Graph Signal Processing Approach to QSAR/QSPR Model Learning of Compounds

Xiaoying Song, Li Chai*, and Jingxin Zhang

***Abstract*—Quantitative relationship between the activity/property and the structure of compound is critical in chemical applications. To learn this quantitative relationship, hundreds of molecular descriptors have been designed to describe the structure, mainly based on the properties of vertices and edges of molecular graph. However, many descriptors degenerate to the same values for different compounds with the same molecular graph, resulting in model failure. In this paper, we design a multidimensional signal for each vertex of the molecular graph to derive new descriptors with higher discriminability. We treat the new and traditional descriptors as the signals on the descriptor graph learned from the descriptor data, and enhance descriptor dissimilarity using the Laplacian filter derived from the descriptor graph. Combining these with model learning techniques, we propose a graph signal processing based approach to obtain reliable new models for learning the quantitative relationship and predicting the properties of compounds. We also provide insights from chemistry for the boiling point model. Several experiments are presented to demonstrate the validity, effectiveness and advantages of the proposed approach.**

***Index Terms*—QSAR/QSPR model learning, compounds, graph signal processing (GSP), multidimensional signal**

I. Introduction

EXTENSIVE chemical and medical experiments have revealed that physicochemical properties of the compound are highly related to its molecular structure. Applying chemical theory and various mathematical analysis methods, the Quantitative Structure Activity/Property Relationship (QSAR/QSPR) model learning attempts to describe this relationship quantitatively [1, 2]. QSAR/QSPR model has become an extensively used tool in computer-aided drug design, toxicity and property prediction of chemicals and pharmaceuticals [3] as well as in different modeling problems in material sciences [4], analytical chemistry and pharmacodynamics profiling of new drug molecules [5].

In most cases, the molecular structure is represented as graph (called molecular graph) with vertices denoting atoms and edges describing chemical bonds between atoms. This graph representation allows for application of graph theoretic algorithms to assess statistical and/or topological properties of networks reconstructed from molecular structures. Based on these statistical and/or topological indices, one can estimate the chemical, biological, medical and pharmacological properties of compounds.

Over the past decade, considerable progress has been made in QSAR/QSPR model learning. It is now feasible to study and predict properties of new compounds from a set of training molecules with known activities/properties/toxicities. Learning QSAR/QSPR models requires three main steps: generating a training set of measured properties of known compounds, encoding the information of compound structures, and building a mathematical model to predict measured properties from the encoded structure [6].

QSAR/QSPR models are regression or classification models. QSAR/QSPR regression models relate a set of "predictor" variables (also called molecular descriptors) to the potency of the response variable, while QSAR/QSPR classification models relate the predictor variables to a categorical value of the response variable. The molecular descriptors consist of two main categories: experimental measurements, such as log P, molar refractivity, dipole moment and physicochemical properties in general, and theoretical molecular descriptors, which are derived from a symbolic representation of the molecule and can be further classified according to the different types of molecular representation. The response variable could be a biological activity or property of the compound. Various learning techniques have been applied to QSAR/QSPR model learning in recent years. Examples include partial least squares (PLS) [7], multiple linear regression (MLR) [3], support vector machine [8], random forest [9], neural networks [10] and so on.

Generally hundreds of molecular descriptors are required to learn a valid model [11, 12]. However, most descriptors are constructed based on topological properties of molecular graph, which may generate the same value for different compounds having the same molecular graph. Therefore, one cannot build useful QSAR/QSPR models to represent different properties of such compounds. To overcome this difficulty, new technical tools are required.

Graph signal processing (GSP) has recently emerged as a powerful new approach to analyzing and processing high-dimensional signals defined on irregular graph domains [13–16]. It treats the data at vertices as the signal on graph, and then analyzes and processes the data from signal processing perspective. The theory of GSP has been growing rapidly in recent years, with development in methods such as graph filtering [17] and graph neural network (GNN) learning [18]. These new theory and techniques have provided new tools for QSAR/QSPR model learning and have motivated this work.

We propose in this paper a new GSP based approach

*Corresponding author. Li Chai is with the Engineering Research Center of Metallurgical Automation and Measurement Technology, Wuhan University of Science and Technology, Wuhan, China. e-mail: chaili@wust.edu.cn.

Xiaoying Song is with the same University. e-mail: xiaoying811@wust.edu.cn.

Jingxin Zhang is with the School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, Australia. e-mail: jingxinzhang@swin.edu.au.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3032718, IEEE Transactions on Pattern Analysis and Machine Intelligence

to QSAR/QSPR model learning. This approach stems from an intuitive idea that different compounds with the same molecular graph can be represented by different signals on the vertices of the graph; such signals are the graph signals in the GSP theory, and hence can be processed using various GSP tools to find new QSAR/QSPR models for different compounds with the same or similar molecular structures.

Based on this idea, we propose the following GSP based approach to QSAR/QSPR model learning: 1) Design new multidimensional (MD) signal at each vertex of molecular graph; 2) Analyze the MD signals (MDS) on all vertices of the graph to derive new MDS descriptors with higher discriminability for QSAR/QSPR modeling; 3) Combine the new and traditional descriptors to construct new QSAR/QSPR models with higher performance and reliability; 4) Treat the descriptors as the signals on a new descriptor graph and construct the graph from descriptor data; 5) Derive a Laplacian graph filter from the descriptor graph and use Laplacian filtering of descriptors to enhance their dissimilarity; 6) Use the dissimilarity enhanced input variables in new QSAR/QSPR model to further enhance model performance and reliability; 7) Use Least Angle Regression with Lasso modification to learn the optimal sparse parameters of the model.

Existing works, in descriptor design and in modeling, are based only on the molecular graph. Our GSP based approach is very different from those of the previous works. To the best of our knowledge, this approach has never been used in QSAR/QSPR model learning of compounds. The contributions of this paper are:

i) A general GSP based approach to QSAR/QSPR model learning of compounds as summarized above.

ii) Application results, showing the advantages of the proposed approach, obtained from two benchmark datasets.

iii) A biological activity model for phenetylamines, the benchmark dataset including 22 different compounds with the same molecular graph. A simple relationship between the MDS and the biological activity and the state-of-the-art fitting results.

iv) A boiling point model for polyaromatic hydrocarbons, the benchmark dataset including 82 different compounds with similar molecular graphs, showing the boiling point of compounds can be estimated by the molecular graph spectral information and the molecular mass.

The paper is organized as follows. Section II provides some background about GSP and QSAR/QSPR model learning. Section III presents the proposed approach and its methods. Sections IV and V present the applications of the proposed approach in two families of compounds. Results obtained from the applications are given in Section VI. Discussions and conclusions on the results of the paper are given in Sections VII and VIII, respectively.

II. Preliminaries

A. Notations and Definitions

A graph is denoted by $G = (V, E, W)$ with vertex set $V = \{v_1, v_2, \cdots, v_N\}$, edge set E and weight matrix W. The number of vertices is $N = |V|$ and the number of edges is $m = |E|$. The *(i,j)*-th element of W is the weight of the edge $\varepsilon_{ij} \in E$. In molecular graph, it is nonzero if there is a chemical bond between vertices v_i and v_j, otherwise it is zero. Different weights are assigned to different chemical bonds, with values of 1, 2, 3 and 1.5 of w_{ij} representing single, double, triple and aromatic bonds, respectively. An illustrate example is shown in Fig. 1. In graph representation, hydrogen atoms are implicit and omitted, which has the advantage of leading to more compact graph architecture and faster training speed.

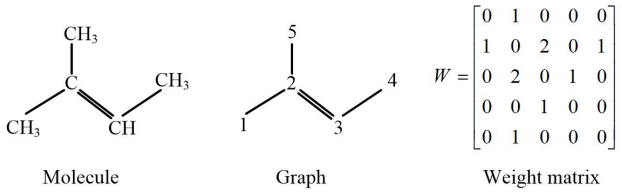


Fig. 1: The molecular graph and weight matrix of 2-methyl-2-butene.

