q^2 > 0.5, \tag{5.4}$$

$$R^2 > 0.6, \tag{5.5}$$

$$\frac{R^2 - R_0^2}{R^2} < 0.1 \tag{5.6}$$

$$0.85 \leq k \leq 1.15 \tag{5.7}$$

where q^2 is the squared leave one out correlation coefficient for the training set, R^2 is the squared Pearson correlation coefficient between the experimental and predicted toxicities for the test set, R_0^2 is the squared correlation coefficient between the experimental and predicted toxicities for the test set with the y-intercept being set to zero so that the regression is given by $Y = kX$. All these metrics will be used to compare performances of toxicity endpoint prediction models.

5.2 Evaluation results

5.2.1 Solvation free energy prediction

5.2.1.1 Microscopic feature parametrization

In earlier hybrid physical and knowledge (HPK) model [4], three types of atomic radii (Amber 6, Amber bondi, and Amber mbondi2 [142]) and three types of charge assignments (OpenEye-AM1-BCC v1 parameters [143], Gasteiger [144], and Mulliken [142]) were used to test the sensitivity and accuracy of models with respect to different parameterizations. For such reason, we continue to use these 9 different parameterizations in order to conveniently compare current models with previous ones.

5.2.1.2 Polar and non-polar descriptors for solvation free energy

Implicit solvent models divide solvation free energies into polar and nonpolar additive contributions, whereas polar and nonpolar interactions are inseparable and non additive. Thus in order to explore how important polar and nonpolar descriptors contribute to solvation process, we use two different set of descriptors to predict small molecule solvation energy.

Descriptor set 1: Polar descriptors with high correlations to solvation free energy. The list of descriptors used can be found in Appendix.

Descriptor set 2: Non-polar atomic surface area descriptors (part of physical descriptors discussed in previous chapter), in addition to polar descriptors with high correlations to solvation free energy in descriptor set 1.

5.2.1.3 Leave-one-out result

Fig. 5.1 contains leave-one-out results using different descriptor sets and HPK models for the 668 set, with different parameterizations. It is evident that the present model outperforms our earlier HPK model with most parameterizations.

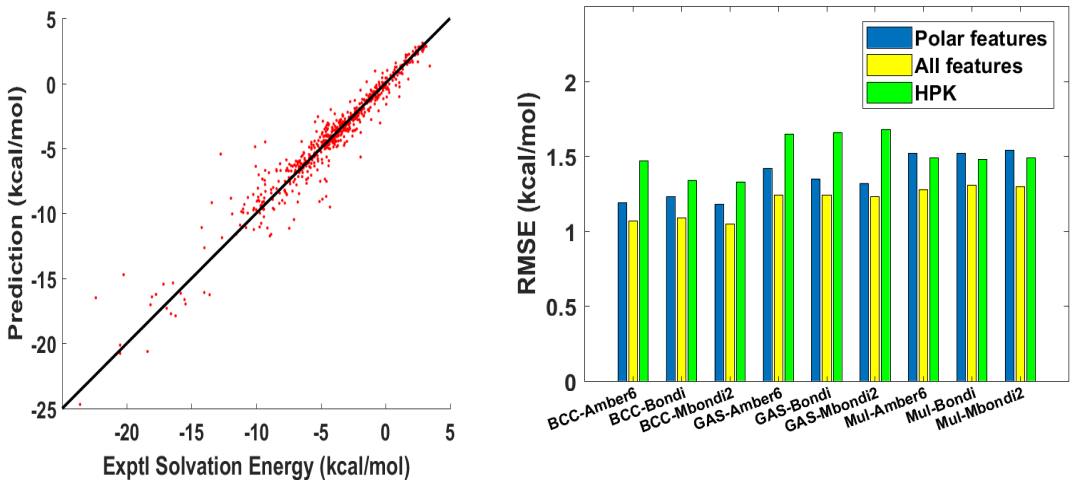


Figure 5.1: Illustration of leave-one-out predictions for the whole set of 668 molecules. Left chart: Correlation between experimental solvation free energies and predictions obtained by BCC charges and Amber MBondi2 using all polar-nonpolar features. Right chart: Comparison of prediction RMSEs obtained by models with polar features and all features against HPK models. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively.

Moreover, leave-one-out method is also used to determine the number of nearest neighbors that should be used to predict a given molecule. For each parameterization, we perform leave-one-out prediction with 1 to 10 nearest neighbors. The best results can be found in Table 5.1. Although complete numerical results are not provided, it is worth to mention that further increasing the number of descriptors does not necessarily increase prediction accuracy. For consideration of consistency, we select 10 nearest neighbors for all following tests.

Table 5.1: The RMSE and ME of the leave-one-out test in the solvation free energy prediction of 668 molecules with descriptor set 1 (the first position), descriptor set 2 (the second position) and HPK model (the last position) [4]. All errors are in unit kcal/mol.

Radius	Charge	BCC	Mulliken	Gasteiger
Amber 6	RMSE	1.19, 1.07,1.47	1.52, 1.28, 1.49	1.42, 1.24, 1.65
	ME	-0.01, 0.01, -0.13	-0.01, 0.00, -0.20	-0.03, -0.01, -0.19
Amber Bondi	RMSE	1.23, 1.09, 1.34	1.52, 1.31, 1.48	1.35, 1.24, 1.66
	ME	-0.02, 0.01, -0.14	-0.01, 0.01, -0.21	-0.02, 0.00, -0.13
Amber MBondi2	RMSE	1.18, 1.05, 1.33	1.54, 1.30, 1.49	1.33, 1.23, 1.68
	ME	0.00, 0.03, -0.14	0.0, 0.01, -0.22	0.00, 0.00, -0.22

5.2.1.4 Blind prediction of SAMPLx challenge molecules

SAMPL0 test First, let us consider the solvation free energy prediction for the SAMPL0 test set, which contains a total of 17 molecules. All structures of this test set are relatively simple. However, the molecule species of this set is quite diverse. Many researchers have reported their solvation free energy predictions for this challenge set [145, 146]. Prior to our work, the optimal prediction for this test set has an RMSE of 1.34 kcal/mol for the whole set [146]. Figure 5.2 depicts the present results for a total of 9 charge and radius combinations. When the BCC charge is used, the RMSEs of our predictions with three radius parametrizations are all smaller than 0.75 kcal/mol. Our optimal prediction has an RMSE of 0.61 kcal/mol, obtained from Amber Bondi radius parametrization in conjugation with the BCC charge assignment with polar features only. When all polar and nonpolar features are used, the results become slightly worse whereas performances over all parametrizations turn out to be more stable especially when the Mulliken charge assignment is used.

SAMPL1 test Having demonstrated the superiority of the proposed model for the prediction of the SAMPL0 challenge set, we further consider the SAMPL1 test set, which is generally believed to be the most difficult one, due to the following two reasons. First, the molecular structures of this test set are extremely complex compared to other molecules with known experimental solvation free energies. Second, the uncertainty of SAMPL1 experimental data is very large. For some molecules the uncertainty is as large as 2.0 kcal/mol [147, 14]. Nevertheless, it is extremely desirable to develop an accurate modeling paradigm for this test set because most molecules in this test set are

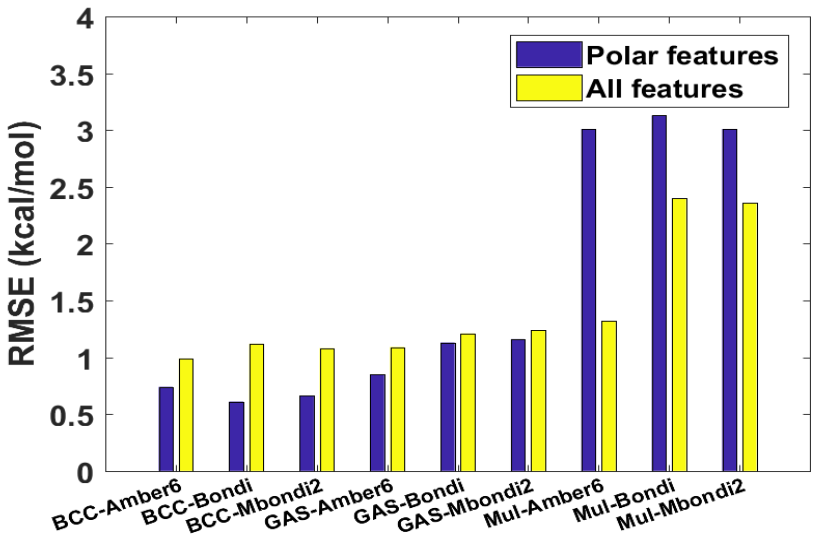


Figure 5.2: Illustration of prediction RMSEs obtained with different molecular parametrizations by the model for SAMPL0 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively.

druggable. The best prediction for the whole set has an RMSE of 2.45 kcal/mol [146]. On a subset of the SAMPL1 test set that contains only 56 molecules, the best performance was shown to give an RMSE of 2.4 kcal/mol [14]. Figure 5.3 illustrates the results of the approach for the whole SAMPL1 test set. It is obvious to see that the model is much more accurate. The optimal prediction with only polar descriptors has an RMSE as small as 2.07 kcal/mol, and adding nonpolar descriptors further improves the RMSE to 1.86 kcal/mol, which is the best to our best knowledge. Additionally, the present model is very robust with respect to the change in force fields. The maximum and minimum prediction RMSEs over 9 sets of parametrizations and 2 feature combinations are 1.86 and 2.82 kcal/mol, respectively. The difference between the maximum and minimum is 0.96 kcal/mol, which is much smaller than experimental uncertainty of 2 kcal/mol for this set [147, 14].

SAMPL2 test Another difficult test set is SAMPL2, which contains a total of 30 molecules [148]. The experimental uncertainty on these molecules is much less than that of the SAMPL1 test set. Nevertheless, accurate solvation prediction for this set is rare. Using all-atom molecular dynamics simulations and multiple starting conformations for prediction, Klimovich and Mobley reported

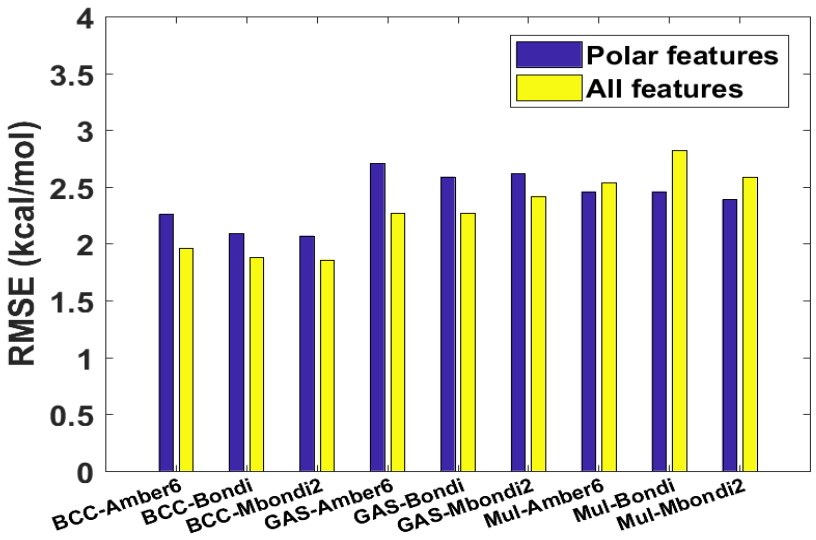


Figure 5.3: Illustration of prediction RMSEs obtained with different molecular parametrizations by the proposed model for SAMPL1 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively.

an RMSE of 2.82 kcal/mol over the whole set and 1.86 kcal/mol over all the molecules except several hydroxyl-rich compounds [148]. Some of the best reported predictions have an RMSE of 1.59 kcal/mol [146]. In our previous test, the molecule containing an I atom (5-iodouracil) is excluded in all calculations due to the lack of appropriate charge force field. In this work, we also ignore this molecule for the same reason. The HPK model gives an optimal prediction with RMSE 1.96 kcal/mol. However, the RMSEs of the prediction vary over a large range, from 1.96 to 4.86 kcal/mol, when different charge and radius force fields are applied. A bar graph of the RMSEs of the predictions is given in Fig. 5.4. Parametrizations based on AM1-BCC charge yield the best results among polar features and adding nonpolar features offers a substantial improvement over polar features, as the first three yellow bars are lower than the first three blue bars. The optimal RMSE for SAMPL2 molecules is 1.64 kcal/mol with AM1-BCC charge and AMBER6 radius, when all features are used to train the models. It is also worthy to note that the variation of RMSEs under different parametrizations is 1.42 kcal/mol (1.64 to 3.06 kcal/mol), which indicates the robustness of the present models compared to HPK models.

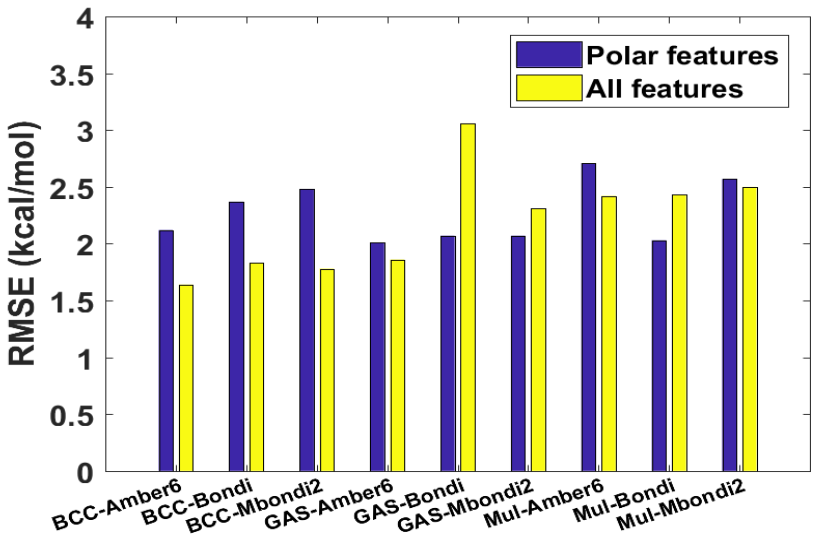


Figure 5.4: Illustration of prediction RMSEs obtained with different molecular parametrizations by the proposed model for SAMPL2 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively.

SAMPL3 test The SAMPL3 test set, which contains 36 molecules, is relatively easy for prediction. The structures of SAMPL3 molecules are relatively simple, and most molecules in this set are chlorinated hydrocarbon molecules [149]. The best prediction in the literature offers an RMSE of 1.29 kcal/mol [146]. Figure 5.5 depicts the RMSEs of the predictions by only polar features and all features. Although the optimal result (RMSe of 0.86 kcal/mol) is generated by polar features with Gasteiger charges, all features combination turns out to be more stable over all parametrizations as Figure 5.5 clearly shows. More specifically, the RMSEs using polar features span over a small range of 0.48 kcal/mol (i.e., from 0.86 to 1.34 kcal/mol) across all 9 different parametrizations, while all features yield a variation of 0.24 kcal/mol. This further verifies the robustness of the current solvation model.

SAMPL4 test Finally, we consider the SAMPL4 test set, which is a very popular one. Many explicit, implicit, integral equation, and hybrid QM/MM approaches [16] have been applied to this set [150]. As shown in Fig. 5.6, the overall performance enhances when all features are used as the blue bars are consistently higher than yellow bars, which indicates the predictive power of nonpolar

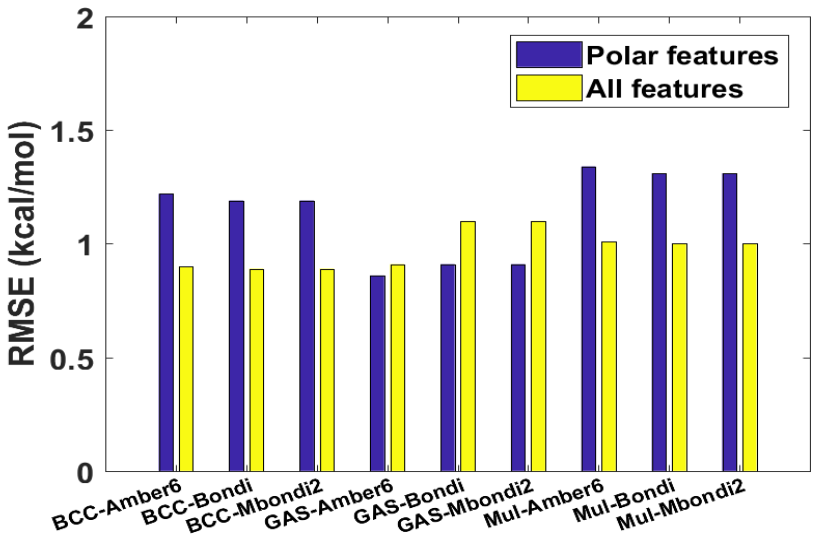


Figure 5.5: Illustration of prediction RMSEs obtained with different molecular parametrizations by the proposed model for SAMPL3 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively.

features. Our proposed model gives an optimal RMSE of 1.14 kcal/mol. It is also easy to see that our current approach is quite robust across different force field and charge parametrizations.

