Recursive Neural Networks for Undirected Graphs for Learning Molecular Endpoints

Ian Walsh, Alessandro Vullo, and Gianluca Pollastri

School of Computer Science and Informatics and
Complex and Adaptive Systems Laboratory
University College Dublin, Belfield, Dublin 4, Ireland

Abstract. Accurately predicting the endpoints of chemical compounds is an important step towards drug design and molecular screening in particular.

Here we develop a recursive architecture that is capable of mapping Undirected Graphs into individual labels, and apply it to the prediction of a number of different properties of small molecules. The results we obtain are generally state-of-the-art.

The final model is completely general and may be applied not only to prediction of molecular properties, but to a vast range of problems in which the input is a graph and the output is either a single property or (with small modifications) a set of properties of the nodes.

1 Introduction

Cost-free, time-efficient computational screening of a large set of compounds capable of excluding a sizeable fraction of them before the testing phase would dramatically reduce the cost of drug design and significantly quicken its pace. The Quantitative Structure-Property/Activity Relationship (QSPR/QSAR) approach dates back as far as forty years ago [1], and relies on finding an appropriate function that maps a molecular compound into a property/activity of interest. Machine learning techniques can be used to tackle this problem, but molecules are inherently structured as graphs and there is no single conclusive solution to design statistical methods that deal with graphs. An early solution to this problem has been to "flatten" a molecule into a fixed-size vector of properties, or features, generally hand-crafted, which are then input to a traditional machine learning tool such as a Neural Network (NN) or a Support Vector Machine (SVM). For instance in [2, 3] this approach is followed to predict aqueous solubility by a Multi-Layer Perceptron (MLP), while in [4] features are incrementally selected to be input to an SVM. In [5] a large number of 2D and 3D features, capturing physiochemical and graph properties of molecules, are input to an NN to predict melting point after being compressed by Principal Component Analysis (PCA). In [6] atomic contributions containing correction factors for intramolecular interactions are derived by multivariate regression, yielding accurate predictions of octanol-wated partition coefficients. Although it is clear that this two-stage approach (encode a molecule as features, map the features into the property) may be successful, it has a number of drawbacks: often, one or more experts need to design the features, thus creating a bottleneck; features are often problem-specific; features may not be optimal; however the features are

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 391–403, 2009.

designed, aspects of the structure/connectivity may be lost or it may be decided arbitrarily which ones to represent, thus potentially missing vital information about the mechanisms involved.

Structural alert (SA) methods search for patterns within datasets that are indicative of the molecules' properties. In [7] mutagenicity classification is predicted with good levels of accuracy using a manual derivation of these substructures. An updated version is introduced in [8] to automatically mine the substructures. In [9] a vector of substructure frequencies is input to an SVM, yielding fairly accurate predictions of cancer toxicity, HIV suppression and potential for anthrax binding. Despite the successes of these methods, their generalisation ability is debatable, and their failure to predict carcinogenicity in a sustained way has been attributed to the evolving nature of chemical datasets [10], in which new, unknown substructures keep appearing. Similarly to homology modelling for protein structure prediction, some molecules will need to be predicted "ab initio" as they contain novel active substructures or neighbourhoods thereof. More recently, kernel methods that integrate some form of structural processing into their kernel function have shown state-of-the-art performances at many tasks [11]. Melting point and octanol-water partition coefficient are predicted in [12] by 2D kernels with minmax similarity and 3D histogram kernels. State-of-the-art results are reported for a the classification of cancer and HIV suppression in [13], by 2D and 3D weighted decomposition kernels with the best results reported for a combination of both. In [14] kernels on molecular fingerprint similarity matching in the 2D case and atomic distances in the 3D case are state-of-the-art for mutagenicity and human tumor suppression.

In this work we design a novel class of machine learning algorithms for processing structured data. We tackle "ab initio" predictions of a number of properties of small molecules (i.e. we do not mine substructures). The algorithms we describe are based on recursive neural networks and they deal with molecules directly as graphs, in that no features are manually extracted from the structure, and the networks automatically identify regions and substructures of the molecules that are relevant for the property in question. The basic structural processing cell we use is similar to those described in [15, 16, 17, 18], and adopted in essentially the same form in applications including molecule regression/classification [19, 20, 21], image classification [22], natural language processing [23], face recognition [24]. In the case of molecules, there are numerous disadvantages in these earlier models: they can only deal with trees, thus molecules (that are more naturally described as Undirected Graphs (UG)) have to be preprocessed before being input; the preprocessing is generally task-dependent; special nodes ("super-sources") have to be defined for each molecule; application domains are generally limited, thus the effectiveness of the models is hard to gauge. In this work, although we loosely build on these previous works, we extend them in two crucial directions: our models deal directly with UG; no preprocessing is necessary, and no part of the molecule has to be marked as a "super-source". We term our model UG-RNN, or Recursive Neural Networks for Undirected Graphs.

We apply UG-RNN to the prediction of aqueous solubility, melting point and octanol water partition coefficient (all regression tasks) and to the classification of mutagenicity. Our results are encouraging, outperforming or matching state-of-the-art kernels on the same regression datasets. We alter the models slightly for mutagenicity prediction in

order to test whether our approach incorporates useful contextual information In this case we show that UG-RNN outperform a state-of-the-art SA method and only perform less accurately than a method based on SVM's fed with a task-specific feature which is not available to our model [25].

UG-RNN are open-ended in that they can be used to learn on any molecular dataset, and are portable to other molecular biology problems that require graph processing such as phylogenetic graph/tree analysis, protein classification when the protein is represented as a graph of contacts, etc.