Define diagonal matrix $D := diag(d_i)$, where $d_i = \sum_j w_{ij}$ is the degree of the vertex v_i. The Laplacian matrix of G is defined as $L = D - W$. L is symmetric and positive semi-definite, and its eigenvalues $\{0 \leq \lambda_1 \leq \lambda_2 \cdots \leq \lambda_N\}$ are defined as the spectra of the graph. Note that the definition of Laplacian matrix is not unique. A different definition will be introduced in Section III-B. A graph signal is a mapping from the vertex set to the real number field, i.e., $x : V \to R$. An operator $H \in R^{N \times N}$ that gives output $H\mathbf{x}$ for a graph signal $\mathbf{x} \in R^{N \times 1}$ (or $\mathbf{x}^T H^T$ for $\mathbf{x}^T \in R^{1 \times N}$) represents a graph filter. In the canonical form, W has a low pass nature and L has a high pass nature in the graph spectrum domain [19].

All graphs considered in this work are assumed to be simple, undirected and weighted, that is, no loop and multiple edge.

B. Molecular Descriptors

Molecular descriptor plays an important role in the development and interpretation of QSAR/QSPR model. It is defined as a positive valued real function $\Psi : G \to R^+$ that maps the molecular graph to a positive real number. There are many degree-based, distance-based and spectrum-based descriptors, which have been widely used in the QSAR/QSPR model to test properties of compounds. In recent years, several research groups have made outstanding contributions to different kinds of descriptors of special molecular structures [20–23]. Lucic et al. [20] showed that the Randic connectivity index and the sum-connectivity index are closely related molecular descriptors. Hayat et al. [21, 22] performed a comparative testing to measure the efficiency of all the well-known valency-based molecular descriptors and proposed an efficient computer-based computational technique to compute descriptors. Vukicevic and Graovac [23] analyzed the first and the second Zagreb eccentricity descriptors. Several well known degree-based and distance-based descriptors are briefly described below [24].

The general Randic index and the general sum connectivity index are stated as

$$R_d(G) = \sum_{\varepsilon_{ij}} (d_i d_j)^d, \tag{1}$$

where d is a real number, d_i and d_j are degrees of vertices v_i and v_j, respectively.

The general sum connectivity index is stated as

$$\chi_d(G) = \sum_{\varepsilon_{ij}} (d_i + d_j)^d. \tag{2}$$

The distance-based version of the atom-bond connectivity index is defined as

$$ABC_2(G) = \sum_{\varepsilon_{ij}} \sqrt{\frac{n_i + n_j - 2}{n_i n_j}}, \tag{3}$$

where n_i is the number of vertices whose distances to vertex v_i are smaller than those to the other vertex v_j of the edge ε_{ij}, and n_j is defined analogously.

Spectrum-based descriptors are also used to describe the molecular structures. Gutman and Zhou [25] defined the Laplacian energy of the graph G as

$$LE(G) = \sum_{i=1}^{N} |\lambda_i - \frac{2m}{N}|, \tag{4}$$

where λ_i is the ith eigenvalue of L.

Dehmer et al. [26] defined some families of eigenvalue-based descriptors, one of them is stated as

$$S_{L,d}(G) = |\lambda_1|^{\frac{1}{d}} + |\lambda_2|^{\frac{1}{d}} + \cdots + |\lambda_N|^{\frac{1}{d}}, \tag{5}$$

where d is a real number.

The above-mentioned descriptors are global measures that describe the overall network topological information. There are some local measures for the molecular graph.

Brandes and Erlebach [27] proposed the stress centrality and the betweenness centrality for each individual vertex of the graph G, which are stated respectively as

$$SC_i = \sum_{jk} n(j,i,k), \tag{6}$$

$$BC_i = \sum_{jk} \frac{n(j,i,k)}{n(j,k)}, \tag{7}$$

where $n(j,i,k)$ is the number of shortest paths between vertices v_j and v_k that pass through vertex v_i. $n(j,k)$ is the total number of shortest paths between vertices v_j and v_k. Stress centrality is used to describe the number of weighted paths that pass through the vertex v_i. Betweenness centrality is used to describe the ratio of paths that pass through the vertex v_i.

Costa et al. [28] introduced the vulnerability efficiency for each individual vertex stated as

$$VE_i = \frac{GE - GE_i}{GE}, \tag{8}$$

where $GE = \frac{1}{N(N-1)} \sum_{\varepsilon_{ij}} d_{ij}$ is the global efficiency, GE_i is the global efficiency after the removal of the vertex v_i and all its associated links, and d_{ij} is the shortest weighted path between vertices v_i and v_j. A smaller GE means that all vertices of the graph are closer. Vulnerability efficiency computes the average efficiency of the graph and measures the importance of vertex v_i on the system performance if vertex v_i and all its associated links are removed.

C. QSAR/QSPR Model

The QSAR/QSPR models in the literature are generally in the form [29, 30]

$$q = \sum_k f_k(\Psi(G)) + c, \tag{9}$$

where $\Psi(G)$ are input variables chosen from the descriptor functions of a compound discussed above, which are only related to the structural information of the graph G, $f_k(\Psi(G))$ are functions of $\Psi(G)$, and c is a constant. The scalar output q of the model is usually biological activity or physicochemical property of a compound. All compounds in the same chemical family are subject to this model, and the input and output variables of the model differ for different compounds.

D. Performance Indices

The performance of QSAR/QSPR model is assessed using two groups of statistical indices, goodness-of-fit metrics and goodness-of-prediction metrics [31]. Goodness-of-fit metrics measure the fitting ability and are used to measure the degree to which the model is able to explain the variance contained in the training set. The three most important metrics are the root mean square error ($RMSE$), the average absolute error (AAE) and the coefficient of determination (R^2). $RMSE$ and AAE are two frequently used metrics of the errors between values predicted by a model and the values observed, which are defined by

$$RMSE = \sqrt{\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} (y_{i\_train} - \hat{y}_i)^2}, \tag{10}$$

$$AAE = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} |y_{i\_train} - \hat{y}_i|, \tag{11}$$

where n_{tr} is the number of compounds in the training set, y_{i_train} and $\hat{y}_i$ are the target value (experimentally observed) and the corresponding predicted value in the training set, respectively.

R^2 is a statistic metric which is independent of the response scale, contrary to $RMSE$ and AAE. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. The most general definition of R^2 is

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{tr}} (y_{i\_train} - \hat{y}_i)^2}{\sum_{i=1}^{n_{tr}} (y_{i\_train} - \bar{y})^2}, \tag{12}$$

where $\bar{y}$ is the average observed value over the entire training set.

Goodness-of-prediction metrics measure the generalization ability of a model. In most cases, model validation by internal cross-validation technique is not enough and validation by an

external test set has been suggested as an effective way of evaluating the model predictive ability. The most important metrics are the root mean square error in prediction ($RMSEP$) and the predictive squared correlation coefficient (R_P^2). $RMSEP$ denotes $RMSE$ on the test set and we use the following R_P^2 defined by Todeschini et al. [32]

$$R_P^2 = 1 - \frac{\sum_{i=1}^{n_{test}}(y_{i\_test} - \hat{y}_i)^2/n_{test}}{\sum_{i=1}^{n_{tr}}(y_{i\_train} - \bar{y})^2/n_{tr}}. \quad (13)$$

As known from [31], the larger the R^2 and the smaller the $RMSE$ and AAE are, the better the model performance is.

III. Methods

A. Problem Statement

The degeneration of traditional molecular descriptors results in the same values for different compounds with the same molecular graph. This can lead to model failure, especially in the following two cases. In the first case, the compounds have the same graph structure and the same weight matrix. The only difference is that their atoms at some vertices are different. If the atomic information is not taken into account, not only the properties of vertices and edges, but also the spectra of the graphs are the same. In the second case, the compounds have the same graph structure and the same heavy atoms, but the types of chemical bonds are different, which means the weight matrices are different. The properties of vertices are the same, if the atomic information is not considered. Illustrative examples for these two cases are shown in Fig. 2. Therefore, descriptor degeneration and reliability and performance of QSAR/QSPR model are the two key problems to be addressed in this paper.

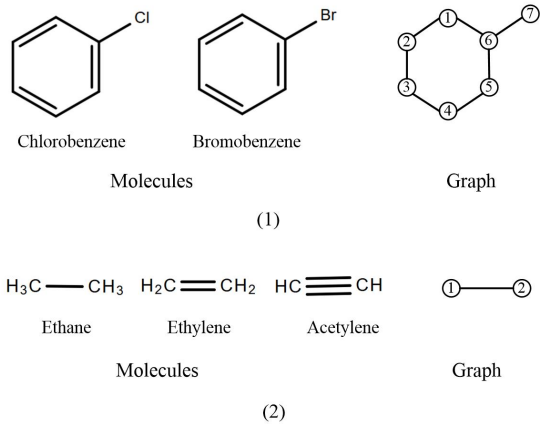


Fig. 2: Two cases that molecular descriptors degenerate.