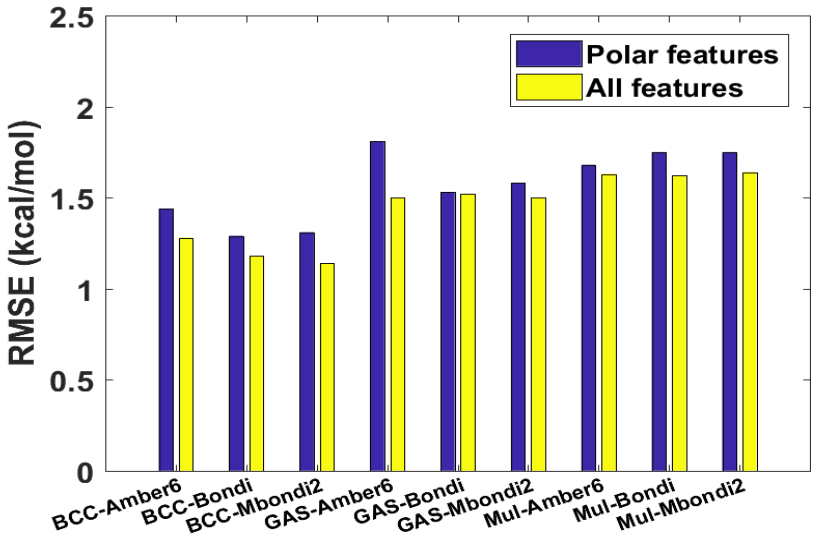


Figure 5.6: Illustration of prediction RMSEs obtained with different molecular parametrizations by the proposed model for SAMPL4 test set. In the plot, GAS and MUL are abbreviations for Gasteiger and Mulliken charges, respectively.

5.2.1.5 Wang's [1] dataset

Finally, we validate our models on the external datasets that Ref [1] used to develop weighted solvent accessible surface area. It is interesting to compare our results with theirs since compounds used for training and testing are essentially independent, which challenges the predictive power of solvation models.

In Model III, they achieved unsigned average errors of 0.50 and 0.66 kcal/mol for the training set and the test set, respectively [1], and an unsigned average error of 0.538 kcal/mol for the entire set. Our models were also trained with 10 nearest neighbors and all polar/nonpolar descriptors, and the average of 50 independent runs yields unsigned average errors of 0.00 and 0.57 kcal/mol, respectively for the training set and the test set, and an unsigned average error of 0.441 kcal/mol for the entire set. In fact, we used a slightly smaller training set of 289 molecules because 4 molecules in the training set have ambiguous chemical names in the PubChem database, while all molecules in the test set are included in our prediction.

Since there are 11 duplicates in the training set and the test set of Ref. [1] (see Table 4.2), we also train models with these 11 duplicated molecules excluded from the training set. The updated model acheives an RMSE of 0.61 kcal/mol for the test set, which is still smaller than that reported in Ref. [1] (i.e., 0.66 kcal/mol).

Therefore, we conclude that our models have a competitive edge over the classic solvation model based on weighted SASA.

5.2.2 Partition coefficient and aqueous solubility prediction

In this section, we present the results of the proposed ESPH methods in conjugation with gradient boosting decision tree and multi-task deep neural networks for variety of data sets, including partition coefficient and solubility test sets. Otherwise stated, different tasks are trained together in the same network. Besides, we would like to introduce some notations for easier reference. ESTD-1 contains 61 Betti-0 bar-based ESTDs, ESTD-2 contains all ESTDs listed in Table 4.9 (a total of 145 ESTDs), and ESTD-3 corresponds to ESTDs discussed in Remark of Section 4.3.1.5.

When molecular descriptors calculated by ChemoPy software [136] are combined with ESTD-1, ESTD-2 and ESTD-3 respectively, two more descriptor sets are created and they are denoted as ESTD$^+$-1, ESTD$^+$-2 and ESTD$^+$-3, respectively.

5.2.2.1 $\log P$ training set cross-validation

In order to have an idea of how our topological representation would work for partition coefficient, a 10-fold cross-validation is performed using baseline method GBDT. Note that 50 runs were done to achieve the final results as randomness is involved and the results are summarized in Table 5.2. Notice that ChemoPy descriptors were not used here. It can be seen that our descriptors

Table 5.2: Results of 10-fold cross validation on the partition coefficient training set, $N = 8199$.

Method	R^2	RMSE	MUE
GBDT-ESTD-2	**0.923**	**0.45**	**0.32**
GBDT-ESTD-1	0.912	0.48	0.35
XLOGP3-AA [5]	0.904	0.50	0.39

perform better than XLOGP3 software [5] given the same training data, and thus demonstrates great predictive power. In addition, we also provide the fluctuation of 10-fold cross-validation of Table 5.3, even though such statistics is not available for XLOGP3 software. It is clear that our ESTDs give quite consistent predictions and their performances are independent of random fold generations since RMSDs of R^2, RMSE and MUE do not essentially fluctuate. Thus it would be very interesting to see the performances of our MT-DNN compared to XLOGP3 and GBDT.

Table 5.3: Performances and fluctuations of fifty 10-fold cross-validation test runs.

Method	Mean R^2 (RMSD)	Mean RMSE (RMSD)	Mean MUE (RMSD)
GBDT-ESTD-2	0.923 (0.001)	0.45 (0.003)	0.32 (0.002)
GBDT-ESTD-1	0.912 (0.001)	0.48 (0.002)	0.35 (0.001)

5.2.2.2 FDA set

The first test set that we would like to apply our model to is the FDA test set. A molecule that contains Hg was dropped due to the difficulty of computation. A major challenge of this set is that its structures are more complex than that of the training set, and the partition coefficient range spans over nearly 12 units. A series of prediction methods [5], including our multi-task neural networks, are applied to this set and their results are summarized in Table 5.4 for a comparison with ours.

Table 5.4: Results of different logP prediction methods on 406 FDA-approved drugs [5], ranked by R^2. Two molecules were dropped for our model evaluation due to feature generation failure of ChemoPy

Method	R^2	RMSE	MUE
GBDT-ESTD+-2-AD	**0.935**	**0.51**	0.24
GBDT-ESTD+-1-AD	0.932	0.52	0.23
MT-ESTD$^+$-1-AD	0.930	0.53	**0.22**
MT-ESTD-1-AD	0.929	0.54	0.26
MT-ESTD$^+$-2-AD	0.928	0.53	0.27
MT-ESTD-1	0.920	0.57	0.28
MT-ESTD-2-AD	0.912	0.59	0.37
GBDT-ESTD$^+$-1	0.910	0.60	0.28
GBDT-ESTD$^+$-2	0.910	0.60	0.30
MT-ESTD$^+$-1	0.909	0.60	0.27
MT-ESTD$^+$-2	0.909	0.60	0.34
ALOGPS	0.908	0.60	0.42
GBDT-ESTD$^+$-1	0.900	0.63	0.39
GBDT-ESTD-1	0.893	0.66	0.41
MT-ESTD-2	0.891	0.66	0.44
GBDT-ESTD$^+$-2	0.883	0.68	0.49
XLOGP3	0.872	0.72	0.51
GBDT-ESTD-2	0.848	0.78	0.57
XLOGP3-AA	0.847	0.80	0.57
CLOGP	0.838	0.88	0.51
TOPKAT	0.815	0.88	0.56
ALOGP98	0.80	0.90	0.64
KowWIN	0.771	1.10	0.63
HINT	0.491	1.93	1.30

As we can see from Table 5.4, our multi-task model gives the best prediction in terms of R^2, RMSE, and MUE. Specifically, the small MUE of our model indicates that our predictions are less biased than other methods tested, except for some outliers. Also, note that the training set is

completely independent of the test set which shows the applicability of our multi-task architecture. We also build models with the same architecture when additional molecules gathered from NIH-database are included as there is no guarantee that the training set of ALOGPS is completely independent of the test set. It turns out that the accuracy can be greatly improved. For instance, the performance of ESTD$^+$-1 can be improved by more than 10% in terms of RMSE (0.60 log units to 0.53 log units). It demonstrates the potential of our MT-DNN architecture when more data become available and it will be more carefully discussed in later section.

5.2.2.3 Star set and non-star set

Star set and Non-star set were proposed by Tetko [6] as two benchmark sets for evaluating partition coefficient models. Over 20 different models were tested on these two sets. It should be emphasized that for these sets, different models are trained on different training sets and their overlap with the test sets is unknown. Thus it makes more sense to merge our 8199 training set with additional molecules in NIH database and see how additional training data can benefit the overall performances. Results of different models on these two sets can be found in Table 5.5. Notice that models trained with additional data from NIH database are labeled with (-AD).

For star set, we achieve RMSE of 0.49 log units with other popular commercial software packages such as ACD/logP and CLOGP, in addition to a high acceptable prediction percentage (77%, rank 2). For non-star set, most methods do not give accurate predictions as the structures in this set are relatively new and complex. Our 51% acceptable rate ranks number 2 among all predictors, though RMSE is relatively high due to a few large outliers. The results are satisfactory, especially when considering commercial software packages generally use a much larger training set than that ours. In general, when there exist more overlapped molecules in the training set, the test results will be significantly improved. Thus as a baseline comparison, it would be more meaningful if we compare our results with XLogP3 software. As Table 5.5 indicates, our MT-ESTD$^+$ models achieve a substantial improvement over XLOP3 for star set, while XLogP3 achieves a lower RMSE for Non-star set. It may be due to XLogP3's corrections terms with relatively new structures. The

performances of our MT-ESTD$^+$ models suggest that our models are able to predict log P accurately. We also would like to know if the predictive power can potentially be further improved once more molecules are incorporated into the training set. Thus we extend our original 8199 training set by adding molecules in both Star set and Non-star set. As shown in Table 5.5, descriptor sets labeled with (-AD) (with additional data from NIH database) generally offer better performances on both Star set and Non-star set.

Table 5.5: Benchmark test results [6] on both star and non-star set.

	Star Set (N = 223)				Non-star Set (N = 43)			
		% of Molecules Within Error Range				% of Molecules Within Error Range		
Method	RMSE	< 0.5	<1	> 1	RMSE	< 0.5	<1	> 1
AB/LogP	0.41	84	12	4	1.00	42	23	35
S+logP	0.45	76	22	3	0.87	40	35	26
MT-ESTD$^+$-1-AD	0.49	77	16	7	0.98	49	19	33
MT-ESTD$^+$-2	0.49	74	21	5	0.97	49	23	28
MT-ESTD$^+$-2-AD	0.50	76	17	7	0.94	51	19	30
ACD/logP	0.50	75	17	7	1.00	44	32	23
GBDT-ESTD$^+$-1-AD	0.51	76	17	6	1.03	44	30	25
GBDT-ESTD$^+$-2-AD	0.51	75	17	7	1.04	41	30	27
CLOGP	0.52	74	20	6	0.91	47	28	26
VLOGP OPS	0.52	64	21	7	1.07	33	28	26
ALOGPS	0.53	71	23	6	0.82	42	30	28
MT-ESTD$^+$-1	0.53	75	17	8	0.97	47	28	26
MT-ESTD-1-AD	0.53	73	18	9	1.00	37	30	33
MT-ESTD-2-AD	0.53	71	19	9	1.01	47	19	35
MT-ESTD-1	0.55	72	18	10	1.01	33	28	40
MT-ESTD-2	0.56	66	23	11	1.06	35	33	33
MiLogP	0.57	69	22	9	0.86	49	30	21
GBDT-ESTD$^+$-2	0.58	75	16	8	1.06	44	25	30
GBDT-ESTD$^+$-1	0.60	74	15	9	1.02	46	23	30

Table 5.5 (cont'd)

XLOGP3	0.62	60	30	10	0.89	47	23	30
KowWIN	0.64	68	21	11	1.05	40	30	30
GBDT-ESTD-2-AD	0.65	62	26	11	1.15	46	16	37
CSLogP	0.65	66	22	12	0.93	58	19	23
GBDT-ESTD-1-AD	0.68	71	16	12	1.16	41	11	46
ALOGP	0.69	60	25	16	0.92	28	40	33
MolLogP	0.69	61	25	14	0.93	40	25	26
ALOGP98	0.70	61	26	13	1.00	30	37	33
GBDT-ESTD-1	0.71	63	22	13	1.07	34	20	44
OsirisP	0.71	59	26	16	0.94	42	26	33
VLOGP	0.72	65	22	14	1.13	40	28	33
GBDT-ESTD-2	0.73	52	30	17	1.23	44	16	39
TLOGP	0.74	67	16	13	1.12	30	37	30
ABSOLV	0.75	53	30	17	1.02	49	28	23
QikProp	0.77	53	30	17	1.24	40	26	35
QuantlogP	0.80	47	30	22	1.17	35	26	40
SLIPPER-2002	0.80	62	22	15	1.16	35	23	42
COSMOFrag	0.84	48	26	19	1.23	26	40	23
XLOGP2	0.87	57	22	20	1.16	35	23	42
QLOGP	0.96	48	26	25	1.42	21	26	53
VEGA	1.04	47	27	26	1.24	28	30	42
CLIP	1.05	41	25	30	1.54	33	9	49
LSER	1.07	44	26	30	1.26	35	16	49
MLOGP(Sim+)	1.26	38	30	33	1.56	26	28	47
NC+NHET	1.35	29	26	45	1.71	19	16	65
SPARC	1.36	45	22	32	1.70	28	21	49
HINTLOGP	1.80	34	22	44	2.72	30	5	65

Remark for aqueous solubility prediction To evaluate the performances of our models for aqueous solubility prediction, several datasets are used, derived from Wang *e*t al. [2] and Hou *e*t al.

[3]. For leave-one-out validation, only the baseline method is used. For 10-fold cross-validation, the 9 remaining folds are trained together with the partition coefficient training set when evaluating the remaining fold with MT-DNN architecture.

5.2.2.4 Wang's 1708 set in ref. [2]

For this dataset, both leave-one-out and 10-fold cross-validation are carried out in order to evaluate the performance of our models.

Leave-one-out As MT-DNN requires a lot of computational resources, only baseline method GBDT is used for leave-one-out prediction. We use 4000 trees and 0.10 learning rate as training parameters to develop models and following results in Table 5.6 are achieved.

Table 5.6: Leave-one-out test on the 1708 solubility data set.

Method	R^2	RMSE	MUE
GBDT-ESTD$^+$-1-AD	**0.931**	**0.543**	**0.389**
GBDT-ESTD$^+$-2-AD	0.929	0.551	0.389
GBDT-ESTD$^+$-2	0.910	0.621	0.457
ASMS-LOGP[2]	0.897	0.664	0.505
GBDT-ESTD$^+$-1	0.893	0.683	0.494
ASMS[2]	0.884	0.707	0.547

10-fold cross-validation As MT-DNN and baseline method GBDT involve randomness, we run MT-DNN and GBDT 50 times and report mean performances for all metrics. The results are summarized in Table 5.7. It is observed that our models yield more accurate and robust predictions than ASMS and ASMS-LOGP models do, improving the R^2 from 0.884 to 0.925. Additionally, we also notice that there generally exists an improvement of MT-ESTD models over GBDT models, though, not as significant as what we see in the previous partition coefficient prediction.

Table 5.7: 10-fold cross-validation on the 1708 solubility data set.

Method	Mean R^2 (RMSD)	Mean RMSE (RMSD)	Mean MUE (RMSD)
MT-ESTD$^+$-1	**0.925** (0.001)	**0.568** (0.005)	**0.393** (0.003)
MT-ESTD$^+$-2	0.924 (0.003)	0.571 (0.010)	0.395 (0.004)
GBDT-ESTD$^+$-1	0.924 (0.002)	0.572 (0.006)	0.408 (0.005)
GBDT-ESTD$^+$-2	0.923 (0.002)	0.571 (0.006)	0.408 (0.005)
MT-ESTD-1	0.908 (0.002)	0.630 (0.005)	0.466 (0.003)
GBDT-ESTD-2	0.904 (0.002)	0.642 (0.008)	0.469 (0.005)
MT-ESTD-2	0.902 (0.002)	0.649 (0.007)	0.466 (0.005)
GBDT-ESTD-1	0.889 (0.003)	0.697 (0.009)	0.502 (0.005)
ASMS[2]	0.884 (0.021)	0.699 (0.054)	0.527 (0.034)
ASMS-LOGP [2]	0.869 (0.022)	0.742 (0.053)	0.570 (0.034)

5.2.2.5 Dataset in ref. [3]

We test our models on dataset proposed by Hou *e*t al. [3], where training and test sets were predefined to cover a variety of molecules. Klopman's test set contains 21 commonly used compounds of pharmaceutical and environmental interest [151] and is to be trained on the original 1290 molecules. Zhu's test set contains 120 molecules that were used to develop Klopman and Zhu's group contribution model [152]. As Hou *e*t al. [3] suggested, we remove 83 molecules that overlap with Zhu's test set from the training set to make predictions independent and unbiased. This reduces the size of the training set for Zhu's test set to 1207.

Klopman's test set Table 5.8 shows the performances of different models on Klopman's test set. Our MT-ESTD models perform similarly to Drug-LOGS method while achieving improvement over Klopman and Zhu's MLR method [152] with ESTDs. It is also evident that the MT-DNN method has an edge over GBDT method, which is consistent with our previous experiments.

Zhu's test set The results of Zhu's test set are summarized in Table 5.9. For this dataset, our MT-ESTD models give satisfactory results with a high Pearson correlation over 0.97 across all ESTD combinations. Such results indicate that our methods are applicable to a wide variety of molecules. Again, the MT-DNN method outperforms the GBDT method.

Table 5.8: Results of Klopman's test set [3], where MUE was not reported.

Method	R	RMSE
MT-ESTD$^+$-1	**0.94**	0.69
Drug-LOGS[3]	**0.94**	**0.64**
GBDT$^+$-2	0.94	0.71
MT-ESTD$^+$-2	0.93	0.75
GBDT$^+$-1	0.93	0.76
MT-ESTD-2	0.92	0.79
Klopman MLR [152]	0.92	0.86
GBDT-2	0.92	0.85
MT-ESTD-1	0.91	0.82
GBDT-1	0.84	1.07

Table 5.9: Results of Zhu's test set.

Method	R	RMSE	MUE
MT-ESTD$^+$-1	**0.97**	**0.65**	**0.47**
MT-ESTD$^+$-2	0.97	0.67	0.48
MT-ESTD-1	0.97	0.70	0.50
MT-ESTD-2	0.97	0.71	0.53
GBDT$^+$-2	0.97	0.73	0.50
GBDT$^+$-1	0.96	0.76	0.52
Drug-LOGS [3]	0.96	0.79	0.57
GBDT-2	0.96	0.79	0.60
GBDT-1	0.96	0.82	0.58
Group contribution [152]	0.96	0.84	0.70

Small Delaney set The small Delaney set has been extensively tested using different approaches, such as ESOL [129] and GSE [153]. Table 5.10 shows the 10-fold cross-validation results of the MT-DNN-ESTD$^+$ model and other methods. Overall, MT-DNN-ChemoPy and MT-DNN-ESTD$^+$ give very similar results in terms of the R^2, RMSE and MUE, which essentially proves the predictive power of our MTL framework. In addition, it is encouraging to notice that the MT-DNN-ESTD model which uses purely ESTDs as input slightly improves 2D kernel model [154]. As a comparison, our baseline method (random forest) underperforms. Again, we notice a substantial accuracy improvement of MT-DNN architecture over RF by direct comparison between their results. It indicates that our MT-DNN models benefit from MTL and there potentially exists an underlying feature representation for partition coefficient and aqueous solubility.

Table 5.10: 10-fold cross-validation results on the small Delaney set.

Methods	R^2	RMSE	MUE
MT-ESTD$^+$-3	**0.93**	**0.54**	**0.37**
MT-ESTD-3	0.92	0.61	0.43
RF-ESTD$^+$-3	0.91	0.63	0.45
RF-ESTD-3	0.88	0.71	0.52
GSE[153]	-	-	0.47
2D kernel[154]	0.91	0.61	0.44

Huuskonen set The Huuskonen set is also a popular solubility set. Similary to the previous small Delaney set, a direct 10-fold cross-validation yileds the results as listed in Table 5.11. Again the results of MT-DNN turn out to be the best in terms of all metrics. When only 121 ESTDs are used, the results of MT-DNN become slightly worse but still perform better than the RBF kernel approach and random forest models.

Table 5.11: 10-fold cross-validation results on the Huuskonen set.

Methods	R^2	RMSE	MUE
MT-ESTD$^+$-3	**0.93**	**0.55**	**0.39**
MT-ESTD-3	0.91	0.60	0.43
RF-ESTD$^+$-3	0.91	0.61	0.45
RF-ESTD-3	0.89	0.69	0.51
RBF[155]	0.90	-	-

Remark We find that ESTD-1 and ESTD-2 do not perform as well as ESTD-3 for Delaney and Huuskonen set, thus their results are skipped. For similar reasons, ESTD-3's results are not included for the other sets.

5.2.3 Toxicity endpoint prediction

5.2.3.1 Fathead minnow LC_{50} test set

The fathead minnow LC_{50} set was randomly divided into a training set (80% of the entire set) and a test set (20% of the entire set) [133], based on which a variety of TEST models were built. Table 5.12 shows the performances of five T.E.S.T. models, the TEST consensus obtained by the

average of all independent TEST predictions, four proposed methods and two consensus results obtained from averaging over present RF, GBDT, ST-DNN and MT-DNN results. TEST consensus gives the best prediction [133] among TEST results, reporting a correlation coefficient of 0.728 and RMSE of 0.768 log(mol/L). As Table 5.12 indicates, our MT-DNN model outperforms TEST consensus both in terms of R^2 and RMSE with only ESTDs as input. When physical descriptors are independently used or combined with ESTDs, the prediction accuracy can be further improved to a higher level, with R^2 of 0.771 and RMSE of 0.705 log(mol/L). The best result is generated by consensus method using all descriptors, with R^2 of 0.789 and RMSE of 0.677 log(mol/L).

Table 5.12: Comparison of prediction results for the fathead minnow LC_{50} test set.

Method	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical [133]	0.710	0.075	0.966	0.801	0.574	0.951
Single Model [133]	0.704	0.134	0.960	0.803	0.605	0.945
FDA [133]	0.626	0.113	0.985	0.915	0.656	0.945
Group contribution [133]	0.686	0.123	0.949	0.810	0.578	0.872
Nearest neighbor [133]	0.667	0.080	1.001	0.876	0.649	0.939
TEST consensus [133]	0.728	0.121	0.969	0.768	0.545	0.951
Results with ESTDs						
RF	0.661	0.364	0.946	0.858	0.638	1.000
GBDT	0.672	0.103	0.958	0.857	0.612	1.000
ST-DNN	0.675	0.031	0.995	0.862	0.601	1.000
MT-DNN	0.738	0.012	1.015	0.763	0.514	1.000
Consensus	0.740	0.087	0.956	0.755	0.518	1.000
Results with only auxiliary molecular descriptors						
RF	0.744	0.467	0.947	0.784	0.560	1.000
GBDT	0.750	0.148	0.962	0.736	0.511	1.000
ST-DNN	0.598	0.044	0.982	0.959	0.648	1.000
MT-DNN	0.771	0.003	1.010	0.705	0.472	1.000
Consensus	0.787	0.105	0.963	0.679	0.464	1.000
Results with all descriptors						
RF	0.727	0.322	0.948	0.782	0.564	1.000
GBDT	0.761	0.102	0.959	0.719	0.496	1.000
ST-DNN	0.692	0.010	0.997	0.822	0.568	1.000
MT-DNN	0.769	0.009	1.014	0.716	0.466	1.000
Consensus	**0.789**	0.076	0.959	**0.677**	**0.446**	1.000

5.2.3.2 Daphnia magna LC_{50} test set

The Daphinia Magna LC_{50} set is the smallest in terms of set size, with 283 training molecules and 70 test molecules, respectively. However, it brings difficulties to building robust QSAR models given the relatively large number of descriptors. Indeed, five independent models in TEST software give significantly different predictions, as indicated by RMSEs shown in Table 5.13 ranging from 0.810 to 1.190 log units. Though the RMSE of Group contribution is the smallest, its coverage is only 0.657 % which largely restricts this method's applicability. Additionally, its R^2 value is inconsistent with its RMSE and MAE. Since Ref. [133] states that "The consensus method achieved the best results in terms of both prediction accuracy and coverage", these usually low RMSE and MAE values might be typos.

We also notice that our non-multitask models that contain ESTDs result in very large deviation from experimental values. Indeed, overfitting issue challenges traditional machine learning approaches especially when the number of samples is less than the number of descriptors. The advantage of MT-DNN model is to extract information from related tasks and our numerical results show that the predictions do benefit from MTL architecture. For models using ESTDs, physical descriptors and all descriptors, the R^2 has been improved from around 0.5 to 0.788, 0.705, and 0.726, respectively. It is worthy to mention that our ESTDs yield the best results, which proves the power of persistent homology. This result suggests that by learning related problems jointly and extracting shared information from different data sets, MT-DNN architecture can simultaneously perform multiple prediction tasks and enhance performances especially on small data sets.

5.2.3.3 Tetraphymena pyriformis IGC_{50} test set

IGC_{50} set is the second largest QSAR toxicity set that we want to study. The diversity of molecules in IGC_{50} set is low and the coverage of TEST methods is relatively high compared to previous LC_{50} sets. As shown in Table 5.14, the R^2 of different TEST methods fluctuates from 0.600 to 0.764 and Test consensus prediction again yields the best result for TEST software with R^2 of 0.764. As for our models, the R^2 of MT-DNN with different descriptors spans a range of 0.038 (0.732 to 0.770),

Table 5.13: Comparison of prediction results for the Daphnia magna LC_{50} test set.

Method	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical [133]	0.695	0.151	0.981	0.979	0.757	0.886
Single Model [133]	0.697	0.152	1.002	0.993	0.772	0.871
FDA [133]	0.565	0.257	0.987	1.190	0.909	0.900
Group contribution [133]	0.671	0.049	0.999	0.803[a]	0.620[a]	0.657
Nearest neighbor [133]	0.733	0.014	1.015	0.975	0.745	0.871
TEST consensus [133]	0.739	0.118	1.001	0.911	0.727	0.900
Results with ESTDs						
RF	0.441	1.177	0.957	1.300	0.995	1.000
GBDT	0.467	0.440	0.972	1.311	0.957	1.000
ST-DNN	0.446	0.315	0.927	1.434	0.939	1.000
MT-DNN	**0.788**	0.008	1.002	**0.805**	**0.592**	1.000
Consensus	0.681	0.266	0.970	0.977	0.724	1.000
Results with only auxiliary molecular descriptors						
RF	0.479	1.568	0.963	1.261	0.946	1.000
GBDT	0.495	0.613	0.959	1.238	0.926	1.000
ST-DNN	0.430	0.404	0.921	1.484	1.034	1.000
MT-DNN	0.705	0.009	1.031	0.944	0.610	1.000
Consensus	0.665	0.359	0.945	1.000	0.732	1.000
Results with all descriptors						
RF	0.460	1.244	0.955	1.274	0.958	1.000
GBDT	0.505	0.448	0.961	1.235	0.905	1.000
ST-DNN	0.459	0.278	0.933	1.407	1.004	1.000
MT-DNN	0.726	0.003	1.017	0.905	0.590	1.000
Consensus	0.678	0.282	0.953	0.978	0.714	1.000

[a] these values are inconsistent with $R^2 = 0.671$.

which indicates that our MT-DNN not only takes care of overfitting problem but also is insensitive to datasets. Although ESTDs slightly underperform compared to physical descriptors, its MT-DNN results are able to defeat most TEST methods except FDA method. When all descriptors are used, predictions by GBDT and MT-DNN outperform TEST consensus, with R^2 of 0.787 and RMSE of 0.455 log(mol/L). The best result is again given by consensus method using all descriptors, with R^2 of 0.802 and RMSE of 0.438 log(mol/L).

Table 5.14: Comparison of prediction results for the Tetraphymena Pyriformis IGC_{50} test set.

Method	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical [133]	0.719	0.023	0.978	0.539	0.358	0.933
FDA [133]	0.747	0.056	0.988	0.489	0.337	0.978
Group contribution [133]	0.682	0.065	0.994	0.575	0.411	0.955
Nearest neighbor [133]	0.600	0.170	0.976	0.638	0.451	0.986
TEST consensus [133]	0.764	0.065	0.983	0.475	0.332	0.983
Results with ESTDs						
RF	0.625	0.469	0.966	0.603	0.428	1.000
GBDT	0.705	0.099	0.984	0.538	0.374	1.000
ST-DNN	0.708	0.011	1.000	0.537	0.374	1.000
MT-DNN	0.723	0.000	1.002	0.517	0.378	1.000
Consensus	0.745	0.121	0.980	0.496	0.356	1.000
Results with only auxiliary molecular descriptors						
RF	0.738	0.301	0.978	0.514	0.375	1.000
GBDT	0.780	0.065	0.992	0.462	0.323	1.000
ST-DNN	0.678	0.052	0.972	0.587	0.357	1.000
MT-DNN	0.745	0.002	0.995	0.498	0.348	1.000
Consensus	0.789	0.073	0.989	0.451	0.317	1.000
Results with all descriptors						
RF	0.736	0.235	0.981	0.510	0.368	1.000
GBDT	0.787	0.054	0.993	0.455	0.316	1.000
ST-DNN	0.749	0.019	0.982	0.506	0.339	1.000
MT-DNN	0.770	0.000	1.001	0.472	0.331	1.000
Consensus	**0.802**	0.066	0.987	**0.438**	**0.305**	1.000

5.2.3.4 Oral rat LD_{50} test set

The oral rat LD_{50} set contains the largest molecule pool with 7413 compounds. However, none of methods is able to provide a 100% coverage of this data set. The results of single model method or group contribution method were not properly built for the entire set [133]. It was noted that LD_{50} values of this data set are relatively difficult to predict as they have a higher experimental uncertainty [156]. As shown in Table 5.15, results of two TEST approaches, i.e., Single Model and Group contribution, were not reported for this problem. The TEST consensus result improves overall prediction accuracy of other TEST methods by about 10 %, however, other non-consensus methods all yield low R^2 and high RMSE.

For our models, all results outperform those of non-consensus methods of TEST. In particular,

GBDT and MT-DNN with all descriptors yield the best (similar) results, giving slightly better results compared to TEST consensus. Meanwhile, our predictions are also relatively stable for this particular set as R^2s do not essentially fluctuate. It should also be noted that our ESTDs have slightly higher coverage than physical descriptors (all combined descriptors) since 2 molecules in the test set that contains As element cannot be properly optimized for energy computation. However this is not an issue with our persistent homology computation. Consensus method using all descriptors again yield the best results for all combinations, with optimal R^2 of 0.653 and RMSE of 0.568 log(mol/kg).

Table 5.15: Comparison of prediction results for the Oral rat LD_{50} test set.

Method	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical [133]	0.578	0.184	0.969	0.650	0.460	0.876
FDA [133]	0.557	0.238	0.953	0.657	0.474	0.984
Nearest neighbor [133]	0.557	0.243	0.961	0.656	0.477	0.993
TEST consensus [133]	0.626	0.235	0.959	0.594	0.431	0.984
Results with ESTDs						
RF	0.586	0.823	0.949	0.626	0.469	0.999
GBDT	0.598	0.407	0.960	0.613	0.455	0.999
ST-DNN	0.601	0.006	0.991	0.612	0.446	0.999
MT-DNN	0.613	0.000	1.000	0.601	0.442	0.999
Consensus	0.631	0.384	0.956	0.586	0.432	0.999
Results with only auxiliary molecular descriptors						
RF	0.597	0.825	0.946	0.619	0.463	0.997
GBDT	0.605	0.385	0.958	0.606	0.455	0.997
ST-DNN	0.593	0.008	0.992	0.618	0.447	0.997
MT-DNN	0.604	0.003	0.995	0.609	0.445	0.997
Consensus	0.637	0.350	0.957	0.581	0.433	0.997
Results with all descriptors						
RF	0.619	0.728	0.949	0.603	0.452	0.997
GBDT	0.630	0.328	0.960	0.586	0.441	0.997
ST-DNN	0.614	0.006	0.991	0.601	0.436	0.997
MT-DNN	0.626	0.002	0.995	0.590	0.430	0.997
Consensus	**0.653**	0.306	0.959	**0.568**	**0.421**	0.997

CHAPTER 6

DISCUSSION

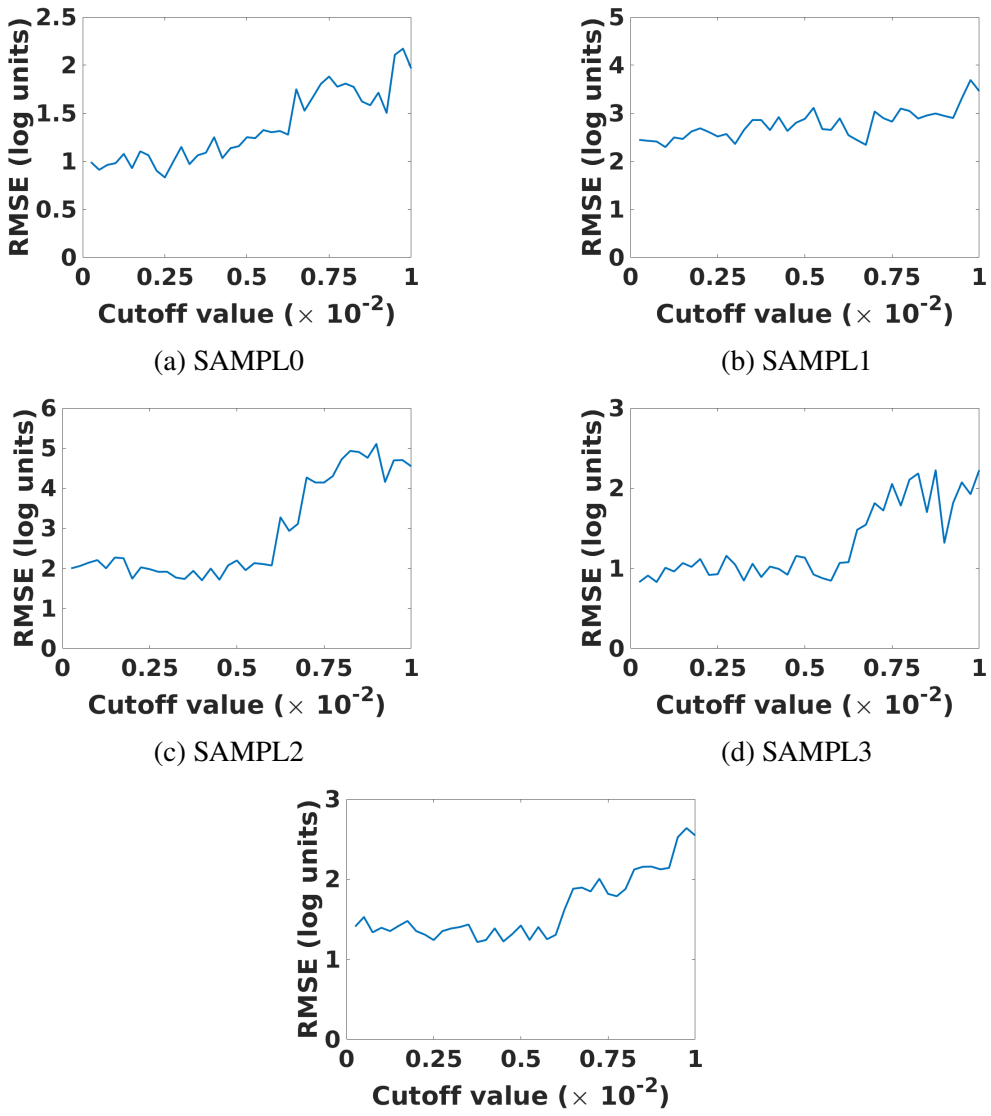
We split the discussion chapter into several parts, namely, solvation free energy, partition coefficient and aqueous solubility, various toxicity endpoints. Within each part, we review model performances of different datasets, and also discuss how current models can be potentially improved.