To address these issues, we propose a novel GSP based approach to QSAR/QSPR model learning. To present this approach, we introduce the following notations.

For a molecular graph $G(V, E, W)$ with N vertices, a real valued M_1-dimensional signal on the vertex v_i is denoted as $\mathbf{s}_i = [s_{i1}, s_{i2}, \cdots, s_{iM_1}]^{\mathrm{T}}$, and the signal matrix $S = [\mathbf{s}_1, \cdots, \mathbf{s}_i, \cdots, \mathbf{s}_N]^{\mathrm{T}} \in R^{N\times M_1}$ collects all vertex signals $\mathbf{s}_i$'s in its N rows.

B. Proposed Approach

In order to solve the above stated problems, we design MD signals on vertices and use them to derive new descriptors, called MDS descriptors. Combining the new MDS descriptors with the traditional descriptors, we propose a new modeling method for the quantitative relationship between the structure and the physicochemical/activity property of the compound. The proposed method includes six steps:

Step 1: *MD signal construction.* Design of MD signal is the first step of the proposed approach. Two basic requirements for the design include: i) be able to distinguish different compounds and ii) can reflect important information relevant to the physicochemical/activity property of the compound. To achieve these, we design the MD signal $\mathbf{s}_i = [s_{i1}, s_{i2}, \cdots, s_{iM_1}]^{\mathrm{T}}$ for each vertex v_i. Both the atomic information of the compound, such as charge, molar mass and chemical bond, and the local measures of the graph can be used to construct the signal s_{ik}, $k = 1, 2, \cdots, M_1$. High correlation of the designed signals with the compound properties improves the model performance.

Step 2: *Input variable construction.* Let $S_k \in R^N$, $k = 1, 2, \cdots, M_1$, be the kth column of S. Then we define new MDS descriptors $\zeta_k := \frac{1}{N}||S_k||_p$, with $\|S_k\|_p$ the p-norm of S_k and $p \geq 1$. Next, we choose M_2 traditional descriptors, such as degree and degree-based indices, Laplacian energy and eigenvalue-based indices, and denote them $\eta_k, k = 1, 2, \cdots, M_2$.

To distinguish these two sets of descriptors, we denote $\Phi(S)$ the set $\{\zeta_1, \cdots, \zeta_k, \cdots, \zeta_{M_1}\}$ to indicate its sole dependence on the MD signal S, and denote $\Psi(G)$ the set $\{\eta_1, \cdots, \eta_k, \cdots, \eta_{M_2}\}$ to indicate its sole dependence on the structural information of the graph G.

Using these descriptors, we define the input variable vector

$$\mathbf{x} = [x_1, \cdots, x_k, \cdots, x_M] := [\zeta_1, \cdots, \zeta_{M_1}, \eta_1, \cdots, \eta_{M_2}] \quad (14)$$

with $M = M_1 + M_2$. Let $\mathbf{x}_i$, $i = 1, 2, \cdots, B$, be the samples of $\mathbf{x}$ from B compounds. We construct the training data matrix

$$X = [\mathbf{x}_1^T, \cdots, \mathbf{x}_i^T, \cdots, \mathbf{x}_B^T]^T \in R^{B\times M}. \quad (15)$$

Step 3: *Input variable GSP*. Inspired by the outstanding work of [33], we treat the variables in $\mathbf{x}$ as the graph signals on a descriptor graph learned from their sample data X, and filter X with the graph Laplacian filter derived from the descriptor graph, using the following procedure.

Let $X_k \in R^B$, $k \in \mathcal{M} = \{1, 2, \cdots, M\}$, be the kth column of X. We first construct a distance matrix E with elements $E_{i,j} = \|X_i - X_j\|_2, i, j \in \mathcal{M}$, and convert E to an asymmetric affinity matrix A using an adaptive Gaussian kernel function $A_{i,j} = exp(-E_{i,j}/\sigma_j)^2$, where σ_j is the $(knn+1)$th smallest $E_{i,j}$ in the jth column of E. Then we symmetrize A to $\tilde{A} = A + A^{\mathrm{T}}$ and column normalize $\tilde{A}$ to a Markov transition matrix K, with $K_{i,j} = \tilde{A}_{i,j}/\sum_i \tilde{A}_{i,j}$ and each column summing to 1. We further construct a normalized Laplacian matrix of the descriptor graph using K and an $M \times M$ identity matrix I_M

$$L_d = I_M - K. \quad (16)$$

The adjacency matrix $\tilde{A}$ defines a graph of $\mathbf{x}$ learned from its sample data X. The columns of Markov transition matrix

K represent the probability distribution of transitioning from a particular descriptor to every other descriptor in one step along the column direction of the graph [33]. The Laplacian matrix L is a derivative operator and its lth power L_d^l defines a high pass graph spectrum filter on the graph, with the order $l \geq 1$ [19]. Roughly speaking, the operation on $\mathbf{x}$ with L_d^l

$$\mathbf{x}L_d^l := \tilde{\mathbf{x}} = [\tilde{x}_1, \cdots, \tilde{x}_k, \cdots, \tilde{x}_M] \tag{17}$$

yields $\tilde{x}_k$'s with enhanced dissimilarity on the graph. A demonstrative example is given in Fig 5 in Section IV-B. In principle, the higher the order l, the stronger the enhancement. In practice, however, the order l should be chosen such that the $|\tilde{x}_k|$'s are not annihilated by filtering. Based on this fact, we filter the training data X by L_d^l to obtain the dissimilarity enhanced training data

$$XL_d^l := \tilde{X} = [\tilde{X}_1, \cdots, \tilde{X}_k, \cdots, \tilde{X}_M]^T. \tag{18}$$

Step 4: *Modeling*. Using the $\Phi(S)$ and $\Psi(G)$ variables defined in step 2, we propose a new QSAR/QSPR model

$$q = \sum_k f_k(\Phi(S), \Psi(G)). \tag{19}$$

Different to the traditional QSAR/QSPR model (9), the new QSAR/QSPR model (19) uses input variables from both $\Phi(S)$ and $\Psi(G)$. Thus it uses not only the structural information of graph G carried by $\Psi(G)$, but also the information from MD signal S carried by $\Phi(S)$. We decompose the property of a molecule as a sum of local contributions, and represent each local atomic environment by MDS descriptors derived from MD signal S that are inherently invariant. This may result in better models and is one of the main contributions of this work.

In this work, we focus on two special cases of the QSAR/QSPR model (19).

Model 1: $f_k(\Phi(S), \Psi(G)) = \theta_k \zeta_k$, $k = 1, 2, \cdots, M_1$, and $f_{k+M_1}(\Phi(S), \Psi(G)) = \theta_{k+M_1} \eta_k$, $k = 1, 2, \cdots, M_2$, with ζ_k and η_k the MDS and traditional descriptors defined in Step 2 and (14) and θ_k their coefficients, that is,

$$q = \sum_{k=1}^{M_1} \theta_k \zeta_k + \sum_{k=1}^{M_2} \theta_{M_1+k} \eta_k + c = [\mathbf{x} \ \ 1]\theta. \tag{20}$$

Model 2: $f_k(\Phi(S), \Phi(S)) = \theta_k \tilde{x}_k$, $k = 1, 2, \cdots, M_1$, and $f_{k+M_1}(\Phi(S), \Psi(G)) = \theta_{k+M_1} \tilde{x}_{k+M_1}$, $k = 1, 2, \cdots, M_2$, with $\tilde{x}_k$ defined in (17) and θ_k their coefficients, that is,

$$q = \sum_{k=1}^{M_1} \theta_k \tilde{x}_k + \sum_{k=1}^{M_2} \beta_k \tilde{x}_{k+M_1} + c = [\tilde{\mathbf{x}} \ \ 1]\theta. \tag{21}$$

In (20) and (21), $M_1 + M_2 = M$ and

$$\theta := [\theta_1, \cdots, \theta_{M_1}, \theta_{M_1+1}, \cdots, \theta_M, c]^T \tag{22}$$

is the coefficient vector of the models, and $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are the regression variables defined in (14) and (17), respectively.