6.1 Solvation free energy

6.1.1 Descriptor importance analysis

An important concern for machine learning is descriptor importance. In order to analyze this issue, we rank all descriptors by their importance and consequently generate 40 different sets of feature combinations. Note that the descriptor importance here refers to Gini importance [157] weighted by the number of trees in a forest calculated by our baseline methods. We train models with different numbers of descriptors to examine their predictive performances on test sets. More specifically, the protocol to select descriptors relies on a series of descriptor importance cutoffs, equally spaced between 0 and 0.01, with features whose importance is greater than the given cutoff value being selected.

Figure 6.1 represents the RMSEs of predicted solvation energy of SAMPL molecules against different descriptor importance cutoffs. When the feature importance cutoff value is large, the number of features is small, and RMSE is typically large too. The performance is getting better when the importance cutoff value is relatively small. However, further reduction in the cutoff value does not necessarily improve the prediction accuracy and may result in worse performance. Indeed, a suitable cutoff value can benefit overall performance. Cutoff value of 2.5×10^{-3} appears to be a good choice in our case according to our descriptor importance analysis.



(a) SAMPL0

(b) SAMPL1

(c) SAMPL2

(d) SAMPL3

(e) SAMPL4

Figure 6.1: Descriptor importance cutoff versus RMSE for all test sets with AM1-BCC charge and MBondi2 radius parametrization. The larger cutoff value is, the smaller number of features is selected.

6.2 Partition coefficient and aqueous solubility

6.2.1 ESTDs for small molecules

As the previous results indicate, there exists a common descriptor representation for partition coefficient and solubility prediction. Our descriptors come from two different categories – one that is computed solely by ESPH, and the other one that has been widely used in the development of

QSAR models. Although the number of ESTDs is small, it turns out that our topological descriptor representation of molecules has a very strong predictive power compared to baseline method. Our ESTDs highlight atom type information and are able to retain intra-molecular interactions via a filtration prcoess.

6.2.2 Multitask learning

The goal of multitask learning is to learn commonalities between different tasks, and to simultaneously improve model performances. Partition coefficient and aqueous solubility are trained jointly and substantial improvements over single task models are observed. Our results suggest that there exists shared information across these two tasks that can benefit prediction accuracy. Indeed, the original motivation for predicting $\log P$ and $\log S$ is that both coefficients closely relate to the extent to which a compound dissolves in solvents. By comparing our MT-DNN with gradient boosting trees, we find that it is beneficial to learn partition coefficient and aqueous solubility models together. Our MT-ESTD models achieve satisfactory results on various partition coefficient and aqueous solubility data sets, some of which are the state-of-the-art to our best knowledge. Moreover, ESTDs alone can give very accurate predictions, bringing us new insights by ESPH computations. In addition to ESTDs, commonly-used 2D descriptors also help to improve the overall accuracy. Learning these two related properties together boosts the overall model performances.

6.2.3 Predictive power for $\log P$ and $\log S$

We have shown that a common set of ESTDs can be used to accurately predict $\log P$ and $\log S$. However, we also notice that the performances of ESTDs on $\log P$ are generally better than those of $\log S$. One major reason is that the size of $\log S$ training set is small comparing to the size of $\log P$ training set, and it is difficult to fulfill the potential of MT-DNN algorithms. Also the (descriptor)/(training sample size) ratio of $\log S$ is much lower than that of $\log P$, and MT-DNN and GBDT models are likely to overfit due to the large number of fitting parameters. However, MT-DNN is still able to take advantage of $\log P$ prediction tasks - for Klopman's and Hou's test

sets, MT-DNN achieves better results than GBDT. We believe that log P and log S can be predicted simultaneously using MT-DNN architectures and it could be beneficial for both prediction tasks.

6.3 Toxicity endpoints prediction

6.3.1 The impact of descriptor selection and potential overfitting

To deal with descriptor redundancy and overfitting, four different sets of high-importance descriptors are selected by a threshold to perform prediction tasks, in a similar way as discussed in previous subsection. More specifically, we rank all descriptors according to their feature importance and use various feature importance thresholds as a selection protocol. Four different values are chosen (2.5e-4, 5e-4, 7.5e-4 and 1e-4) and the results using MT-DNN are shown in Table 6.1. Results for the other three remaining sets are provided in Appendix.

Table 6.1: Results of selected descriptor groups for LC50 set

Threshold	# of descriptors	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
0.0	1030	0.769	0.009	1.014	0.716	0.466	1.000
2.5e-4	411	0.784	0.051	0.971	0.685	0.459	1.000
5e-4	308	0.764	0.062	0.962	0.719	0.470	1.000
7.5e-4	254	0.772	0.064	0.958	0.708	0.468	1.000
1e-3	222	0.764	0.063	0.963	0.717	0.467	1.000

Table 6.1 shows performance with respect to different numbers of descriptors. When the number of descriptors is increased from 222, 254, 308, 411 to 1030, RMSE does not increase and R^2 does not change much. This behavior suggests that our models are essentially insensitive to the number of descriptors and thus there is little overfitting. MT-DNN architecture takes care of overfitting issues by successive feature abstraction, which naturally mitigates noise generated by less important descriptors. MT-DNN architecture can also potentially take advantage across related tasks, which in turn reduces the potential overfitting on single data set by the alternative training procedure.

Similar behaviors have also been observed for the remaining three data sets, as presented in Appendix. Therefore our MT-DNN architecture is very robust against descriptor selection and can avoid overfitting.

6.3.2 The predictive power of ESTDs for toxicity

One of the main objectives of this study is to understand toxicity of small molecules from a topological point of view. It is important to see if ESTDs alone can match those methods proposed in T.E.S.T. software. When all ESTDs and MT-DNN architecture are used for toxicity prediction, we observe following results:

- LC50 set and LC50DM set. Models using only ESTDs achieve higher accuracy than T.E.S.T. consensus method.
- LD50 set. Consensus result of ESTDs tops T.E.S.T. software in terms of both R^2 and RMSE and MT-DNN results outperform all non-consensus T.E.S.T methods.
- IGC50 set. ESTDs are slightly underperformed than T.E.S.T consensus. However, MT-DNN with ESTDs still yield better results than most non-consensus T.E.S.T methods except FDA.

It is evident that our ESTDs along with MT-DNN architecture have a strong predictive power for all kinds of toxicity endpoints. The ability of MT-DNN to learn from related toxicity endpoints has resulted in a substantial improvement over ensemble methods such as GBDT. Along with physical descriptors calculated by our in-house MIBPB, we can obtain state-of-the-art results for all four quantitative toxicity endpoints.

6.3.3 Alternative element specific networks for generating ESTDs

Apart from the element specific networks proposed in Table 4.8, we also use alternative element specific networks listed below in Table 6.2 to perform the same prediction tasks. Instead of using two types of element-specific networks, we only consider two-element networks, which essentially puts more emphasis on intra-molecular interaction aspect. Eventually, this new construction yields 30 different element specific networks (9+8+7+6), and a total of 840 ESTDs (30×28) are calculated and used for prediction. On LC50 set, IGC50 set and LD50 set, overall performances of the new ESTDs can be improved slightly. However on LC50-DM set, the accuracy is comparably lower (still

Table 6.2: Alternative element specific networks used to characterize molecules

Network type	Element specific networks
Two-element	$\{b_i, c_j\}$, where $b_i \in \mathcal{B}$, $c_j \in \mathcal{C}$, $i \in \{1 \ldots 3\}$, $j \in \{1 \ldots 9\}$, and $i < j$, where $\mathcal{B}$={H, C, N, O} and $\mathcal{C}$={H, C, N, O, F, P, S, Cl, Br, I}.

higher than T.E.S.T consensus). Detailed performances of these ESTDs are presented in Appendix. Thus the predictive power of our ESTDs is not sensitive to the choice of element specific networks as long as reasonable element types are included.

6.3.4 A potential improvement with consensus tools

In this work, we also propose consensus method as discussed in Section 5.2.3. The idea of consensus is to train different models on the same set of descriptors and average across all predicted values. The underlying mechanism is to take advantage of system errors generated by different machine learning algorithms with the potential to reduce bias for the final prediction.

As we notice from Section 5.2.3, consensus method offers a considerable boost in prediction accuracy. For reasonably large sets except LC50-DM set, consensus models turn out to give the best predictions. When it comes to small set (LC50-DM set), consensus models perform worse than MT-DNN. It is likely due to the fact that large number of descriptors may cause overfitting issues for most machine learning algorithms, and consequently generate large deviations, which eventually result in a large error of consensus method. Thus, it should be a good idea to perform prediction tasks with both MT-DNN and consensus methods, depending on the size of data sets, to take advantage of both approaches.

CHAPTER 7

THESIS CONTRIBUTION AND FUTURE WORK

In this chapter, thesis contribution is highlighted for predictions of small molecule properties, in terms of topological modeling and machine learning algorithms. The thesis then ends with some perspectives on future work.

7.1 Solvation free energy

Implicit solvent models intuitively split the total solvation free energies into polar and nonplar contributions. However, polar and nonpolar interactions are coupled and interdependent during solvation process. A novel framework is proposed to break the polar-nonpolar division used in implicit solvent models and treat polar and nonpolar contributions on an equal footing, based on the assumption that there exists a microscopic descriptor vector that can uniquely characterize a molecule and distinguish it from other molecules.

To validate the proposed method, we adopt a large dataset of 668 molecules collected in our earlier work [4]. We propose two sets of descriptors to train the quantitative models: one set with polar features are highly correlated with solvation free energies of this dataset and the other set with both polar and onpolar features. Although non-polar features such as atomic area are not highly correlated with solvation free energy, the inclusion of nonpolar features improves the overall performance of the present method. Highly accurate solvation free energy prediction is confirmed by both the leave-one-out test over 668 molecules and the prediction of five SAMPL test sets, namely, SAMPL0, SAMPL1, SAMPL2, SAMPL3 and SAMPL4. Finally, we consider a test set of 94 molecules and its associated training set [1] for a comparison of the present method and a classic solvation model based on weighted solvent accessible surface area [1].

7.2 Partition coefficient and aqueous solubility

Partition coefficient and aqueous solubility are among the most important physical properties of small molecules and have significant applications to drug design and discovery in terms of lipophilic efficiency. Based on chemical and physical models, a wide variety of computational methods has been developed in the literature for the theoretical predictions of partition coefficient and aqueous solubility.

Present work introduces an algebraic topology based method, element specific persistent homology (ESPH), for simultaneous partition coefficient and aqueous solubility predictions. ESPH offers an unconventional representation of small molecules in terms of multiscale and multicomponent topological invariants. Here the multiscale representation is inherited from persistent homology, while the multicomponent formulation is developed to retain essential chemical information during the topological simplification of molecular geometric complexity. Therefore, the present ESPH gives a unique representation of small molecules that cannot be obtained by any other methods. Although ESPH representation of molecules cannot be literally translated into a physical interpretation, it systematically and comprehensively enciphers chemical and physical information of molecules into scalable topological invariants, and thus is ideally suited for machine learning/deep learning algorithms to decipher such information.

To predict partition coefficient and aqueous solubility, we integrate ESPH with advanced machine learning methods, including gradient boosting tree, random forest, and deep neural networks to construct topological learning strategies. Since partition coefficient and aqueous solubility are highly correlated to each other, we develop a common set of ESPH based descriptors, called element specific topological descriptors (ESTDs), to represent both properties. This approach enables us to perform simultaneous predictions of partition coefficient and aqueous solubility using a topology based multi-task deep learning strategy.

To test the representational of ESPH and the predictive power of the proposed topological multi-task deep learning strategy, we consider some commonly used data sets, including two benchmark test sets, for partition coefficient, as well as additional solubility data sets. Extensive

cross validations and benchmark tests indicate that the proposed topological multi-task strategy offers some of the most accurate predictions of partition coefficient and aqueous solubility.

7.3 Toxicity endpoints

Toxicity refers to the degree of damage a substance on an organism, such as an animal, bacterium, or plant, and can be qualitatively or quantitatively measured by experiments. Experimental measurement of quantitative toxicity is extremely valuable, but is typically expensive and time consuming, in addition to potential ethic concerns. Theoretical prediction of quantitative toxicity has become a useful alternative in pharmacology and environmental science. A wide variety of methods has been developed for toxicity prediction in the past. The performances of these methods depend not only on the descriptors, but also on machine learning algorithms, which makes the model evaluation a difficult task.

We introduce a series of novel descriptors, called element specific topological descriptor (ESTD), for the characterization and prediction of toxicity endpoints. Additionally, physical descriptors based on established physical models are also developed to enhance the predictive power of ESTDs. These new descriptors are then combined with a variety of advanced machine learning algorithms to demonstrate their capability in quantitative toxicity analysis.

Four quantitative toxicity data sets, i.e., 96 hour fathead minnow LC_{50} data set (LC_{50} set), 48 hour Daphnia magna LC_{50} data set (LC_{50}-DM set), 40 hour Tetrahymena pyriformis IGC_{50} data set (IGC_{50} set), and oral rat LD_{50} data set (LD_{50} set), are used in the present study. Comparison has also been made to the state-of-art approaches given in the T.E.S.T website at `https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test` by United States Environmental Protection Agency. Our numerical experiments indicate that the proposed ESTDs are as competitive as individual methods in T.E.S.T. Aided with physical descriptors and MT-DNN architecture, ESTDs are able to establish state-of-the-art predictions for quantitative toxicity data sets. Additionally, MT-DNN models are typically more accurate than ensemble methods.

It is worthy to note that the proposed new descriptors are very easy to generate and thus have

almost 100% coverage for all molecules, indicating their broader applicability to practical toxicity analysis and prediction. In fact, our topological descriptors are much easier to construct than physical descriptors, which depend on physical models and force fields. The present work indicates that ESTDs are a new class of powerful descriptors for small molecules.

7.4 Future work

In this section, we propose the future work of this thesis from several different aspects, following cheminformatics, topological modeling and machine learning point of view.

First of all, there are still many other useful chemical/phyiscal properties to be studied. Properties such as boiling point, viscosity, boiling point, and melting point of a chemical can all potentially be predicted following the same QSAR pipeline. Furthermore, qualitative properties can also be incorporated into our current framework. For instance, Tox21 data challenge (`https://tripod.nih.gov/tox21/challenge/`), although also focuses on toxicity predictions, proposes 12 different qualitative toxicity endpoints for participants to predict. After all, both quantitative and qualitative properties would be essential for drug design and drug discovery.

Secondly, our current ESPH framework can be further improved. From the idea of primitive PH, we have developed ESPH, and further modified ESPH to better characterize small molecules by eliminating covalent bonding. Proposed ESPH, however, mainly takes care of geometric information. A potential approach is to incorporate electrostatic persistence or resolution based persistence to generate more ESTDs. Moreover, it may also be beneficial to construct complexes (such as Alpha complex) in addition to Vietoris-Rips complex or to use different sets of distance metrics. We firmly believe that topological modeling has tremendous potential in cheminformatics and bioinformatics applications.

Last but not least, it is promising that our current molecular descriptors can benefit from the development of machine learning techniques. Neural network based models, such as convolutional neural networks (CNN), generative adversarial networks (GAN) and capsule networks (CN), have all demonstrated success or achieved preliminary results when applying to biomolecule modeling.

For example, one can take advantage of image processing by splitting atoms of different element types into different channels (analogous to RBG channel for image) and analyzing molecular descriptors based on locations. The potential of machine learning can never be overestimated.

APPENDICES

APPENDIX A

SUPPLEMENTARY MATERIALS FOR SOLVATION FREE ENERGY PREDICTION

A.1 A list of polar descriptors used for solvation free energy prediction

Table A.1: Microscopic polar features with high correlations to the solvation free energy used in this study

Feature Name
Sum of atomic reaction field energy
Sum of the absolute value of atomic reaction field energy
Sum of H atomic reaction field energy
Sum of the absolute value of H atomic reaction field energy
Sum of O atomic reaction field energy
Sum of the absolute value of O atomic reaction field energy
Minimum value of atomic reaction field energy
Maximum of the absolute value of reaction field energy
Minimum value of H atomic reaction field energy
Maximum of the absolute value of H atomic reaction field energy
Mean of atomic reaction field energy
Mean of the absolute value of atomic reaction field energy
Variance of atomic reaction field energy
Variance of the absolute value of reaction field energy
Variance of H atomic reaction field energy
Variance of the absolute value of H atomic reaction field energy
Sum of the absolute value of atomic charge
Sum of H atomic charge
Sum of the absolute value of H atomic charge
Sum of O atomic charge
Sum of the absolute value of O atomic charge
Minimum of atomic charge
Maximum of the absolute value of atomic charge
Maximum of H atomic charge
Maximum of the absolute value of H atomic charge
Mean of the absolute value of atomic charge
Variance of the atomic charge
Variance of the absolute value of atomic charge
Variance of the absolute value of H atomic charge
Variance of H atomic charge

A.2 Leave-one-out results

Table A.2 shows the query of each molecule and leave one out results of AM1-BCC charge and MBondi2 radius parameterizations with selected polar features.

Table A.2: Molecules with corresponding query number, and eave one out results with AM1-BCC charge and MBondi2 radius

PubChem ID	Query Number	Experimental Value	Predicted Value
10008	1	-0.8	-0.535
10041	1	2.51	1.893
101	2	-9.52	-7.801
1031	2	-4.85	-5.018
1032	2	-6.46	-6.244
10326	1	-1.46	-0.877
10371	3	-7.1	-7.277
10399	2	-5.48	-3.895
10405	2	-2.65	-2.505
10422	3	-5.56	-4.679
10430	2	-6.09	-6.315
10461	5	-3.92	-3.353
10486	5	1.09	-0.379
1049	3	-4.69	-5.326
10541	3	-8.7	-10.266
10553	1	1.31	0.767
10566	1	0.4	1.337
1068	6	-1.61	-1.418
10686	1	-1.21	-0.935
10687	2	-6.16	-6.429
10722	2	-3.93	-4.715
10740	1	-2.78	-3.035
107828	3	-17.17	-18.016
10822	3	-4.73	-4.581
10823	3	-4.59	-4.856
10824	3	-7.29	-7.762
10864	1	-0.23	-0.442
10870	1	0.16	-0.167
10882	2	-2.49	-2.445
10899	5	-0.33	-0.260
10903	2	-2.1	-2.414
10907	1	2.89	2.773
10926	5	-0.7	-1.395
10943	5	-0.98	-1.125
10974	3	-1.88	-2.334
10977	5	-0.07	-0.515

Table A.2 (cont'd)

10992	5	0.0	0.087
11	1	-1.79	-1.228
11000	5	-0.99	-0.912
11095291	2	0.1	-1.943
11124	6	-2.93	-2.928
11126	3	-1.38	-1.475
11137	2	-9.62	-6.729
11182	1	-1.66	-1.559
11229	1	2.56	2.535
11239	2	1.83	1.502
11251	1	-3.24	-3.015
11260	2	2.52	2.545
11264	1	-3.88	-3.905
11269	2	2.56	2.870
11271	1	-2.74	-3.442
11304	1	-2.82	-2.428
11327	2	-2.47	-2.770
11335	1	-5.26	-5.531
11386	1	-2.78	-2.485
11387	1	-2.63	-2.800
1140	1	-0.9	-0.797
11416	3	1.58	2.105
11417	3	-5.22	-4.831
11420	2	-4.82	-4.853
11428	3	-4.35	-4.182
1146	5	-3.2	-5.502
11507	1	2.71	2.815
11513	2	-4.01	-3.224
11519	1	2.97	2.870
11523	1	2.11	1.670
11526	3	-4.72	-4.959
11542	1	2.88	2.547
11543	2	-3.92	-3.999
11565	3	-4.84	-4.516
11574	1	0.67	1.255
11582	1	2.93	2.585
11583	2	-3.28	-3.102
11587	1	0.93	1.170
11597	1	1.58	1.760
11598	1	1.01	1.170
11610	1	1.66	1.680
11611	1	1.68	1.675
11638	1	-0.22	1.047

Table A.2 (cont'd)

11643	5	-0.77	-0.775
11670	2	-3.91	-4.364
11721	2	-2.01	-2.215
1174	3	-16.59	-17.440
11903	1	-0.85	-0.535
11904	1	-1.24	-1.244
11940	1	-3.78	-2.800
12009	3	-14.0	-13.380
12016	5	0.04	0.132
12083	3	-6.88	-6.300
12101	2	-6.25	-5.703
12160	1	-0.95	-1.170
12178	2	-4.06	-4.539
12180	2	-2.83	-2.650
12206	2	-2.56	-2.593
12217	2	-2.11	-2.116
12232	6	-1.83	-1.535
12245	5	0.07	0.045
12257	2	-2.02	-2.802
12264	6	-1.21	-1.314
12287	3	-4.01	-4.811
12291	3	-3.88	-3.671
12302	3	-3.09	-2.827
12309	1	0.01	0.248
12309460	4	-4.23	-4.692
12321	6	-1.43	-2.143
12332	3	-9.31	-8.857
12340	2	-1.81	-1.912
12348	2	-2.51	-2.350
12350	1	0.6	0.248
12370	1	0.71	0.675
12371	5	0.29	-0.232
12375	2	-3.54	-3.127
12418	5	-1.43	-1.047
12463	5	-1.34	-1.943
12468	5	-1.62	-1.262
12508	2	-2.22	-2.412
12580	2	-5.21	-6.061
12586	2	-4.1	-3.166
126	2	-8.83	-9.038
12720	1	0.04	-0.250
12724	1	1.91	1.680
12732	1	0.29	0.530

Table A.2 (cont'd)

12741	2	-2.34	-2.765
12896	3	-11.24	-9.358
12986	1	1.47	1.680
13	2	-1.12	-1.047
13004	3	-4.61	-3.668
13019	7	-10.17	-9.746
13081	2	-6.5	-5.217
13187	2	-2.49	-2.709
13207	3	-3.9	-3.442
13238893	3	-1.92	-6.087
13263	3	-7.65	-7.663
13389	3	-7.0	-6.837
13394	3	-9.65	-8.318
13450	5	-6.68	-8.389
135191	2	-20.52	-23.559
13529	2	-3.77	-4.357
13567	3	-2.09	-2.012
13855	2	-2.92	-1.993
138747	3	-4.29	-4.243
138975	3	-11.95	-7.774
1390	3	-8.41	-6.564
14109	1	-0.04	-0.535
141897	2	-5.73	-5.918
14215	3	-3.64	-4.378
14276	3	-4.4	-3.162
14282	3	-7.62	-7.206
14315	1	-2.4	-2.800
144381	3	-6.4	-7.051
144702	4	-4.59	-4.107
15050	2	-4.42	-4.368
15413	2	-2.21	-2.220
15546	3	-16.43	-15.724
15600	1	3.16	2.930
15625	4	-3.37	-3.758
156391	2	-10.21	-9.925
15758	3	-17.74	-16.703
16003	3	-4.13	-3.644
16270	1	2.13	2.585
16295	1	-0.18	0.247
16318	5	-2.98	-0.978
16441	2	-4.09	-3.290
16628	1	1.58	1.993
16666	2	-3.2	-3.507

Table A.2 (cont'd)

1672	3	-6.93	-7.207
16900	2	-3.43	-3.437
17190	3	-8.94	-11.688
1732	4	-6.79	-6.848
174	2	-9.3	-7.136
176	2	-6.69	-7.501
177	2	-3.5	-3.250
17739	2	-5.26	-6.844
17756	2	-6.4	-5.745
178	3	-9.71	-10.312
180	2	-3.8	-3.945
18636	5	-4.53	-1.125
18927701	1	-0.8	0.246
18937	1	1.05	0.597
19041	1	2.93	2.895
19540	1	2.55	2.502
19878	3	-4.46	-4.072
20419	3	-7.44	-10.787
20528	3	-14.21	-9.851
20748	2	-3.3	-2.412
2078	3	-8.21	-8.459
20848	3	-7.98	-9.430
21075956	3	-4.95	-5.145
21210	2	-5.73	-4.771
21269179	1	-1.29	-0.768
2160	3	-7.43	-6.135
220639	3	-2.82	-2.896
221525	5	-2.32	-2.815
22188	3	-11.14	-9.860
222	2	-4.29	-6.906
22227	2	-3.75	-3.752
222536	7	-4.97	-5.298
223106	1	1.07	0.247
22386	2	-4.39	-4.420
2244	2	-9.94	-9.319
2268	7	-10.03	-6.451
2319	3	-3.51	-2.982
2331	3	-11.0	-10.803
240	2	-4.02	-4.978
244	2	-6.62	-5.745
249266	5	-2.69	-2.933
25146	7	-5.74	-8.238
2519	3	-12.64	-10.790

Table A.2 (cont'd)

2566	3	-9.61	-9.117
261	2	-3.18	-3.181
263	2	-4.72	-4.731
26331	3	-5.45	-5.476
264	2	-6.35	-6.268
26447	2	-2.53	-3.442
2730	7	-5.04	-6.853
27588	5	-2.28	-2.661
2879	2	-6.13	-6.548
297	1	2.0	1.680
3017	7	-6.48	-6.172
30209	3	-1.66	-1.928
3030	4	-9.86	-7.827
3031	3	-4.71	-3.673
3048	4	-4.82	-6.154
3059	2	-9.4	-9.804
3100	3	-9.34	-7.454
31234	2	-6.92	-6.060
31242	2	-6.13	-5.747
31246	2	-2.92	-2.986
31249	2	-5.71	-5.298
31260	2	-4.42	-4.568
31265	2	-2.23	-2.178
31268	3	-5.48	-4.395
31272	2	-2.64	-2.573
31275	2	-5.06	-3.895
31276	2	-2.21	-2.505
31285	1	2.06	1.670
31289	2	-2.07	-2.303
31297	7	-4.87	-6.475
31347	7	-8.61	-9.740
31373	5	0.1	-2.514
31420	3	-11.85	-9.586
31423	1	-4.52	-3.741
31645	3	-9.41	-10.178
32594	3	-3.68	-4.959
3283	2	-1.59	-2.125
3286	7	-6.1	-6.834
3301	3	-7.6	-6.557
33135	3	-4.4	-4.702
335	3	-5.9	-6.011
33500	2	-2.45	-3.264
3385	3	-16.92	-18.056

Table A.2 (cont'd)

3394	2	-8.42	-8.897
342	2	-5.49	-6.429
34468	3	-5.66	-7.407
34586	4	-3.56	-4.243
34591	3	-5.04	-5.461
35454	4	-3.81	-4.010
356	1	2.88	2.930
3589	5	-2.55	-3.352
36401	5	-3.48	-2.516
36613	4	-3.67	-3.539
3672	2	-7.0	-6.699
36980	5	-2.46	-2.394
37037	5	-4.4	-2.973
37207	4	-3.52	-3.580
37247	5	-2.16	-2.517
3776	2	-4.74	-4.272
38019	5	-3.04	-3.352
3825	4	-10.78	-19.721
38251	2	-3.71	-3.920
38252	4	-3.1	-3.684
38253	4	-4.05	-3.757
38254	4	-4.15	-3.758
38306	5	-3.17	-3.610
39253	5	-4.61	-2.270
398	3	-3.13	-3.178
4004	7	-8.15	-6.873
402	6	-0.7	-1.207
4044	3	-6.78	-9.428
40818	3	-5.73	-4.518
4101	3	-4.56	-8.587
4109	3	-10.65	-9.717
4116	5	-1.12	-0.640
4130	7	-7.19	-6.462
4156	7	-4.87	-9.685
442474	2	-2.49	-3.586
447466	2	-4.22	-4.218
447907	3	-10.91	-9.807
454	2	-2.29	-2.783
460	2	-5.94	-5.930
46174049	4	-4.82	-4.554
4684	4	-7.03	-6.792
4685	5	-1.01	-1.228
4790	7	-4.37	-6.853

Table A.2 (cont'd)

4837	3	-7.4	-7.050
48889	4	-3.84	-3.888
4929	3	-8.43	-6.704
4933	3	-7.78	-8.305
5216	3	-10.22	-9.381
527	2	-3.43	-3.438
5281168	2	-3.68	-4.445
5283324	2	-3.44	-3.434
52997	3	-20.25	-14.017
52999	3	-15.54	-16.091
53167	3	-7.77	-6.563
5326160	1	1.31	1.217
5326161	1	1.34	0.850
53476	4	-8.68	-7.291
53479	4	-7.78	-8.449
5377791	7	-7.07	-6.513
5541	2	-8.84	-15.756
5569	3	-3.25	-2.799
56160	3	-14.01	-17.548
5793	2	-25.47	-22.599
5802	3	-18.17	-17.100
5853	7	-12.74	-9.281
5899	3	-15.46	-18.053
5943	5	0.08	-1.396
5993	5	-3.44	-2.812
6027	2	-20.52	-23.556
6053	3	-6.66	-7.217
6054	2	-6.79	-6.588
6115	3	-5.49	-6.007
6129	3	-9.45	-10.469
61362	2	-4.51	-3.290
6184	2	-2.81	-2.505
6212	5	-1.08	-0.640
6213	7	-10.08	-6.889
6214	5	-0.64	-1.263
6228	3	-7.81	-7.456
6251	2	-23.62	-23.886
6276	2	-4.57	-4.633
6278	5	-0.19	-1.138
63079	5	-4.38	-2.972
63088	5	-3.61	-2.972
6324	1	1.83	2.105
6325	1	1.28	1.091

Table A.2 (cont'd)

6326	1	-0.01	-0.024
6327	5	-0.55	-0.525
6329	3	-4.55	-5.738
6334	1	2.0	2.105
6335	1	-0.48	-0.197
6337	5	-0.63	-0.455
6338	5	-0.59	-0.105
6341	3	-4.5	-4.712
6342	3	-3.88	-4.517
6343	6	-1.14	-1.125
6344	5	-1.31	-1.047
6351	1	0.75	1.670
6360	1	2.3	2.379
6361	5	-0.25	-0.188
6365	5	-0.84	-1.047
6366	5	0.25	-0.168
6368	1	-0.11	-0.766
6372	5	-0.5	1.343
6373	1	0.81	0.000
6375	3	-4.02	-4.096
637564	2	-4.63	-3.786
637566	2	-4.45	-4.802
638186	5	-0.78	-0.685
6386	2	-4.47	-5.047
6391	5	1.69	0.994
6392	5	2.52	1.717
639662	1	1.66	1.680
6403	1	2.51	2.502
6405	2	-4.43	-4.364
6408	5	0.06	-1.394
6409	2	-4.31	-4.595
64151	2	-4.46	-3.226
6416	2	-3.11	-3.920
6419	5	-1.23	-0.845
6423	5	-1.45	-4.483
6427	5	0.82	0.042
6428	5	1.77	0.973
6429	5	2.32	2.123
6430	5	2.87	0.614
643820	2	-4.78	-4.485
643833	5	-1.17	-1.262
6441	2	-4.88	-5.480
64689	2	-25.47	-22.599

Table A.2 (cont'd)

6497	7	-5.1	-7.770
6544	2	-5.18	-3.819
6556	1	2.38	2.547
6557	1	0.68	0.597
6560	2	-4.5	-4.897
6561	2	-2.86	-3.194
6563	5	0.0	-0.105
6564	5	-1.27	-2.370
6568	2	-4.62	-4.701
6569	2	-3.71	-3.763
6574	5	-1.99	-1.465
6575	5	-0.44	-1.138
6578	3	-9.4	-9.803
6582	3	-10.0	-9.778
6584	2	-3.13	-2.803
6587	3	-3.71	-3.265
6589	1	2.34	2.259
6591	5	-2.37	-1.574
66750	2	-7.75	-8.895
6710	3	-9.44	-9.808
6720	5	-5.22	-0.585
6734	1	-3.15	-1.894
6736	3	-5.88	-5.578
674	3	-4.29	-5.145
679	7	-9.28	-4.596
6809	3	-9.61	-10.127
6845	5	-3.32	-2.810
68510	3	-0.41	-2.980
6853	1	-3.35	-2.427
688400	3	-6.23	-6.264
6895	5	-1.24	-2.670
6896	3	-5.21	-4.913
69027	2	-2.4	-3.058
6944	3	-3.58	-3.645
6946	3	-7.37	-7.436
6947	3	-4.58	-9.511
6950	3	-6.23	-6.003
69689	3	-13.6	-11.641
69720	2	-4.04	-5.281
6993809	2	-4.2	-4.183
6997	2	-5.66	-6.348
6998	2	-4.68	-7.800
7000	3	-6.12	-7.456

Table A.2 (cont'd)

7002	1	-2.44	-2.800
7005	2	-7.67	-7.315
702	2	-5.0	-4.837
7041	2	-6.96	-6.053
7043	2	-5.33	-4.366
7047	3	-5.72	-5.836
7057	3	-7.47	-6.834
7095	1	-2.7	-3.035
712	2	-2.75	-4.071
7144	2	-5.8	-6.918
7150	2	-3.92	-4.183
7165	2	-3.64	-4.943
7175	2	-9.37	-9.042
7184	2	-8.72	-9.318
7237	1	-0.9	-0.935
7238	5	-1.14	-0.977
7239	5	-1.36	-1.315
7240	3	-4.91	-5.894
7242	3	-5.53	-5.736
7245	4	-4.55	-5.665
7247	1	-0.86	-1.232
7249	2	-6.5	-6.128
7258	4	-7.29	-6.731
7267	2	-5.91	-5.240
727	5	-5.44	-2.660
7270	5	-1.34	-1.573
7282	1	2.51	2.547
7288	2	-3.41	-3.102
7295	4	-4.0	-4.549
7296	1	1.59	1.680
7298	2	-5.49	-4.980
7301	2	-3.3	-2.802
7304	3	-2.89	-3.436
73272	3	-15.83	-13.950
7351	2	-1.69	-2.481
7366	1	-0.44	-0.465
7368	1	-0.25	-0.000
7393	2	-5.91	-5.801
7406	1	-0.3	-0.797
7407	1	-1.24	-1.308
7410	2	-4.58	-4.714
7416	3	-4.12	-4.098
7422	3	-3.45	-3.675

Table A.2 (cont'd)