Step 5: *Sparse coefficient learning*. Define $Q := [q_1, \cdots, q_i, \cdots, q_B]^T$ consisting of the output variables of B compounds. Let X in the form (15) and $\tilde{X}$ in the form (18) be the corresponding data matrix and filtered data matrix of these compounds. Define $Y := [X \ \ \mathbf{1}]$ for (20) and $Y := [\tilde{X} \ \ \mathbf{1}]$ for (21), with $\mathbf{1} \in R^B$ an all 1 vector. Then we can write the following matrix form regression equation for learning the models (20) and (21)

$$Q = Y\theta. \tag{23}$$

When many descriptors are used as input variables in the model, that is, when M is large, some of them may be interdependent on each other or may have mere correlation with the output variable q. To single out and eliminate these variables and find the optimal coefficients for the reserved variables, we use the Least Angle Regression with Lasso modification [34] to find an optimal sparse θ with the least-squares of prediction error and minimum number of nonzero coefficients. The computation procedure is summarized in Algorithm 1, where $\bar{Q}$ denotes the model prediction of Q and Y_j denotes the jth column of Y in (23).

Step 6: *Performance evaluation*. The performance of our model is assessed using the metrics defined in Section II-D. $RMSE$, AAE and R^2 are used to measure the fitting ability of the model; and $RMSEP$ and R_P^2 are used to measure the generalization ability of the model. Low errors between predicted and experimental values indicate a good QSAR/QSPR model, whereas high errors indicate a poor one.

The pseudo-code of the proposed method is given in Algorithm 2.

Algorithm 1 Least Angle Regression with Lasso [34]

1: Standardize the variables to have mean zero and unit norm. Start with the residual $r = Q - \bar{Q}$ and $\theta_1, \theta_2, \cdots, \theta_M = 0$.
2: Find the variable Y_j most correlated with r.
3: Move θ_j from 0 towards its least-squares coefficient $\langle Y_j, r \rangle$, until some other competitor Y_k has as much correlation with the current residual as does Y_j.
4: Move θ_j and θ_k in the direction defined by their joint least squares coefficient of the current residual on $\langle Y_j, Y_k \rangle$, until some other competitor Y_l has as much correlation with the current residual.
5: If a nonzero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.
6: Continue in this way until all M variables and the constant variable $\mathbf{1}$ have been entered. After $min(B-1, M)$ steps, we arrive at the full least-squares solution and minimum number of nonzero θ_i's.

IV. Application in Phenetylamines

In order to evaluate the predictive ability and better understand the proposed models (23), we apply them to two benchmark datasets recommended by the International Academy of Mathematical Chemistry[1] for verification and comparison of new descriptors.

The first benchmark dataset, Phenet for short, is constituted by 22 phenetylamines with two substituent sites and its property of biological activity is given. A set of 110 molecular

[1] http://www.moleculardescriptors.eu/dataset/dataset.htm

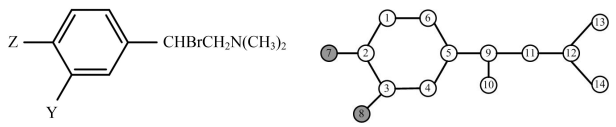
Algorithm 2 Pseudo-code of the proposed method

Require: Atomic information of compound, such as molar mass and charge; Global and local descriptors of the molecular graph, such as SC_i, BC_i.

1: Use atomic information and local descriptors to construct the graph signal $S \in \mathbb{R}^{N\times M_1}$.
2: Compute MDS descriptors $\zeta_k = \frac{1}{N}||S_k||_p$, $k = 1, 2, \cdots, M_1$.
3: Choose M_2 global descriptors as η_j, $j = 1, 2, \cdots, M_2$.
4: Define $\mathbf{x} = [\zeta_1, \cdots, \zeta_{M_1}, \eta_1, \cdots, \eta_{M_2}] \in R^{1\times M}$, and use B samples of $\mathbf{x}$ to form training data matrix $X = [\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_B^T]^T \in R^{B\times M}$.
5: For Model 1, let $Y = [X\ \mathbf{1}]$ then go to 11, otherwise continue.
6: Use the ith and jth columns of X to calculate the distance matrix $E_{i,j} = \| X_i^{\mathrm{T}} - X_j^{\mathrm{T}} \|_2, i, j \in \{1, 2, \cdots, M\}$.
7: Compute the affinity matrix A using adaptive Gaussian kernel $A_{i,j} = exp(-E_{i,j}/\sigma_j)^2$, σ_j = the $(knn + 1)$th smallest $E_{i,j}$ in the jth column of E.
8: Symmetrize A to $\tilde{A} = A + A^T$ and column normalize $\tilde{A}$ to Markov transition matrix K with $K_{i,j} = \tilde{A}_{i,j} / \sum_k \tilde{A}_{i,k}$.
9: Construct the Laplacian matrix $L_d = I_M - K$.
10: Filter the training data matrix X to obtain $\tilde{X} = XL_d^l$ and let $Y = [\tilde{X}\ \mathbf{1}]$.
11: Use Algorithm 1 to find the optimal sparse θ.
12: Assess $RMSE$, AAE and R^2 of the model.

descriptors is also given, which are calculated by DRAGON software (version 5.4).

Phenet dataset contains data on the adrenergic blocking potencies of N, N-dimethyl-2-bromo-phenetylamines in the rat with varying structures as illustrated in Fig. 3(a). The Z and Y, substituents on the ring, are H, F, Cl, Br, I, or CH_3. Compounds show different biological activities according to the different combinations of substituents. The same graph G for this dataset is shown in Fig. 3(b), in which the two gray vertices are the substituents Z and Y.



(a) Varying structures (b) The same graph G

Fig. 3: Phenet dataset.

First, we design the MD signal $\mathbf{s}_i = [s_{i1}, s_{i2}, \cdots, s_{iM_1}]^{\mathrm{T}}$, with s_{i1} = betweenness centrality (BC_i), s_{i2} = stress centrality (SC_i), s_{i3} = vulnerability efficiency (VE_i), s_{i4} = atomic mass and s_{i5} = atomic charge. This yields a 5-dimensional vector signal $\mathbf{s}_i$ for each vertex v_i and the ith row of the signal matrix S described in Step 2. Using S_k, the kth column of S, in $\zeta_k = \frac{1}{N}\|S_k\|_p$, we obtain five MDS descriptors $\zeta_k, k = 1, 2, \cdots, 5$. Then, we use two spectrum-based descriptors, LE and $S_{L,2}$, and the 110 traditional topological descriptors given in the benchmark dataset Phenet to obtain $\eta_k, k = 1, 2, \cdots, 112$.

The descriptors ζ_k and η_k thus obtained constitute the input variable vector $\mathbf{x} \in R^M$ defined in (14), with $M = 117$, $x_k = \zeta_k$ for $k = 1, 2, \cdots, 5$, and $x_{k+5} = \eta_k =$ for $k = 1, 2, \cdots, 112$. Filtering $\mathbf{x}$ with L_d^l gives the $\tilde{\mathbf{x}} \in R^{117}$ defined in (17). Using the $\mathbf{x}$ and $\tilde{\mathbf{x}}$ respectively in (20) and (21) gives the following two models that relate the biological activity (BA) to the structure of the compound represented by these variables.

$$BA = \underbrace{\sum_{k=1}^{5} \theta_k x_k}_{\text{MDS descriptors}} + \underbrace{\sum_{k=1}^{112} \theta_{k+5} x_{k+5}}_{\text{traditional topological descriptors}} + c. \tag{24}$$

$$BA = \sum_{k=1}^{117} \theta_k \tilde{x}_k + c. \tag{25}$$

The model (24) has two parts: MDS descriptors and traditional topological descriptors, and the model (25) contains the graph filtered these two parts in each $\tilde{x}_k$ (see (18) for details). Thus, their performance is expected to be better than that of the model with only traditional topological descriptors in the second part of (24).

V. Application in Polyaromatic Hydrocarbons

The second benchmark dataset, PAH for short, are generally highly toxic and carcinogenic compounds and ubiquitous contaminants of aquatic and atmospheric ecosystems. PAH dataset is constituted by 82 polyaromatic hydrocarbons. The properties of melting point, boiling point and octanol-water partition coefficient are given. A set of 112 molecular descriptors is also given, which are calculated by DRAGON software (version 5.4).

The number of fused benzene rings contained in the compounds in this dataset is from 2 to 11, and some compounds also have substituents. Examples are shown in Fig. 4.

Since PAH dataset does not provide information about the atomic charge, we use s_{i1} = betweenness centrality (BC_i), s_{i2} = stress centrality (SC_i), s_{i3} = vulnerability efficiency (VE_i) and s_{i4} = atomic mass to form a 4-dimensional signal $\mathbf{s}_i$ for each vertex v_i and the ith row of S. Following the same procedure as that used for Phenetylamines, we use $\zeta_k = \frac{1}{N}\|S_k\|_p$ to obtain four MDS descriptors $\zeta_k, k = 1, 2, 3, 4$; and use two spectrum-based descriptors, LE and $S_{L,2}$, and the 112 traditional topological descriptors given in the benchmark dataset PAH to obtain $\eta_k, k = 1, 2, \cdots, 114$. Letting $x_k = \zeta_k$ for $k = 1, 2, 3, 4$, and $x_{k+4} = \eta_k$ for $k = 1, 2, \cdots, 114$, we obtain the input variable vector $\mathbf{x} \in R^{118}$ and the L_d^l filtered variable vector $\tilde{\mathbf{x}} \in R^{118}$.

Using the variables obtained above, we construct two models in (26) and (27) that relate the boiling point (BP) to the structure of the compound represented by these variables. The same as (24), (26) consists of the MDS descriptor part and the traditional descriptor part, while each $\tilde{x}_k$ in (27) contains graph filtered these two parts.

$$BP = \underbrace{\sum_{k=1}^{4} \theta_k x_k}_{\text{MDS descriptors}} + \underbrace{\sum_{j=1}^{114} \theta_{k+4} x_{k+4}}_{\text{traditional topological descriptors}} + c. \tag{26}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3032718, IEEE Transactions on Pattern Analysis and Machine Intelligence

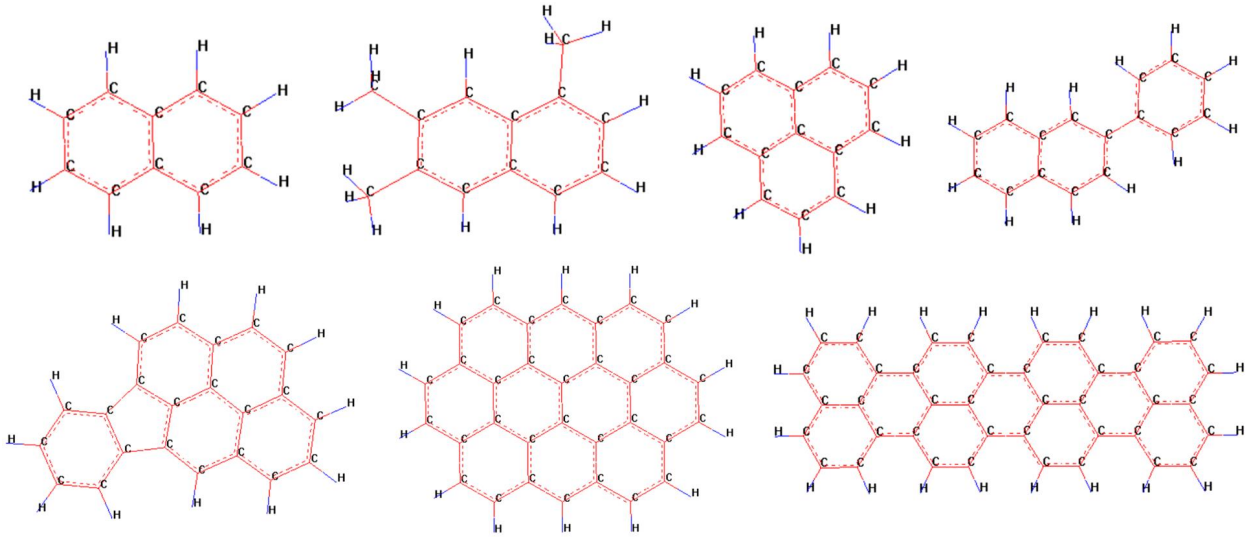


Fig. 4: Example compounds of PAH dataset.

$$BP = \sum_{k=1}^{118} \theta_k \tilde{x}_k + c. \tag{27}$$

VI. Results

In this section, we first present the uniqueness of molecular descriptors derived from the MD signal on compounds of different families with the same graph structure. Then, we evaluate our method by numerical experiments. The models (24)-(27) presented in Sections IV and V are learned from training data to study their behavior, and the performance indices introduced in Section II-D are used to evaluate their performance. All the computations are performed with Matlab.

A. Differentiation of Ethane/Ethylene/Acetylene

This subsection demonstrates the uniqueness of MDS descriptors on the compounds in different families with the same graph structure, which is the second case shown in Fig. 2. The only difference in molecular graphs of these compounds is their different weights. Different types of chemical bonds make them fall into different categories, so we just need to design non-degenerative molecular descriptors to distinguish them.

Ethane, ethylene and acetylene, as compounds with different chemical bonds containing two carbon atoms, are analyzed. Traditional topological descriptors always degenerate into the same values for these three compounds. For any real number d, the general Randic indices for ethane, ethylene and acetylene will always be the same value $R_{C_2H_6} = R_{C_2H_4} = R_{C_2H_2} = 1$. For any real number d, the general sum connectivity indices always have the same value $\chi_{C_2H_6} = \chi_{C_2H_4} = \chi_{C_2H_2} = 2^d$.

In our work, we design MDS descriptors based on GSP. A simple 2-dimensional signal based on eigenvalue and atomic mass is designed. The signal matrices for ethane, ethylene and acetylene are respectively

$$S_{C_2H_6} = \begin{bmatrix} 0 & 15 \\ 2 & 15 \end{bmatrix}, S_{C_2H_4} = \begin{bmatrix} 0 & 14 \\ 4 & 14 \end{bmatrix}, S_{C_2H_2} = \begin{bmatrix} 0 & 13 \\ 6 & 13 \end{bmatrix}.$$

When the 1-norm of each column is used to calculate $\zeta := [\zeta_1, \zeta_2]$ for the above three matrices, we get $\zeta_{C_2H_6} = [2, 30]$, $\zeta_{C_2H_4} = [4, 28]$ and $\zeta_{C_2H_2} = [6, 26]$, respectively. Similarly, when the 2-norm of each column is used, we get $\zeta_{C_2H_6} = [2, 21.21]$, $\zeta_{C_2H_4} = [4, 19.80]$ and $\zeta_{C_2H_2} = [6, 18.38]$, respectively. By defining a simple 2-dimensional signal on each vertex, these three compounds are easily distinguished. The design of the MD signal based on the important information (structural information, spatial information, etc.) of the compound guarantees the uniqueness of the new MDS descriptors.

This method of constructing molecular descriptors by means of GSP is also effective for other molecular structures with different chemical bonds and the same atomic type. High-dimensional signals can carry more information and are more suitable for complex compounds.

B. Phenetylamine Models

Phenet dataset is used as training dataset to learn the models (24) and (25). The BA values of the compounds in the dataset are used to form the output data vector Q. The MDS descriptors $x_k = \zeta_k = \frac{1}{N}||S_k||_1, k = 1, \cdots, 5$, are calculated for each compound in the dataset, and the values of 112 traditional descriptors of these compounds are used as $x_{k+5} = \eta_k, k = 1, 2, \cdots, 112$. The x_k's thus obtained for all the compounds are used to construct the data matrices X as in (15), and to calculate $\tilde{X}$ according to Steps 6-10 of Algorithm 2, with empirically determined $knn = 1$ and $l = 1$ in Steps 7 and 10, respectively. To compare with the traditional models using only topological descriptors, we also construct a data matrix X_d consisting of the last 112 columns of X, i.e., the values of 112 traditional descriptors of all the compounds.