7423	3	-8.84	-7.406
7456	2	-9.51	-8.900
74626	2	-3.82	-4.512
7463	1	-0.68	-0.935
7475	3	-9.82	-9.629
7476	2	-4.4	-4.541
7498	6	-2.38	-1.958
7500	1	-0.79	-0.397
7501	1	-1.24	-0.935
7503	5	-1.93	-2.668
7505	3	-4.1	-5.145
7506	3	-6.02	-6.781
7515	3	-4.69	-6.085
7519	2	-2.45	-2.599
7520	6	-2.73	-1.636
7523	3	-4.33	-4.703
753	2	-13.43	-10.059
7560	3	-9.13	-10.800
7580	1	-2.82	-2.882
7583	2	-2.87	-3.393
76122	2	-4.1	-3.128
7668	1	-0.53	-0.513
7674	2	-2.22	-2.490
7705	1	-0.4	-0.475
7732	3	-7.48	-6.724
7749	2	-2.68	-2.650
7761	2	-6.0	-5.741
7762	2	-2.49	-2.540
7765	3	-3.28	-2.930
77650	2	-9.8	-7.340
7770	2	-2.28	-2.540
7771	2	-6.01	-5.703
77918	3	-7.63	-8.403
7797	2	-2.3	-2.080
7803	2	-2.44	-2.481
7809	1	-0.8	-1.232
7812	3	-5.9	-5.462
7813	3	-5.57	-5.462
7818	3	-7.58	-6.381
7824	2	-2.49	-2.468
7843	1	2.1	2.313
7844	1	1.38	1.320
7845	1	0.56	0.353

Table A.2 (cont'd)

7846	1	-0.16	-0.250
7848	6	-1.1	-1.135
7850	5	-0.57	-0.640
7852	3	-4.39	-4.516
7854	3	-3.84	-3.376
7858	2	-5.03	-5.480
7865	2	-2.78	-3.080
7892	1	2.51	2.088
7895	2	-3.52	-3.275
79	3	-6.75	-6.088
7903	1	-3.34	-3.563
7907	2	2.83	2.625
7909	2	-3.05	-3.508
7910	3	-3.73	-3.427
7912	3	-3.22	-4.020
79123	2	-1.82	-2.083
7914	3	-0.53	-1.687
79143	2	-18.4	-20.045
7915	1	-2.64	-2.540
7929	3	-0.83	-0.535
7932	3	-5.82	-5.462
7933	4	-6.62	-6.891
7936	3	-4.86	-4.327
7937	3	-4.59	-4.517
7947	1	-0.9	-0.349
7948	2	-6.27	-5.915
795	5	-9.63	-9.141
7950	3	-0.78	-1.110
7956	3	-18.06	-16.249
7962	1	1.7	1.583
7963	3	-4.93	-4.703
79639	3	-22.4	-17.631
7964	5	-1.12	-1.395
7965	3	-4.59	-3.989
7966	2	-5.46	-5.018
7967	2	-4.91	-3.427
7970	3	-4.77	-4.651
7972	3	-6.32	-4.958
7975	3	-4.63	-5.007
7976	3	-5.51	-5.671
7977	3	-4.39	-5.241
7991	2	-6.16	-6.126
7997	2	-2.79	-2.650

Table A.2 (cont'd)

8003	1	2.3	1.893
8004	1	1.68	1.338
8005	5	-0.16	0.186
8007	3	-4.24	-4.441
8008	3	-3.64	-3.192
8012	6	-0.99	-1.162
8018	3	-6.55	-5.543
8019	2	-6.62	-6.428
8020	2	-2.93	-4.000
8021	3	-4.07	-4.858
8025	2	-2.56	-2.783
8027	3	-4.78	-4.412
8028	2	-3.47	-3.081
8030	6	-1.4	-1.535
8038	2	-2.36	-2.650
8051	2	-3.04	-3.077
8052	2	-2.13	-2.540
8058	1	2.48	2.693
8059	5	-2.32	-2.272
8060	3	-4.09	-4.543
8061	3	-3.52	-3.436
8063	2	-3.03	-2.842
8071	2	-4.84	-2.650
8076	2	-6.69	-6.588
8077	6	-1.64	-1.418
8078	1	1.23	1.137
8079	1	0.14	1.378
8082	3	-5.11	-4.287
8083	3	-7.17	-6.226
8091	2	-2.04	-2.282
8093	2	-2.88	-2.842
8095	2	-5.31	-3.943
8102	3	-3.95	-4.286
8103	2	-4.4	-4.110
8114	2	-1.16	-1.509
8115	5	-4.23	-2.270
8118	6	-1.28	-1.418
8121	2	-6.34	-5.969
81226	3	-5.99	-7.609
8125	1	2.08	1.818
8127	3	-3.79	-3.376
8129	2	-4.21	-4.540
8130	2	-2.67	-2.599

Table A.2 (cont'd)

8133	2	-6.25	-6.222
8141	1	3.13	2.870
8143	3	-3.65	-3.989
8148	3	-3.24	-3.222
8163	2	-2.16	-2.283
81713	3	-11.01	-10.472
8174	2	-3.64	-3.507
8252	1	1.32	1.335
8254	2	-1.91	-2.350
8255	1	1.16	1.320
8263	1	3.43	1.552
8302	1	2.31	1.680
8323	3	-9.53	-8.664
8341	3	-11.53	-11.107
8370	5	-2.33	-0.684
84179	3	-9.76	-9.008
8418	1	-3.95	-3.883
8434	2	-9.2	-9.591
84440	2	-5.23	-4.566
8452	2	-4.7	-4.365
8454	3	-3.98	-3.888
8471	3	-3.22	-2.980
8500	2	-4.7	-4.709
85254	3	-8.18	-9.703
8606	3	-9.01	-6.731
8640	3	-7.28	-7.497
8663	2	-8.11	-6.914
8680	1	-0.45	-0.349
8723	2	-4.42	-4.485
878	6	-1.2	-0.983
8857	2	-2.94	-2.883
887	2	-5.1	-5.064
8881	5	-1.89	-2.428
8882	1	0.56	0.665
8892	2	-6.21	-6.347
8894	2	-3.12	-3.167
8900	1	2.67	2.547
8902	3	-3.65	-3.260
8908	2	-2.26	-2.125
8909	2	-0.83	-1.687
8914	2	-3.89	-3.443
9005	2	-4.42	-4.455
9007	2	-7.66	-6.348

Table A.2 (cont'd)

9033	3	-5.42	-5.782
91662	5	-1.96	-2.970
91729	3	-16.23	-16.916
9216	2	-3.15	-3.509
9253	1	1.2	1.046
9265	1	0.8	1.553
9266	1	0.86	1.091
931	1	-2.4	-3.539
9321	7	-9.3	-4.566
93462	4	-6.44	-6.922
9411	3	-9.73	-10.890
94221	2	-4.44	-4.485
949	3	-3.45	-4.793
957	3	-4.09	-4.150
9570071	2	-9.84	-9.119
9589	3	-8.26	-6.727
9595287	3	-10.18	-13.603
9609	6	-1.46	-1.372
962	2	-6.3	-5.047
96257	3	-4.8	-4.351
9707	2	-5.29	-5.746
9774	2	-4.16	-4.193
9775	5	2.51	2.074
980	3	-10.64	-8.973
9818	1	1.07	0.806
9872	2	-4.15	-4.408
9893	2	-1.1	-3.016
991	7	-6.74	-6.860
995	1	-3.88	-3.558
996	2	-6.61	-6.347

A.3 SAMPLx challenge results

Table A.3, A.4, A.5, A.6 and A.7 represent the solvation energy prediction for SAMPL0, SAMPL1, SAMPL2, SAMPL3, SAMPL4 molecules calculated with polar features, respectively. Table A.8, A.9, A.10, A.11 and A.12 represent the solvation energy prediction for SAMPL0, SAMPL1, SAMPL2, SAMPL3, SAMPL4 molecules calculated with all available polar and nonpolar features, respectively.

Table A.3: Solvation energy prediction results for SAMPL0 molecules using selected polar features

PubChem ID	Experimental Value	Predicted Value
223106	1.07	0.155
31275	-5.06	-4.735
7765	-3.28	-2.783
8020	-2.93	-3.574
222536	-4.97	-5.138
7761	-6.0	-6.423
8121	-6.34	-5.941
5541	-8.84	-10.141
12375	-3.54	-3.057
74626	-3.82	-4.409
795	-9.63	-9.872
84179	-9.76	-9.798
81713	-11.01	-10.798
7503	-1.93	-2.429
8115	-4.23	-2.964
9609	-1.46	-1.372
7498	-2.38	-1.958

Table A.4: Solvation energy prediction results for SAMPL1 molecules using selected polar features

PubChem ID	Experimental Value	Predicted Value
6544	-5.18	-4.032
14215	-3.64	-4.749
16003	-4.13	-4.768
6053	-6.66	-7.696
3031	-4.71	-3.257
34468	-5.66	-7.334
5569	-3.25	-2.982
2319	-3.51	-2.982
33500	-2.45	-2.982
56160	-14.01	-16.171
15546	-16.43	-18.067
107828	-17.17	-18.016
91729	-16.23	-18.021
52997	-20.25	-16.157
52999	-15.54	-14.734
9411	-9.73	-10.395
10974	-1.88	-3.220
13567	-2.09	-2.984
79123	-1.82	-2.698
40818	-5.73	-4.807
21075956	-4.95	-5.463
7560	-9.13	-10.910
17190	-8.94	-14.287
20419	-7.44	-11.639
12896	-11.24	-10.456
2566	-9.61	-9.836
6950	-6.23	-6.174
81226	-5.99	-7.206
85254	-8.18	-9.963
31645	-9.41	-11.945
6129	-9.45	-10.908
2078	-8.21	-9.919
20848	-7.98	-9.437
13263	-7.65	-8.699
13450	-6.68	-9.082
8606	-9.01	-7.367
4929	-8.43	-8.577
4933	-7.78	-7.700
22188	-11.14	-10.395
9570071	-9.84	-9.373
9595287	-10.18	-13.643
4109	-10.65	-10.412

Table A.4 (cont'd)

5216	-10.22	-10.923
12309460	-4.23	-4.181
3030	-9.86	-7.827
3048	-4.82	-4.326
46174049	-4.82	-4.328
3589	-2.55	-3.583
5993	-3.44	-3.395
727	-5.44	-2.790
6423	-1.45	-4.702
3286	-6.1	-9.321
4004	-8.15	-9.307
4790	-4.37	-9.313
25146	-5.74	-9.308
13081	-6.5	-9.315
2730	-5.04	-9.311
5853	-12.74	-9.309
5377791	-7.07	-9.301
2268	-10.03	-9.316
991	-6.74	-9.310
4130	-7.19	-9.313
3017	-6.48	-9.308

Table A.5: Solvation energy prediction results for SAMPL2 molecules using selected polar features

PubChem ID	Experimental Value	Predicted Value
18927701	-0.8	0.530
8263	3.43	2.312
3394	-8.42	-6.588
3672	-7.0	-6.217
753	-13.43	-8.779
3825	-10.78	-7.796
3059	-9.4	-9.119
156391	-10.21	-7.508
2244	-9.94	-7.084
135191	-20.52	-19.263
5793	-25.47	-20.759
7175	-9.37	-7.797
7184	-8.72	-7.797
7456	-9.51	-7.797
8434	-9.2	-7.797
15758	-17.74	-17.666
73272	-15.83	-17.666
3385	-16.92	-17.268
5899	-15.46	-17.262
5802	-18.17	-17.675
1174	-16.59	-16.315
7956	-18.06	-17.619
6809	-9.61	-10.142
2519	-12.64	-9.295
10541	-8.7	-11.448
8370	-2.33	-2.969
6720	-5.22	-3.463
31347	-0.64	0.063
6214	-8.61	-9.867

Table A.6: Solvation energy prediction results for SAMPL3 molecules using selected polar features

PubChem ID	Experimental Value	Predicted Value
6324	1.83	2.035
7095	-2.7	-2.428
9216	-3.15	-4.715
37207	-3.52	-4.694
38252	-3.1	-4.694
34586	-3.56	-4.695
36613	-3.67	-4.695
38253	-4.05	-4.695
15625	-3.37	-4.696
35454	-3.81	-4.696
48889	-3.84	-4.696
38251	-3.71	-4.634
38254	-4.15	-4.632
6214	-0.64	0.087
12418	-1.43	-1.443
6419	-1.23	-0.757
6337	-0.63	-0.515
11	-1.79	-2.225
6365	-0.84	-0.657
6574	-1.99	-1.372
6591	-2.37	-1.372
249266	-2.69	-1.940
36980	-2.46	-2.115
37247	-2.16	-2.115
27588	-2.28	-2.225
36401	-3.48	-3.722
63088	-3.61	-3.307
91662	-1.96	-2.225
38019	-3.04	-2.380
63079	-4.38	-2.910
37037	-4.4	-2.380
38306	-3.17	-3.726
39253	-4.61	-3.726
6278	-0.19	0.187
16318	-2.98	-4.206
18636	-4.53	-4.715

Table A.7: Solvation energy prediction results for SAMPL4 molecules using selected polar features

PubChem ID	Experimental Value	Predicted Value
8079	0.14	1.337
10740	-2.78	-2.780
8418	-3.95	-3.726
11903	-0.85	-0.475
11940	-3.78	-2.780
31275	-5.06	-3.427
9216	-3.15	-4.465
10722	-3.93	-4.940
8908	-2.26	-2.275
26447	-2.53	-4.200
16666	-3.2	-3.166
6251	-23.62	-23.886
6997	-5.66	-6.390
8894	-3.12	-3.166
7583	-2.87	-3.245
7043	-5.33	-6.849
16441	-4.09	-2.845
6998	-4.68	-5.932
442474	-2.49	-1.839
8095	-5.31	-3.667
22227	-3.75	-3.165
61362	-4.51	-3.783
94221	-4.44	-4.530
637566	-4.45	-4.715
643820	-4.78	-4.271
21210	-5.73	-4.825
460	-5.94	-6.267
7144	-5.8	-6.635
17739	-5.26	-6.245
7041	-6.96	-5.184
8082	-5.11	-4.020
2160	-7.43	-5.108
77918	-7.63	-9.305
4044	-6.78	-8.255
30209	-1.66	-1.928
96257	-4.8	-5.623
138747	-4.29	-5.093
31420	-11.85	-8.441
6710	-9.44	-10.139
8341	-11.53	-10.140
20528	-14.21	-10.151
8323	-9.53	-7.613

Table A.7 (cont'd)

3100	-9.34	-8.408
53479	-7.78	-7.944
53476	-8.68	-7.883
7258	-7.29	-7.346
93462	-6.44	-4.693

Table A.8: Solvation energy prediction results for SAMPL0 molecules using all features

PubChem ID	Experimental Value	Predicted Value
223106	1.07	-0.137
31275	-5.06	-4.717
7765	-3.28	-2.650
8020	-2.93	-3.370
222536	-4.97	-5.924
7761	-6.0	-6.265
8121	-6.34	-4.709
5541	-8.84	-25.855
12375	-3.54	-2.533
74626	-3.82	-4.686
795	-9.63	-10.133
84179	-9.76	-9.596
81713	-11.01	-9.598
7503	-1.93	-0.662
8115	-4.23	-2.170
9609	-1.46	-1.330
7498	-2.38	-1.372

Table A.9: Solvation energy prediction results for SAMPL1 molecules using all features

PubChem ID	Experimental Value	Predicted Value
6544	-5.18	-4.385
14215	-3.64	-4.262
16003	-4.13	-4.262
6053	-6.66	-8.247
3031	-4.71	-2.672
34468	-5.66	-3.900
5569	-3.25	-2.695
2319	-3.51	-2.703
33500	-2.45	-2.727
56160	-14.01	-16.387
15546	-16.43	-16.068
107828	-17.17	-18.361
91729	-16.23	-18.285
52997	-20.25	-15.588
52999	-15.54	-14.088
9411	-9.73	-11.020
10974	-1.88	-3.029
13567	-2.09	-2.159
79123	-1.82	-3.022
40818	-5.73	-4.990
21075956	-4.95	-5.421
7560	-9.13	-11.112
17190	-8.94	-12.631
20419	-7.44	-11.561
12896	-11.24	-10.571
2566	-9.61	-10.656
6950	-6.23	-6.238
81226	-5.99	-8.910
85254	-8.18	-8.932
31645	-9.41	-12.282
6129	-9.45	-10.462
2078	-8.21	-9.347
20848	-7.98	-11.169
13263	-7.65	-9.850
13450	-6.68	-8.214
8606	-9.01	-7.741
4929	-8.43	-8.199
4933	-7.78	-8.772
22188	-11.14	-11.029
9570071	-9.84	-9.651
9595287	-10.18	-13.044
4109	-10.65	-9.827

Table A.9 (cont'd)

5216	-10.22	-9.443
12309460	-4.23	-3.702
3030	-9.86	-7.995
3048	-4.82	-7.429
46174049	-4.82	-7.451
3589	-2.55	-4.050
5993	-3.44	-3.070
727	-5.44	-2.285
6423	-1.45	-0.243
3286	-6.1	-4.572
4004	-8.15	-4.822
4790	-4.37	-4.503
25146	-5.74	-4.772
13081	-6.5	-4.658
2730	-5.04	-4.649
5853	-12.74	-4.825
5377791	-7.07	-4.783
2268	-10.03	-4.789
991	-6.74	-4.916
4130	-7.19	-5.018
3017	-6.48	-4.373

Table A.10: Solvation energy prediction results for SAMPL2 molecules using all features