Applying Algorithm 1 to the model output-input data pairs Q-X_d, Q-X and Q-$\tilde{X}$, respectively, we obtain the following three models, where most of the 117 input variables have been found insignificant and hence eliminated by Algorithm

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3032718, IEEE Transactions on Pattern Analysis and Machine Intelligence

1, resulting in very sparse optimal coefficient vectors θ's with dimension ≤ 5.

$$BA = 0.253S_{L,2} + 3.704WA - 1.269Dz - 1.772, \quad (28)$$

$$BA = 6.078\zeta_3 + 3.680WA - 1.259Dz - 7.586, \quad (29)$$

$$BA = 41.130\tilde{x}_1 + 16.121\tilde{x}_4 + 2.817\tilde{x}_5 + 8.157\tilde{x}_7 + 0.963. \quad (30)$$

The model (28) uses only traditional descriptor variables $(S_{L,2}, WA, Dz)$, the model (29) uses both traditional descriptor (WA, Dz) and new MDS descriptor (ζ_3) variables, while the model (30) uses graph filtered both types of descriptor variables. These input variables are singled out by Algorithm 1 as the significant ones among the 112 traditional descriptors in X_d, the 117 variables in X, and the 117 variables in $\tilde{X}$, respectively.

TABLE I: Performance comparison of models for Phenet dataset

	model (28)	model (29)	model (30)
R^2	0.955	0.960	0.966
$RMSE$	0.118	0.112	0.102
AAE	0.105	0.101	0.088

Table I compares the performance indices R^2, $RMSE$ and AAE of the above three models. It is clear from the table that the models (29) and (30), using both MDS and traditional descriptors as inputs, outperform the model (28) using only traditional descriptor inputs, and the model (30) using the graph (L_d^l) filtered input variables $\tilde{x}_k$ performs best.

The better performance of model (30) stems from the enhanced dissimilarity of the graph filtered input variables $\tilde{x}_k$ as compared with the original input variables x_k. This is shown in Fig. 5 using the first eight elements of the input vector $\mathbf{x}$ for a compound in Phenet dataset. It can be seen that the correlation of input variables is significantly reduced after graph filtering.

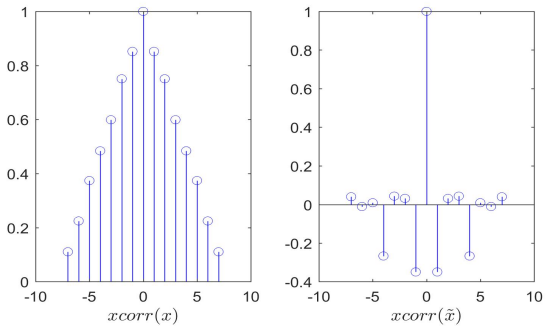


Fig. 5: Correlation coefficients of the first eight input variables for a compound in Phenet dataset before (left) and after (right) graph filtering.

For the model (30), the absolute errors are plotted in Fig. 6 against the experimental biological activity. Its AAE is 0.088 with the minimum absolute error 0.01 and the maximum absolute error 0.23. The minimum and maximum values of the corresponding relative error are 0.11% and 2.72%, respectively. The range of biological activity of these compounds is 7.56∼9.52. When this model is used for clustering, that is, dividing Phenet dataset into two categories on average based on the median 8.86, all compounds can be classified correctly. This further verifies the validity of the model.

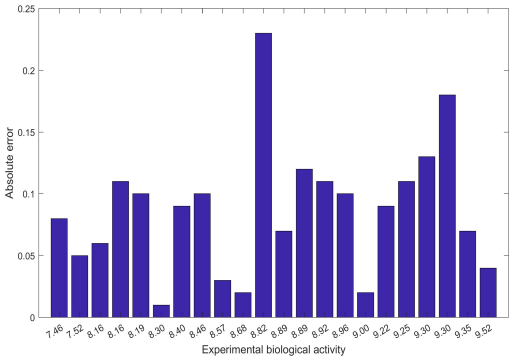


Fig. 6: Absolute errors of the model (30) of Phenet dataset.

Each dimension of the MD signal S reflects a certain character (chemical/structure) of the compound, with different compounds having different MD signals. Therefore, when very few variables are used to distinguishing or modeling different compounds, the proposed method still performs well.

Table II compares the QSAR models of the Phenet dataset reported in [35–38] with the models (29) and (30), where PCA/NN is a neural network model that uses principle components derived from the full data set; MLR-EM is a multiple linear regression model employing an expectation maximization method; EM/NN is a neural network model using sparse descriptors derived from the MLR-EM algorithm. The comparison reveals the superiority of the models (29) and (30) in predicting the biological activity of phenetylamines. As seen from the table, the proposed models (29) and (30) outperform previously reported models. Using fewer descriptors, the proposed models still outperform the neural network model using more descriptors.

TABLE II: Comparison of the proposed models with other works for Phenet dataset

	Model	R^2	$RMSE$	No. of descriptors
Vukicevic et al. [35]	MLR	0.5405	–	1
Burden et al. [36]	MLR-EM χ=0.035	0.750	0.265	3
Burden et al. [36]	PCA/NN 1 node	0.936	0.130	10
Burden et al. [36]	EM/NN 3 nodes	0.959	0.260	7
Unger and Hansch [37]	MLR	0.964	0.164	3
Unger and Hansch [37]	MLR	0.944	0.197	2
Todeschini et al. [38]	MLR	0.845	0.177	3
This work	Model (29)	0.960	0.112	3
This work	Model (30)	0.966	0.102	4

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3032718, IEEE Transactions on Pattern Analysis and Machine Intelligence

C. Polyaromatic Hydrocarbon Models

PAH dataset is used as training dataset to learn the models (26) and (27). The BP values of the compounds in the dataset are used to form the output data vector Q. The $x_k = \zeta_k = \frac{1}{N}||S_k||_1, k = 1, \cdots, 4$, are calculated for each compound in the dataset, and the values of 114 traditional descriptors of these compounds are used as $x_{k+5} = \eta_k, k = 1, 2, \cdots, 112$. The x_k's thus obtained for all the compounds are used to construct the data matrices X as in (15), and to calculate $\tilde{X}$ according to Steps 6-10 of Algorithm 2, with empirically determined $knn = 3$ and $l = 2$ in Steps 7 and 10, respectively. To compare with the traditional models using only topological descriptors, we also construct a data matrix X_d consisting of the last 114 columns of X.

Applying Algorithm 1 to the output-input data pairs Q-X_d, Q-X and Q-$\tilde{X}$, respectively, we obtain the following three models. Again, most of the 118 input variables have been found insignificant and hence eliminated by Algorithm 1, resulting in very sparse optimal coefficient vectors θ's with dimension ≤ 6.

$$BP = 0.842ZM1V + 0.150LE + 0.023, \tag{31}$$

$$BP = 0.901\zeta_4 + 0.946S_{L,2} - 0.799, \tag{32}$$

$$\begin{aligned} BP = \quad & 10.299\tilde{x}_1 - 19.172\tilde{x}_3 - 17.303\tilde{x}_4 + \\ & 4.737\tilde{x}_7 + 3.763\tilde{x}_8 + 1.029. \end{aligned} \tag{33}$$

Of these models, (31) is a model with only traditional descriptor variables ($ZM1V, LE$); (32) is a model with both traditional descriptor ($S_{L,2}$) and new MDS descriptor (ζ_3) variables; and (33) is a model with graph filtered both types of variables. These input variables are singled out by Algorithm 1 as the significant ones among the 114 traditional descriptors in X_d, the 118 variables in X, and the 118 variables in $\tilde{X}$, respectively.

Table III compares the performance indices R^2, $RMSE$ and AAE of the above three models. Similar to Phenetylamines case, the models (32) and (33), using both new MDS and traditional descriptors as inputs, outperform the model (31) using only traditional descriptor inputs; and the model (33) using graph (L_d^l) filtered input variables $\tilde{x}_k$ performs best, showing again the advantage of graph filtering of input variables in performance enhancement.

TABLE III: Performance comparison of models for PAH dataset

	model (31)	model (32)	model (33)
R^2	0.981	0.982	0.985
$RMSE$	11.198	11.177	10.196
AAE	7.766	7.044	6.774

Fig. 7 plots the absolute errors of the model (33) against the experimental boiling point. The wide range of boiling point of this dataset, 178∼519, results in relatively large errors. This model shows an excellent performance and the average value of the absolute relative error is merely 2.08%. The absolute errors of 1-phenylnaphthalene (BP = 334) and azulene (BP = 270) are greater than 35. The absolute errors of and 2-7-dimethylpyrene (BP = 396) and acenaphthylene (BP = 270) are greater than 20. With the exception of these four compounds, the absolute errors of the others are less than 15 and the corresponding AAE drops to 4.34. Considering the large value of boiling point, this model is good for fitting.

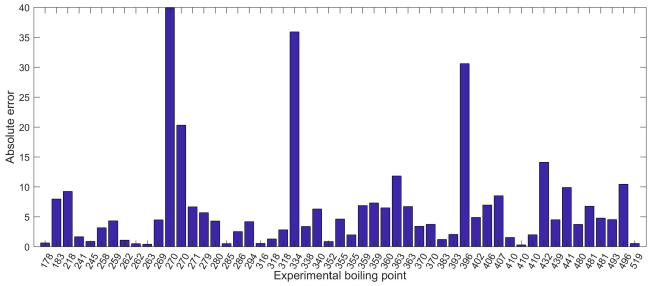


Fig. 7: Absolute errors of the model (33) of PAH dataset.

The two input variables in model (32) have clear chemistry meaning. The traditional eigenvalue-based descriptor $S_{L,2}$ describes the information of molecular graph spectrum of the compound, while the MDS descriptor ζ_4 is actually the average atomic mass representing the information of molecular mass. Fig. 8 plots the 3D surfaces of boiling point, $S_{L,2}$ and ζ_4 of the compounds in PAH dataset. The apparent linear dependence of BP on $S_{L,2}$ and ζ_4 shown in the plot is consistent with the linear model (32) we have obtained, and asserts the validity of the model. From this model, we may conclude that the boiling point of a compound can be predicted quite precisely with the graph spectral information $S_{L,2}$ and the molecular mass ζ_4.

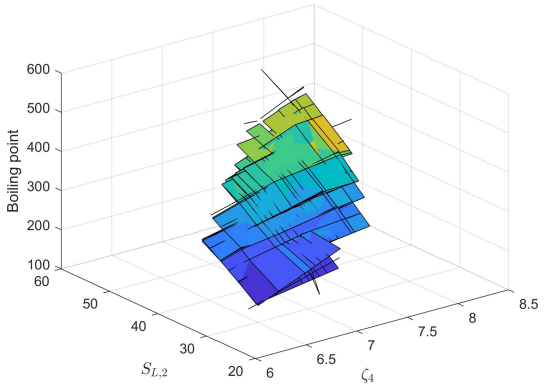


Fig. 8: Relationship of variables and boiling point for model (32) of PAH dataset.

Table IV compares the QSPR models of polyaromatic hydrocarbons reported in [35, 39, 40] with the models (32) and (33). In [39], a set of 67 non-substituted PAHs containing 2 to 7 fused rings with five and six carbon atoms were studied. The 23 non-substituted PAHs studied in [40] contain 1 to 6 fused rings. Whereas, the dataset used in this paper is more complex, with the number of fused rings ranging from 2 to 11 and some PAHs also having substituents. This makes QSPR modeling harder. As seen from the table, the prediction performance R^2 of the models (32) and (33) on this harder dataset is

comparable to that of [35, 39, 40] on those easier datasets. For compounds in the same families with the same or similar graph structure, the models (32) and (33) not only solve the problem of degeneration of molecular descriptors, but also can predict properties of compounds efficiently.

TABLE IV: Comparison of the proposed models with other works for polyaromatic hydrocarbons

	Model	R^2	No. of descriptors	No. of training set
Vukicevic et al. [35]	MLR	0.980	1	53
Ribeiro et al. [39]	PCR	0.995	3	36
Ribeiro et al. [39]	PLS	0.995	2	36
Ferreira [40]	PLS	0.998	3	23
This work	Model (32)	0.982	2	53
This work	Model (33)	0.985	5	53

VII. Discussions

A challenging issue in QSAR/QSPR research is that its performance depends mainly on the quality of molecular descriptors used. Based only on the properties of vertices and edges, traditional molecular descriptors always degenerate to the same values for different compounds having the same or similar molecular graph. Therefore it is difficult to obtain reliable and effective QSAR/QSPR models using these traditional descriptors.

To solve the problem of descriptor degeneration, we have used the GSP's foundation concept of signals on graph to design an MD signal with distinctive information of compound for each vertex of molecular graph. By GSP analysis of the MD signals on vertices, we have derived new MDS descriptors that can better distinguish the compounds with the same or similar molecular graph structures.

To solve the problem in model reliability and performance, we have combined new MDS descriptors with traditional molecular descriptors in linear regression to derive the new QSAR/QSPR model (20) with enhanced reliability and performance. To further enhance model reliability and performance, we have introduced the descriptor graph to derive the Laplacian graph filter, and used the highpass nature of the filter to enhance the dissimilarity of descriptors. Using the dissimilarity enhanced descriptors in linear regression, we have obtained the QSAR/QSPR model (21) with further enhanced performance and reliability. The model (20) with descriptors as regression variables is useful in revealing the QSAR/QSPR between a compound and its descriptors, as shown in Section VI-C; while the model (21), with descriptors hidden in the filtered input variables, is useful for reliable prediction of the biological activity or physicochemical property of compounds in practice.

Existing works, whether in descriptor design or in modeling, are based solely on the molecular graph. Our GSP-based approach, as summarized above, is very different from those of previous works. This novel approach has resulted in significant performance and reliability improvement in QSAR/QSPR modeling as shown in two application examples in Section IV. To the best of our knowledge, this work is the first attempt to study and solve QSAR/QSPR modeling problem from the perspective of GSP. The graph filtering of input variables used in our approach might be useful for other regression based model learning problems.

This work intends to draw the attention of researchers in GSP and machine learning communities for further research on GSP based QSAR/QSPR model learning. In this work, GSP has been applied in the modeling of biological activity and boiling point. Other information can be included in the proposed method for analyzing more complex properties, such as binding affinity, lethal dose, and octanol/water partition coefficient. Learning approaches, such as GNN, might be used to establish nonlinear models and explore deeper relationships between the molecular structure and these complex properties.

VIII. Conclusions

A new paradigm for QSAR/QSPR modeling based on GSP has been proposed to learn the quantitative relationship between the physicochemical/activity property and the structure of compounds.

An MD signal has been introduced for each vertex of the molecular graph. By analyzing the MD signals, a number of new (MDS) molecular descriptors with higher discriminability have been designed. Combination of these new descriptors and traditional descriptors as the inputs in linear regression has resulted in new QSAR/QSPR models with enhanced performance.

Treating the combined descriptor variables as the signals on a descriptor graph, a novel approach has been presented to derive the descriptor graph and its Laplacian graph filter from the descriptor data. The Laplacian filtered input variables with enhanced dissimilarity have been used as inputs in linear regressions to derive new QSAR/QSPR models with further enhanced performance.

For datasets containing the compounds with the same or similar graph structure, our proposed models have shown better performance. We have also provided a new insight from chemistry into the boiling point model of compounds.

The results of this paper have provided deeper understanding and new approaches for compound prediction and classification, with application potential in various areas such as biochemical and pharmaceutical engineering. These results have also shed some new lights on regression based learning.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (grants 61625305, 61801338 and 61801339).

References

[1] S. M. Hosamani, B. B. Kulkarni, R. G. Boli, et al., "QSPR analysis of certain graph theocratical matrices and their corresponding energy," *Applied Mathematics and Nonlinear Sciences*, vol. 2, no. 1, pp. 131-150, 2017.

[2] M. Dehmer, F. Emmert-Streib, and Y. Shi, "Quantitative graph theory," *Information Sciences: an International Journal*, vol. 418(C), pp. 575-580, 2017.

[3] O. B. Ghanem, M. A. Mutalib, J. M. Leveque, et al., "Development of QSAR model to predict the ecotoxicity of Vibrio fischeri using COSMO-RS descriptors," *Chemosphere*, vol. 170, pp. 242-250, 2017.

[4] T. Le, V. C. Epa, F. R. Burden, and D. A. Winkler, "Quantitative structure-property relationship modeling of diverse materials properties," *Chemical Reviews*, vol. 112, no. 5, pp. 2889-2919, 2012.

[5] A. Cooper, T. Potter, and T. Luker, "Prediction of efficacious inhalation lung doses via the use of in silico lung retention quantitative structure-activity relationship models and in vitro potency screens," *Drug Metabolism and Disposition*, vol. 38, no. 12, pp. 2218-2225, 2010.

[6] P. Liu and W. Long, "Current mathematical methods used in QSAR/QSPR studies," *International Journal of Molecular Sciences*, vol. 10, no. 5, pp. 1978-1998, 2009.