PubChem ID	Experimental Value	Predicted Value
18927701	-0.8	0.603
8263	3.43	1.318
3394	-8.42	-6.511
3672	-7.0	-5.764
753	-13.43	-9.164
3825	-10.78	-13.958
3059	-9.4	-8.420
156391	-10.21	-8.337
2244	-9.94	-7.683
135191	-20.52	-20.937
5793	-25.47	-23.679
7175	-9.37	-7.329
7184	-8.72	-6.887
7456	-9.51	-7.729
8434	-9.2	-7.702
15758	-17.74	-16.623
73272	-15.83	-13.920
3385	-16.92	-16.831
5899	-15.46	-15.115
5802	-18.17	-17.318
1174	-16.59	-17.474
7956	-18.06	-14.373
6809	-9.61	-11.710
2519	-12.64	-11.015
10541	-8.7	-14.316
8370	-2.33	-3.180
6720	-5.22	-3.070
31347	-0.64	-0.723
6214	-8.61	-9.778

Table A.11: Solvation energy prediction results for SAMPL3 molecules using all features

PubChem ID	Experimental Value	Predicted Value
6324	1.83	2.310
7095	-2.7	-3.083
9216	-3.15	-4.577
37207	-3.52	-4.479
38252	-3.1	-4.479
34586	-3.56	-4.531
36613	-3.67	-4.532
38253	-4.05	-4.579
15625	-3.37	-4.631
35454	-3.81	-4.623
48889	-3.84	-4.630
38251	-3.71	-4.564
38254	-4.15	-4.675
6214	-0.64	-1.590
12418	-1.43	-2.416
6419	-1.23	-1.545
6337	-0.63	-0.608
11	-1.79	-1.115
6365	-0.84	-0.801
6574	-1.99	-2.376
6591	-2.37	-2.414
249266	-2.69	-3.299
36980	-2.46	-2.137
37247	-2.16	-2.194
27588	-2.28	-2.024
36401	-3.48	-2.217
63088	-3.61	-3.091
91662	-1.96	-2.293
38019	-3.04	-2.805
63079	-4.38	-3.006
37037	-4.4	-2.733
38306	-3.17	-2.719
39253	-4.61	-2.626
6278	-0.19	-1.679
16318	-2.98	-2.528
18636	-4.53	-1.355

Table A.12: Solvation energy prediction results for SAMPL4 molecules using all features

PubChem ID	Experimental Value	Predicted Value
8079	0.14	0.870
10740	-2.78	-3.063
8418	-3.95	-3.848
11903	-0.85	-0.888
11940	-3.78	-3.011
31275	-5.06	-3.679
9216	-3.15	-5.498
10722	-3.93	-4.864
8908	-2.26	-2.107
26447	-2.53	-3.283
16666	-3.2	-2.917
6251	-23.62	-25.954
6997	-5.66	-6.143
8894	-3.12	-2.278
7583	-2.87	-3.669
7043	-5.33	-4.852
16441	-4.09	-3.572
6998	-4.68	-6.869
442474	-2.49	-2.027
8095	-5.31	-4.956
22227	-3.75	-3.463
61362	-4.51	-3.635
94221	-4.44	-4.034
637566	-4.45	-4.185
643820	-4.78	-4.031
21210	-5.73	-4.109
460	-5.94	-5.257
7144	-5.8	-7.423
17739	-5.26	-6.888
7041	-6.96	-5.976
8082	-5.11	-5.291
2160	-7.43	-5.097
77918	-7.63	-9.472
4044	-6.78	-8.697
30209	-1.66	-2.491
96257	-4.8	-4.599
138747	-4.29	-3.502
31420	-11.85	-9.133
6710	-9.44	-9.773
8341	-11.53	-11.820
20528	-14.21	-11.259
8323	-9.53	-8.767

Table A.12 (cont'd)

3100	-9.34	-7.481
53479	-7.78	-8.435
53476	-8.68	-9.121
7258	-7.29	-7.877
93462	-6.44	-4.586

APPENDIX B

SUPPLEMENTARY MATERIALS FOR TOXICITY ENDPOINT PREDICTION

This part of supplementary materials contain MT-DNN results with selected descriptors for all four datasets. Table B.1 - B.4 correspond to results calculated with descriptors whose importance are higher than 2.5e-4, 5e-4, 7.5e-4 and 1e-3, respectively. Table B.5 - B.12 contain the results with the ESTDs proposed in Discussion section using different algorithms.

Table B.1: Performances of different descriptor groups with importance threshold 2.5e-4

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.784	0.051	0.971	0.685	0.459	1.000
LC50DM	0.760	0.145	0.943	0.850	0.550	1.000
IGC50	0.760	0.078	0.981	0.482	0.334	1.000
LD50	0.617	0.306	0.954	0.598	0.433	0.997

Table B.2: Performances of different descriptor groups with importance threshold 5e-4

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.764	0.062	0.962	0.719	0.470	1.000
LC50DM	0.742	0.152	0.951	0.877	0.552	1.000
IGC50	0.757	0.075	0.984	0.486	0.338	1.000
LD50	0.612	0.325	0.954	0.601	0.437	0.997

Table B.3: Performances of different descriptor groups with importance threshold 7.5e-4

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.772	0.064	0.958	0.708	0.468	1.000
LC50DM	0.729	0.157	0.945	0.899	0.565	1.000
IGC50	0.751	0.076	0.983	0.491	0.341	1.000
LD50	0.600	0.341	0.953	0.611	0.439	0.997

Table B.4: Performances of different descriptor groups with importance threshold 1e-3

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.764	0.063	0.963	0.717	0.467	1.000
LC50DM	0.677	0.204	0.945	0.981	0.611	1.000
IGC50	0.746	0.075	0.982	0.497	0.342	1.000
LD50	0.607	0.325	0.955	0.605	0.437	0.997

Table B.5: Performances of RFon different datasets using ESTDs only proposed in Discussion section

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.685	0.324	0.947	0.829	0.608	1.000
LC50DM	0.418	1.749	0.960	1.327	1.023	1.000
IGC50	0.649	0.407	0.967	0.584	0.414	1.000
LD50	0.585	0.936	0.947	0.629	0.471	0.999

Table B.6: Performances of RF on different datasets using ESTDs proposed in Discussion section along with physical descriptors

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.732	0.314	0.948	0.776	0.561	1.000
LC50DM	0.449	1.517	0.957	1.289	0.973	1.000
IGC50	0.737	0.245	0.979	0.510	0.370	1.000
LD50	0.615	0.766	0.948	0.607	0.455	0.997

Table B.7: Performances of GBDT on different datasets using ESTDs only proposed in Discussion section

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.687	0.119	0.959	0.829	0.583	1.000
LC50DM	0.490	0.489	0.969	1.261	0.917	1.000
IGC50	0.737	0.072	0.987	0.508	0.348	1.000
LD50	0.609	0.431	0.958	0.604	0.449	0.999

Table B.8: Performances of GBDT on different datasets using ESTDs proposed in Discussion section along with physical descriptors

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.764	0.103	0.960	0.715	0.485	1.000
LC50DM	0.498	0.563	0.957	1.236	0.913	1.000
IGC50	0.791	0.053	0.994	0.451	0.313	1.000
LD50	0.635	0.328	0.959	0.583	0.438	0.997

Table B.9: Performances of MT-DNN on different datasets using ESTDs only proposed in Discussion section

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.738	0.012	1.004	0.763	0.497	1.000
LC50DM	0.723	0.001	0.993	0.907	0.629	1.000
IGC50	0.736	0.001	1.008	0.506	0.365	1.000
LD50	0.611	0.000	0.998	0.602	0.442	0.999

Table B.10: Performances of MT-DNN on different datasets using ESTDs proposed in Discussion section along with physical descriptors

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.771	0.012	1.017	0.716	0.467	1.000
LC50DM	0.725	0.000	0.999	0.903	0.597	1.000
IGC50	0.768	0.000	1.002	0.473	0.335	1.000
LD50	0.632	0.000	1.000	0.585	0.427	0.997

Table B.11: Performances of Consensus (MT-DNN and GBDT) on different datasets using ESTDs only proposed in Discussion section

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.745	0.084	0.962	0.745	0.496	1.000
LC50DM	0.653	0.297	0.968	1.017	0.736	1.000
IGC50	0.765	0.105	0.979	0.476	0.336	1.000
LD50	0.635	0.401	0.956	0.584	0.431	0.999

Table B.12: Performances of Consensus (MT-DNN and GBDT) different datasets using ESTDs proposed in Discussion section along with physical descriptors

Dataset	R^2	$\frac{R^2-R_0^2}{R^2}$	k	RMSE	MAE	Coverage
LC50	0.792	0.074	0.958	0.674	0.444	1.000
LC50DM	0.674	0.288	0.960	0.985	0.711	1.000
IGC50	0.802	0.068	0.987	0.437	0.304	1.000
LD50	0.657	0.321	0.957	0.565	0.420	0.997

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Junmei Wang, Wei Wang, Shuanghong Huo, Matthew Lee, and Peter A Kollman. Solvation model based on weighted solvent accessible surface area. *The Journal of Physical Chemistry B*, 105(21):5055–5067, 2001.

[2] Junmei Wang, George Krudy, Tingjun Hou, Wei Zhang, George Holland, and Xiaojie Xu. Development of reliable aqueous solubility models and their application in druglike analysis. *Journal of chemical information and modeling*, 47(4):1395–1404, 2007.

[3] TJ Hou, Ke Xia, Wei Zhang, and XJ Xu. Adme evaluation in drug discovery. 4. prediction of aqueous solubility based on atom contribution approach. *Journal of chemical information and computer sciences*, 44(1):266–275, 2004.

[4] Bao Wang, Zhixiong Zhao, and G. W. Wei. Automatic parametrization of non-polar implicit solvent models for the blind prediction of solvation free energies. *Journal of Chemical Physics*, 145:124110, 2016.

[5] Tiejun Cheng, Yuan Zhao, Xun Li, Fu Lin, Yong Xu, Xinglong Zhang, Yan Li, Renxiao Wang, and Luhua Lai. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *Journal of chemical information and modeling*, 47(6):2140–2148, 2007.

[6] Raimund Mannhold, Gennadiy I Poda, Claude Ostermann, and Igor V Tetko. Calculation of molecular lipophilicity: state-of-the-art and comparison of log P methods on more than 96,000 compounds. *Journal of pharmaceutical sciences*, 98(3):861–893, 2009.

[7] Joey W. Storer, David J. Giesen, Gregory D. Hawkins, Gillian C. Lynch, Christopher J. Cramer, Donald G. Truhlar, and Daniel A. Liotard. Solvation modeling in aqueous and nonaqueous solvent, new techniques and a reexamination of the claisen rearrangement. In C. J. Cramer and D. G. Truhlar, editors, *Structure, Energetics, and Reactivity in Aqueous Solution: Characterization of Chemical and Biological Systems*, 568, pages 24–49. American Chemical Society, 1994.

[8] R. Daudel. Quantum theory of chemical reactivity. In *Quantum Theory of Chemical Reactivity*, 1973.

[9] C. Reichardt. Solvents and solvent effects in organic chemistry. In *Solvents and Solvent Effects in Organic Chemistry*. Wiley-VCH: New York, 1990.

[10] Mikhail Borisover, Minolen Reddy, and Ellen R. Graber. Solvation effect on organic compound interactions in soil organic matter. *Environ. Sci. Technol.*, 35(12):2518âĂŞ2524, 2001.

[11] M.M. Kreevoy and D.G. Truhlar. In investigation of rates and mechanisms of reactions, part i. In C.F. Bernasconi, editor, *In Investigation of Rates and Mechanisms of Reactions, Part I*, page 13. Wiley: New York, 1986.

[12] M. E. Davis and J. A. McCammon. Electrostatics in biomolecular structure and dynamics. *Chemical Reviews*, 94:509–21, 1990.

[13] A. Warshel and A. Papazyan. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Current Opinion in Structural Biology*, 8(2):211–217, 1998.

[14] Aleksandr V. Marenich, Christopher J. Cramer, , and Donald G. Truhlar. Performance of SM6 SM8 and SMD on the SAMPL1 test set for the prediction of small-molecule solvation free energies. *Journal of Physical Chemistry B*, 113:4538âĂŞ4543, 2009.

[15] Silvia A. Martins, Sergio F. Sousa, Maria Joao Ramos, and Pedro A. Fernandes. Prediction of solvation free energies with thermodynamic integration using the general amber force field. *Journal of Chemical Theory and Computation*, 10:3570–3577, 2014.

[16] Gerhard König, Frank C Pickard, Ye Mei, and Bernard R Brooks. Predicting hydration free energies with a hybrid qm/mm approach: an evaluation of implicit and explicit solvation models in sampl4. *Journal of computer-aided molecular design*, 28(3):245–257, 2014.

[17] B. Jayaram, D. Sprous, and D. L. Beveridge. Solvation free energy of biomacromolecules: Parameters for a modified generalized Born model consistent with the AMBER force field. *Journal of Physical Chemistry B*, 102(47):9571–9576, 1998.

[18] J. A. Grant, B. T. Pickup, M. T. Sykes, C. A. Kitchen, and A. Nicholls. The Gaussian Generalized Born model: application to small molecules. *Physical Chemistry Chemical Physics*, 9:4913–22, 2007.

[19] H. Gohlke and D. A. Case. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex ras-raf. *Journal of Computational Chemistry*, 25(2):238–250, 2004.

[20] M. Feig, W. Im, and C. L. Brooks III. Implicit solvation based on generalized Born theory in different dielectric environments. *Journal of Chemical Physics*, 120(2):903–911, 2004.

[21] B. N. Dominy and C. L. Brooks, III. Development of a generalized Born model parameterization for proteins and nucleic acids. *Journal of Physical Chemistry B*, 103(18):3765–3773, 1999.

[22] D. Bashford and D. A. Case. Generalized Born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51:129–152, 2000.

[23] Jian J. Tan, Wei Z. Chen, and Cun X. Wang. Investigating interactions between HIV-1 gp41 and inhibitors by molecular dynamics simulation and MM-PBSA/GBSA calculations. *Journal of Molecular Structure: Theochem.*, 766(2-3):77–82, 2006.

[24] C. J. Cramer and D. G. Truhlar. Implicit solvation models: equilibria, structure, spectra, and dynamics. *Chemical Reviews*, 99(8):2161–2200, 1999.

[25] Christopher J. Cramer and Donald G. Truhlar. A universal approach to solvation modeling. *Accounts of chemical research*, 41:760–768, 2008.

[26] Insook Park, Yun Hee Jang, Sungu Hwang, and Doo Soo Chung. Poisson-boltzmann continuum solvation models for nonaqueous solvents i. 1-octanol. *Chemistry Letters*, 32:4, 2003.

[27] K. A. Sharp and B. Honig. Calculating total electrostatic energies with the nonlinear Poisson-Boltzmann equation. *Journal of Physical Chemistry*, 94:7684–7692, 1990.

[28] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules - theory and applications. *Annual Review of Biophysics and Biophysical Chemistry*, 19:301–332, 1990.

[29] Lin Li, Chuan Li, Subhra Sarkar, Jie Zhang, Shawn Witham, Zhe Zhang, Lin Wang, Nicholas Smith, Marharyta Petukh, and Emil Alexov. Delphi: a comprehensive suite for delphi software and associated resources. *BMC biophysics*, 5(1):9, 2012.

[30] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–9, 1995.

[31] M. K. Gilson, M. E. Davis, B. A. Luty, and J. A. McCammon. Computation of electrostatic forces on solvated molecules using the Poisson-Boltzmann equation. *Journal of Physical Chemistry*, 97(14):3591–3600, 1993.

[32] Lin Li, Chuan Li, and Emil Alexov. On the modeling of polar component of solvation energy using smooth gaussian-based dielectric function. *Journal of Theoretical and Computational Chemistry*, 13(03):1440002, 2014.

[33] Lin Li, Chuan Li, Zhe Zhang, and Emil Alexov. On the dielectric âĂIJconstantâĂİ of proteins: smooth dielectric function for macromolecular modeling and its implementation in delphi. *Journal of chemical theory and computation*, 9(4):2126–2136, 2013.

[34] K. Lum, D. Chandler, and J. D. Weeks. Hydrophobicity at small and large length scales. *Journal of Physical Chemistry B*, 103(22):4570–7, 1999.

[35] David M Huang and David Chandler. Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding. *Proceedings of the National Academy of Sciences*, 97(15):8324–8327, 2000.

[36] E. Gallicchio, L. Y. Zhang, and R. M. Levy. The SGB/NP hydration free energy model based on the surface generalized Born solvent reaction field and novel nonpolar hydration free energy estimators. *Journal of Computational Chemistry*, 23(5):517–29, 2002.

[37] E. Gallicchio and R. M. Levy. AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *Journal of Computational Chemistry*, 25(4):479–499, 2004.

[38] Niharendu Choudhury and B Montgomery Pettitt. On the mechanism of hydrophobic association of nanoscopic solutes. *Journal of the American Chemical Society*, 127(10):3556–3567, 2005.

[39] J. A. Wagoner and N. A. Baker. Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Proceedings of the National Academy of Sciences of the United States of America*, 103(22):8331–6, 2006.

[40] J. Dzubiella, J. M. J. Swanson, and J. A. McCammon. Coupling hydrophobicity, dispersion, and electrostatics in continuum solvent models. *Physical Review Letters*, 96:087802, 2006.

[41] G. W. Wei. Differential geometry based multiscale models. *Bulletin of Mathematical Biology*, 72:1562 – 1622, 2010.

[42] Z. Chen, N. A. Baker, and G. W. Wei. Differential geometry based solvation models I: Eulerian formulation. *J. Comput. Phys.*, 229:8231–8258, 2010.

[43] Z. Chen and G. W. Wei. Differential geometry based solvation models III: Quantum formulation. *J. Chem. Phys.*, 135:194108, 2011.