[7] Z. Cheng, A. S. Zhu, L. Q. Zhang, "Quantitative analysis of electronic absorption spectroscopy by piecewise orthogonal signal correction and partial least square," *Spectroscopy & Spectral Analysis*, vol. 28, no. 4, pp. 860-864, 2008.

[8] A. Baghban, J. Sasanipour, S. Habibzadeh, et al., "Estimating solubility of supercritical H2S in ionic liquids through a hybrid LSSVM chemical structure model," *Chinese Journal of Chemical Engineering*, 2018.

[9] A. L. Teixeira, J. P. Leal, and A. O. Falcao, "Random forests for feature selection in QSPR Models-an application for predicting standard enthalpy of formation of hydrocarbons," *Journal of Cheminformatics*, vol. 5, no. 1, pp. 1-15, 2013.

[10] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, et al., "Convolutional networks on graphs for learning molecular fingerprints," *Advances in Neural Information Processing Systems*, pp. 2224-2232, 2015.

[11] M. Dehmer, M. Grabner, and K. Varmuza, "Information indices with high discriminative power for graphs," *PLoS One*, vol. 7, no. 2, pp. e31214, 2012.

[12] M. V. Diudea, A. Ilic, K. Varmuza, and M. NDehmer, "Network analysis using a novel highly discriminating topological index," *Complexity*, vol. 16, no. 6, pp. 32-39, 2011.

[13] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1644-1656, 2013.

[14] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3042-3054, 2014.

[15] D. I. Shuman, S. K. Narang, P. Frossard, et al., "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83-98, 2013.

[16] A. Ortega, P. Frossard, J. Kovacevic, et al., "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808-828, 2018.

[17] H. Bahonar, A. Mirzaei, S. Sadri, et al., "Graph embedding using frequency filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, DOI: 10.1109/TPAMI.2019.2929519.

[18] M. M. Bronstein, J. Bruna, Y. LeCun, et al., "Geometric deep learning: going beyond Euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18-42, 2017.

[19] A. Kheradmand, P.Milanfar, "A general framework for regularized, similarity-based image restoration," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5136-5151, 2014.

[20] B. Lucic, N. Trinajstic, and B. Zhou, "Comparison between the sum-connectivity index and product-connectivity index for benzenoid hydrocarbons," *Chemical Physics Letters*, vol. 475, no. 1-3, pp. 146-148, 2009.

[21] S. Hayat, M. Imran, and J. B. Liu, "Correlation between the Estrada index and π-electronic energies for benzenoid hydrocarbons with applications to boron nanotubes," *International Journal of Quantum Chemistry*, vol. 119, no. 23, pp. e26016, 2019.

[22] S. Hayat, S. Khan, A. Khan, et al., "Valency-based molecular descriptors for measuring the π-electronic energy of lower polycyclic aromatic hydrocarbons," *Polycyclic Aromatic Compounds*, pp. 1-17, 2020.

[23] D. Vukicevic and A. Graovac, "Note on the comparison of the first and second normalized zagreb eccentricity indices," *Acta Chimica Slovenica*, vol. 57, no. 3, pp. 524-528, 2010.

[24] R. Todeschini and C. Viviana, *Handbook of Molecular Descriptors*, vol. 11. John Wiley & Sons, 2008.

[25] I. Gutman and B. Zhou, "Laplacian energy of a graph," *Linear Algebra and Its Applications*, vol. 414, no. 1, pp. 29-37, 2006.

[26] M. Dehmer, L. Sivakumar, and K. Varmuza, "Uniquely discriminating molecular structures using novel eigenvalue-based descriptors," *Match-Communications in Mathematical and Computer Chemistry*, vol. 67, no. 1, pp. 147, 2012.

[27] U. Brandes and T. Erlebach, *Network Analysis: Methodological Foundations*, Berlin: Springer, 2005.

[28] L. F. Costa, F. A. Rodrigues, G. Travieso, et al., "Characterization of complex networks: A survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167-242, 2007.

[29] C. B. Santiago, J. Y. Guo, and M. S. Sigman, "Predictive and mechanistic multivariate linear regression models for reaction development," *Chemical Science*, vol. 9, no. 9, pp. 2398-2412, 2018.

[30] A. Lusci, G. Pollastri, and P. Baldi, "Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules," *Journal of Chemical Information and Modeling*, vol. 53, no. 7, pp. 1563-1575, 2013.

[31] K. Mansouri, C. M. Grulke, R. S. Judson, et al., "OPERA models for predicting physicochemical properties and environmental fate endpoints," *Journal of Cheminformatics*,

vol. 10, no. 1, pp. 10, 2018.

[32] R. Todeschini, D. Ballabio, and F. Grisoni, "A comparative study of regression metrics for predictivity assessment of QSAR models," *Journal of Chemical Information and Modeling*, vol. 56, no. 10, pp. 1905-1913, 2016.

[33] D. Van Dijk, J. Nainys, R. Sharma, et al., "MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data," *bioRxiv*,2017.

[34] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media, 12th Printing, 2017.

[35] D. Vukicevic, N. Trinajstic, "Bond-additive modeling. 3. Comparison between the product-connectivity index and sum-connectivity index," *Croatica Chemica Acta*, vol. 83, no. 3, pp. 349-351, 2010.

[36] F. R. Burden, D. A. Winkler, "Optimal sparse descriptor selection for QSAR using Bayesian methods," *QSAR & Combinatorial Science*, vol. 28, no. 6-7, pp. 645-653, 2009.

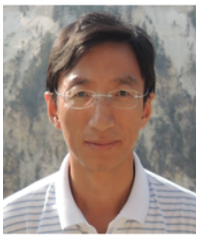
[37] S. H. Unger, C. Hansch, "Model building in structure-activity relations. Reexamination of adrenergic blocking activity of. beta.-halo-. beta.-arylalkylamines," *Journal of Medicinal Chemistry*, vol. 16, no. 7, pp. 745-749, 1973.

[38] R. Todeschini, G. Paola, "3D-modelling and prediction by WHIM descriptors. Part 6. Application of WHIM descriptors in QSAR studies," *Quantitative Structure-Activity Relationships*, vol. 16, no. 2, pp. 120-125, 1997.

[39] F. A. de Lima Ribeiro, M. MC. Ferreira, "QSPR models of boiling point, octanol-water partition coefficient and retention time index of polycyclic aromatic hydrocarbons," *Journal of Molecular Structure: THEOCHEM*, vol. 663, no. 1-3, pp. 109-126, 2003.

[40] M. MC. Ferreira, "Polycyclic aromatic hydrocarbons: a QSPR study," *Chemosphere*, vol. 44, no. 2, pp. 125-146, 2001.

Xiaoying Song received the B.S. degree in electronic science and technology and the Ph.D. degree in microelectronics and solid-state electronics from Wuhan University, Wuhan, China, in 2012 and 2017, respectively. Since July 2017, she has been with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan. Her research interests include image compression, graph signal processing.



Li Chai received the B.S. degree in applied mathematics and the M.S. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1994 and 1997, respectively, and the Ph.D. degree in Electrical engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2002. In September 2002, he joined Hangzhou Dianzi University, China. He worked as a postdoctoral research fellow at the Monash University, Australia, from May 2004 to June 2006. In 2008, he joined Wuhan University of Science and Technology, where he is currently a Chutian Chair Professor. He was a visiting researcher at Newcastle University in 2009, Central Queensland University in 2011, and Harvard University in 2015. His research interests include distributed optimization, filter bank frames, graph signal processing, and networked control systems. Professor Chai is the recipient of the Distinguished Young Scholar of the National Science Foundation of China. He is currently an associate editor for the Decision and Control.

Jingxin Zhang (M02) received the M.E. and Ph.D. degrees in electrical engineering from Northeastern University, Shenyang, China. Since 1989, he has held research and academic positions in Northeastern University, China, the University of Florence, Italy, the University of Melbourne, the University of South Australia, Deakin University and Monash University, Australia. He is currently Associate Professor of Electrical Engineering, Swinburne University of Technology, and Adjunct Associate Professor of Electrical and Computer Systems Engineering, Monash University, Melbourne, Australia. His research interests include signals and systems and their applications to biomedical and industrial systems. He is the recipient of the 1989 Fok Ying Tong Educational Foundation (Hong Kong) for the Outstanding Young Faculty Members in China, and 1992 China National Education Committee Award for the Advancement of Science and Technology.