[44] Z. Chen, N. A. Baker, and G. W. Wei. Differential geometry based solvation models II: Lagrangian formulation. *J. Math. Biol.*, 63:1139– 1200, 2011.

[45] B. Wang and G. W. Wei. Parameter optimization in differential geometry based solvation models. *Journal Chemical Physics*, 143:134119, 2015.

[46] A Leo, DH Hoekman, and C Hansch. Hydrophobic, electronic, and steric constants, 1995.

[47] Han Van De Waterbeemd and Eric Gifford. ADMET in silico modelling: towards prediction paradise? *Nature reviews Drug discovery*, 2(3):192–204, 2003.

[48] Corwin Hansch and Toshio Fujita. p-σ-π analysis. a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86(8):1616–1626, 1964.

[49] Toshio Fujita, Junkichi Iwasa, and Corwin Hansch. A new substituent constant, π, derived from partition coefficients. *Journal of the American Chemical Society*, 86(23):5175–5180, 1964.

[50] Albert Leo, Corwin Hansch, and David Elkins. Partition coefficients and their uses. *Chemical reviews*, 71(6):525–616, 1971.

[51] Arup K Ghose and Gordon M Crippen. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. modeling dispersive and hydrophobic interactions. *Journal of chemical information and computer sciences*, 27(1):21–35, 1987.

[52] William M Meylan and Philip H Howard. Atom/fragment contribution method for estimating octanol–water partition coefficients. *Journal of pharmaceutical sciences*, 84(1):83–92, 1995.

[53] William M Meylan and Philip H Howard. Estimating log P with atom/fragments and water solubility with log P. *Perspectives in drug discovery and design*, 19(1):67–84, 2000.

[54] Albert J Leo. Calculating log poct from structures. *Chemical Reviews*, 93(4):1281–1306, 1993.

[55] JT Chou and Peter C Jurs. Computer-assisted computation of partition coefficients from molecular structures using fragment constants. *Journal of Chemical Information and Computer Sciences*, 19(3):172–178, 1979.

[56] Alanas A Petrauskas and Eduard A Kolovanov. ACD/Log P method description. *Perspectives in drug discovery and design*, 19(1):99–116, 2000.

[57] Matthew J Walker. Training ACD/LogP with experimental data. *QSAR & Combinatorial Science*, 23(7):515–520, 2004.

[58] Hao Zhu, Aleksander Sedykh, Suman K Chakravarti, and Gilles Klopman. A new group contribution approach to the calculation of LogP. *Current Computer-Aided Drug Design*, 1(1):3–9, 2005.

[59] Aleksandr Y Sedykh and Gilles Klopman. A structural analogue approach to the prediction of the octanol-water partition coefficient. *Journal of chemical information and modeling*, 46(4):1598–1603, 2006.

[60] Igor V Tetko and Vsevolod Yu Tanchuk. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *Journal of chemical information and computer sciences*, 42(5):1136–1145, 2002.

[61] Igor V Tetko and Pierre Bruneau. Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *Journal of pharmaceutical sciences*, 93(12):3103–3110, 2004.

[62] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

[63] Li Di and Edward H Kerns. Biological assay challenges from compound solubility: strategies for bioassay optimization. *Drug discovery today*, 11(9):446–451, 2006.

[64] Bao-Gen DUAN, Yan LI, Jie LI, Tie-Jun CHENG, and Ren-Xiao WANG. An empirical additive model for aqueous solubility computation: success and limitations. *Acta Physico-Chimica Sinica*, 28(10):2249–2257, 2012.

[65] Junmei Wang, Tingjun Hou, and Xiaojie Xu. Aqueous solubility prediction based on weighted atom type counts and solvent accessible surface areas. *Journal of Chemical Information and Modeling*, 49(3):571–581, 2009.

[66] Samuel H Yalkowsky and Shri C Valvani. Solubility and partitioning i: solubility of nonelectrolytes in water. *Journal of pharmaceutical sciences*, 69(8):912–922, 1980.

[67] John C Dearden. In silico prediction of aqueous solubility. *Expert opinion on drug discovery*, 1(1):31–52, 2006.

[68] RM Dannenfelser, M Paric, M White, and Samuel H Yalkowsky. A compliation of some physico-chemical properties for chlorobenzenes. *Chemosphere*, 23(2):141–165, 1991.

[69] William L Jorgensen and Erin M Duffy. Prediction of drug solubility from structure. *Advanced drug delivery reviews*, 54(3):355–366, 2002.

[70] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

[71] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33:249–274, 2005.

[72] K. L. Xia and G. W. Wei. Persistent homology analysis of protein structure, flexibility and folding. *International Journal for Numerical Methods in Biomedical Engineerings*, 30:814–844, 2014.

[73] K. L. Xia, X. Feng, Y. Y. Tong, and G. W. Wei. Persistent homology for the quantitative prediction of fullerene stability. *Journal of Computational Chemsitry*, 36:408–422, 2015.

[74] K. L. Xia, Z. X. Zhao, and G. W. Wei. Multiresolution topological simplification. *Journal Computational Biology*, 22:1–5, 2015.

[75] K. L. Xia, Z. X. Zhao, and G. W. Wei. Multiresolution persistent homology for excessively large biomolecular datasets. *Journal of Chemical Physics*, 143:134103, 2015.

[76] B. Wang and G. W. Wei. Object-oriented persistent homology. *Journal of Computational Physics*, 305:276–299, 2016.

[77] Zixuan Cang, Lin Mu, Kedi Wu, Kris Opron, Kelin Xia, and Guo-Wei Wei. A topological approach to protein classification. *Molecular based Mathematical Biologys*, 3:140–162, 2015.

[78] Z. X. Cang and G. W. Wei. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics (arXiv preprint arXiv:1703.10966)*, Revised, 2017.

[79] Z. X. Cang and G. W. Wei. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction . *International Journal for Numerical Methods in Biomedical Engineering*, Accepted, 2017.

[80] Z. X. Cang and G. W. Wei. TopologyNet: Topology based deep convolutional neural networks for biomolecular property predictions . *Plos Computational Biology (arXiv preprint arXiv:1704.00063)*, Submitted, 2017.

[81] Omar Deeb and Mohammad Goodarzi. In silico quantitative structure toxicity relationship of chemical compounds: some case studies. *Current drug safety*, 7(4):289–297, 2012.

[82] Gregory W Kauffman and Peter C Jurs. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *Journal of chemical information and computer sciences*, 41(6):1553–1560, 2001.

[83] Subhash Ajmani, Kamalakar Jadhav, and Sudhir A Kulkarni. Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *Journal of chemical information and modeling*, 46(1):24–31, 2006.

[84] Hongzong Si, Tao Wang, Kejun Zhang, Yun-Bo Duan, Shuping Yuan, Aiping Fu, and Zhide Hu. Quantitative structure activity relationship model for predicting the depletion percentage of skin allergic chemical substances of glutathione. *Analytica chimica acta*, 591(2):255–264, 2007.

[85] Hongying Du, Jie Wang, Zhide Hu, Xiaojun Yao, and Xiaoyun Zhang. Prediction of fungicidal activities of rice blast disease based on least-squares support vector machines and project pursuit regression. *Journal of agricultural and food chemistry*, 56(22):10785–10792, 2008.

[86] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[87] Peixun Liu and Wei Long. Current mathematical methods used in qsar/qspr studies. *International journal of molecular sciences*, 10(5):1978–1998, 2009.

[88] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[89] Christian Widmer and Gunnar Rätsch. Multitask learning in computational biology. In *ICML Unsupervised and Transfer Learning*, pages 207–216, 2012.

[90] Qian Xu and Qiang Yang. A survey of transfer and multitask learning in bioinformatics. *Journal of Computing Science and Engineering*, 5(3):257–268, 2011.

[91] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[92] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[93] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, 2012.

[94] Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8599–8603. IEEE, 2013.

[95] Richard Socher, Yoshua Bengio, and Christopher D Manning. Deep learning for nlp (without magic). In *Tutorial Abstracts of ACL 2012*, pages 5–5. Association for Computational Linguistics, 2012.

[96] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[97] George E Dahl, Navdeep Jaitly, and Ruslan Salakhutdinov. Multi-task neural networks for qsar predictions. *arXiv preprint arXiv:1406.1231*, 2014.

[98] Junshui Ma, Robert P Sheridan, Andy Liaw, George E Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 55(2):263–274, 2015.

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

[100] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

[101] Z. X. Wu and G. W. Wei. Comparison of multi-task convolutional neural network (MT-CNN) and a few other methods for toxicity prediction . *Toxicological Sciences*, Submitted, 2017.

[102] Alessandro Lusci, Gianluca Pollastri, and Pierre Baldi. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling*, 53(7):1563–1575, 2013.

[103] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 454–463. IEEE, 2000.

[104] G. Carlsson, A. Zomorodian, A. Collins, and L. J. Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11(2):149–187, 2005.

[105] R. Ghrist. Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc.*, 45:61–75, 2008.

[106] William M Haynes. *CRC handbook of chemistry and physics*. CRC press, 2014.

[107] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[108] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[109] Pang-Ning Tan and Vipin Kumar. Michael steinbach: Introduction to data mining.

[110] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[111] Leo Breiman. Arcing the edge. Technical report, Technical Report 486, Statistics Department, University of California at Berkeley, 1997.

[112] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[113] L Mason, J Baxter, P Bartlett, and M Frean. Boosting algorithms as gradient descent in function space (technical report). *RSISE, Australian National University*, 1999.

[114] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000.

[115] François Chollet. Keras. https://github.com/fchollet/keras, 2015.

[116] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[117] Bao Wang, Chengzhang Wang, Kedi Wu, and Guo-Wei Wei. Breaking the polar-nonpolar division in solvation free energy prediction. *Journal of computational chemistry*, 39(4):217–233, 2018.

[118] S. Cabani, P. Gianni, V Mollica, and L Lepori. Group Contributions to the Thermodynamic Properties of Non-Ionic Organic Solutes in Dilute Aqueous Solution. *Journal of Solution Chemistry*, 10(8):563–595, 1981.

[119] Junmei Wang, Wei Wang, Shuanghong Huo, Matthew Les, and Peter A. Kollman. Solvation model based on weighted solvent ccessible surface area. *J. Phys. Chem. B*, 105:5055–5067, 2001.

[120] Jiabo Li, Tianhai Zhu, Gergory D. Hawkins, Paul Winget, Daniel A. Liotard, Christopher J. Cramer, and Donald G. Truhlar. Extension of the platform of applicability of the sm5.42r universal solvation model. *Theoretical Chemistry Accounts*, 103:9–63, 1999.

[121] Matthew T Geballe, A Geoffrey Skillman, Anthony Nicholls, J Peter Guthrie, and Peter J Taylor. The sampl2 blind prediction challenge: introduction and overview. *Journal of computer-aided molecular design*, 24(4):259–279, 2010.

[122] A Geoffrey Skillman. Sampl3: blinded prediction of host–guest binding affinities, hydration free energies, and trypsin inhibitors. *Journal of computer-aided molecular design*, pages 1–2, 2012.

[123] Hari S Muddana, C Daniel Varnado, Christopher W Bielawski, Adam R Urbach, Lyle Isaacs, Matthew T Geballe, and Michael K Gilson. Blind prediction of host–guest binding affinities: a new sampl3 challenge. *Journal of computer-aided molecular design*, 26(5):475–487, 2012.

[124] Hari S Muddana, Andrew T Fenley, David L Mobley, and Michael K Gilson. The sampl4 host–guest blind prediction challenge: an overview. *Journal of computer-aided molecular design*, 28(4):305–317, 2014.

[125] David L Mobley, Karisa L Wymer, Nathan M Lim, and J Peter Guthrie. Blind prediction of solvation free energies from the sampl4 challenge. *Journal of computer-aided molecular design*, 28(3):135–150, 2014.

[126] Corwin Hansch, Albert Leo, DH Hoekman, et al. *Exploring QSAR: fundamentals and applications in chemistry and biology*, volume 557. American Chemical Society Washington, DC, 1995.

[127] Alex Avdeef. *Absorption and drug development: solubility, permeability, and charge state*. John Wiley & Sons, 2012.

[128] Junmei Wang, George Krudy, Tingjun Hou, Wei Zhang, George Holland, and Xiaojie Xu. Development of reliable aqueous solubility models and their application in druglike analysis. *Journal of chemical information and modeling*, 47(4):1395–1404, 2007.

[129] John S Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005, 2004.

[130] Jarmo Huuskonen. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *Journal of Chemical Information and Computer Sciences*, 40(3):773–777, 2000.

[131] SH Yalkowsky and RM Dannelfelser. The arizona database of aqueous solubility. *Tuscon, AZ, USA*, 1990.

[132] P Howard and W Meylan. Physical/chemical property database (physprop), 1999.

[133] T. Martin. *User's Guide for T.E.S.T. (version 4.2) (Toxicity Estimation Software Tool): A Program to Estimate Toxicity from Molecular Structure*.

[134] Kevin S Akers, Glendon D Sinks, and T Wayne Schultz. Structure–toxicity relationships for selected halogenated aliphatic chemicals. *Environmental Toxicology and Pharmacology*, 7(1):33–39, 1999.

[135] Hao Zhu, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatica, Tomas Oberg, Phuong Dao, Artem Cherkasov, and Igor V Tetko. Combinatorial qsar modeling of chemical toxicants tested against tetrahymena pyriformis. *Journal of Chemical Information and Modeling*, 48(4):766–784, 2008.

[136] Dong-Sheng Cao, Qing-Song Xu, Qian-Nan Hu, and Yi-Zeng Liang. Chemopy: freely available python package for computational biology and chemoinformatics. *Bioinformatics*, page btt105, 2013.

[137] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

[138] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling*, 25(2):247–260, 2006.

[139] Beibei Liu, Bao Wang, Rundong Zhao, Yiying Tong, and Guo Wei Wei. ESES: software for Eulerian solvent excluded surface. *Journal of Computational Chemistry*, 38:446–466, 2017.

[140] Duan Chen, Zhan Chen, Changjun Chen, W. H. Geng, and G. W. Wei. MIBPB: A software package for electrostatic analysis. *J. Comput. Chem.*, 32:657 – 670, 2011.

[141] Alexander Golbraikh, Min Shen, Zhiyan Xiao, Yun-De Xiao, Kuo-Hsiung Lee, and Alexander Tropsha. Rational selection of training and test sets for the development of validated QSAR models. *Journal of computer-aided molecular design*, 17(2):241–253, 2003.

[142] D. A. Case, J. T. Berryman, R. M. Betz, D. S. Cerutti, T. E. Cheatham III, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K. M. Merz, G. Monard, P. Needham, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, R. Salomon-Ferrer, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R.M. Wolf, X. Wu, D. M. York, and P. A. Kollman. Amber 2015. *University of California, San Francisco*, 2015.

[143] Araz Jakalian, Bruce L. Bush, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method. *Journal of Computational Chemistry*, 21(2):132–146, 2000.

[144] J. Gasteiger and M. Marsili. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron*, 36:3219–3228, 1980.

[145] Anthony Nicholls, David L. Mobley, J. Peter Guthrie, John D. Chodera, Chridtopher I. Bayly, Matthew D. Cooper, and Vijay S. Pande. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.*, 51:769–799, 2008.

[146] Charles W. Kehoe, Christopher J. Fennell, and Ken A. Dill. Testing the semi-explicit assembly solvation model in the sampl3 community blind test. *J Comput Aided Mol Des*, 26:563–568, 2012.

[147] J. Peter Guthrie. A blind challenge for computational solvation free energies: Introduction and overview. *Journal of Physical Chemistry B*, 113:4501–4507, 2009.

[148] Pavel V. Klimovich and David L. Mobley. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *J. Comput Aided Mol Des.*, 24:307–316, 2010.

[149] Matthew T. Geballe and J. P. Guthrie. The SAMPL3 blind prediction challenge: transfer energy overview. *Journal of Computer-Aided Molecular Design*, 26:489 –496, 2012.

[150] David L. Mobley, Karisa L. Wymer, Nathan M. Lim, and J. Peter Guthrie. Blind prediction of solvation free energies from the sampl4 challenge. *J. Comput Aided Mol Des*, 28:135–150, 2014.

[151] Gilles Klopman, Shaomeng Wang, and Donald M Balthasar. Estimation of aqueous solubility of organic molecules by the group contribution approach. application to the study of biodegradation. *Journal of chemical information and computer sciences*, 32(5):474–482, 1992.

[152] Gilles Klopman and Hao Zhu. Estimation of the aqueous solubility of organic molecules by the group contribution approach. *Journal of chemical information and computer sciences*, 41(2):439–445, 2001.

[153] Neera Jain and Samuel H Yalkowsky. Estimation of the aqueous solubility i: Application to organic nonelectrolytes. *Journal of pharmaceutical sciences*, 90(2):234–252, 2001.

[154] Chloé-Agathe Azencott, Alexandre Ksikes, S Joshua Swamidass, Jonathan H Chen, Liva Ralaivola, and Pierre Baldi. One-to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties. *Journal of chemical information and modeling*, 47(3):965–974, 2007.

[155] Holger Fröhlich, Jörg K Wegner, and Andreas Zell. Towards optimal descriptor subset selection with support vector machines in classification and regression. *QSAR & Combinatorial Science*, 23(5):311–318, 2004.

[156] Hao Zhu, Todd M Martin, Lin Ye, Alexander Sedykh, Douglas M Young, and Alexander Tropsha. Quantitative structure- activity relationship modeling of rat acute toxicity by oral exposure. *Chemical research in toxicology*, 22(12):1913–1921, 2009.

[157] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.