2 Methods

A molecule is naturally described as a UG, possibly with cycles, where atoms represent vertices and bonds represent edges. Here we factorise the UG representing a molecule into N Directed Acyclic Graphs, where: N is the total number of atoms/nodes in the molecule; the k^{th} DAG is obtained from the UG describing the molecule by directing all its edges along the shortest path to its k^{th} atom v_k. The order of the atoms in the molecule is unimportant, as the result of processing is independent on it. Figure 1 shows how the undirected graph of nitrobenzene can be represented as 9 DAG's.

Let $ch^1_{[v,k]}, \ldots, ch^n_{[v,k]}$ be the children of atom/node v in the k^{th} DAG, then we assume that there is a hidden vector $\mathbf{G}_{v,k} \in \mathbb{R}^m$ describing the contextual information *upstream* of node v as:

$$\mathbf{G}_{v,k} = \mathcal{M}^{(G)}\left(i_v, \mathbf{G}_{ch^1_{[v,k]}}, \ldots, \mathbf{G}_{ch^n_{[v,k]}}\right) \tag{1}$$

where $i_v \in \mathbb{R}^l$ is the label associated with node v (i.e. essentially, all the information to input about the atom v). When a node has no children, or fewer children than the maximum allowed (n) then the empty arguments in $\mathcal{M}^{(G)}()$ are set to vectors of zeroes

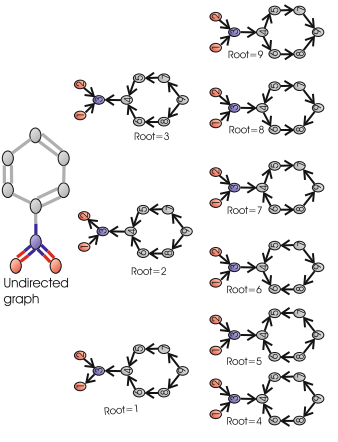

Fig. 1. The undirected graph of nitrobenzene and the 9 DAG's derived from the molecule

(boundary conditions). The maximum number of children n is set to the maximum out-degree of all the vertices in the structures, and is normally $n = 4$ for molecules. We realise the function $\mathcal{M}^{(G)}$, or state transition function, by a two layered perceptron. In the most basic model we assume *stationarity*, that is the same function (thus network) is used to process all vertices in all DAG's. This may be regarded as a form of weight sharing, and helps keep the number of free parameters in the model low.

Given there are as many DAG's as there are nodes in the UG describing the molecule, and given each node v_k is a root in a graph, there will be N vectors associated with root nodes: $\mathbf{G}_{v_k,k}$. Each of these vectors provides a description of the molecule "as seen" from v_k, and may be regarded as a way of "localising" the computations. Although it may be possible to analyse individual vectors $\mathbf{G}_{v_k,k}$ to point out which parts of a molecule are relevant to determine a property, we have not focussed our work on this task at present. To map the complex of these vectors into a single property, we first add them up:

$$\mathbf{G}_{structure} = \sum_{k=1}^{N} \mathbf{G}_{v_k,k} \tag{2}$$

Notice how: each atom in the molecule is within one transition function (i.e. one Two-Layered Perceptron) from the vector representing the whole molecule, thus minimising the well know vanishing gradient problem, which affects recursive neural networks and deep networks in general [26]; given that all atoms compete to be represented in $\mathbf{G}_{structure}$, if this vector is selected with the purpose of predicting a given property, then it effectively represents a task-dependent encoding (compression) of the molecule.

We map $\mathbf{G}_{structure}$ into the property of interest as:

$$o = \mathcal{M}^{(O)}(\mathbf{G}_{structure}) \tag{3}$$

We implement $\mathcal{M}^{(O)}$ by a Two-Layered Perceptron with a linear output when we predict real-valued properties, and softmax units in case of classification. The error function we minimise is, respectively, a sum of squared differences between target and network output, and the relative entropy between target and output. The whole network (all DAG's, sum of hidden vectors of root nodes, and output function) is trained by gradient descent. The gradient of the error is computed, exactly, by the backpropagation algorithm, which we can apply here given that the overall network including the molecule has no cycles.

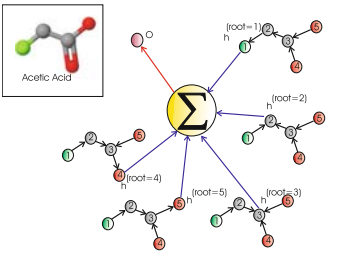

Fig. 2. An example of the model used on acetic acid. All the variables are explained in the text.

Figure 2 shows how the 5 DAG's of acetic acid are composed to build a UG-RNN. Each DAG produces a distinct contextual vector at each root. These vectors are then added and mapped into the final output representing the desired property.

2.1 Relaxing Stationarity

We also implement a system in which we relax the stationarity hypothesis. In this case we search for common bonding patterns for an atom (or common neghbourhoods for a node), and we implement dedicated transition functions for the most frequent ones. For example the same transition network/function will represent every carbon atom with a single nitrogen and double bonded oxygen. A neighbourhood token is created by investigating each atom in the *training set* and storing a the atom symbol and its immediate neighbours. Neighbourhood tokens that contain all carbon symbols are removed because they are considered non-informative. The T most frequent tokens yield special transition functions, while all the other patterns are handled by a general function:

$$\begin{aligned}
\mathbf{G}^{1}_{v,k} &= \mathcal{M}^{(1)}\left(i_v, \mathbf{G}^{u(1)}_{ch^1_{[v,k]}}, \ldots, \mathbf{G}^{u(n)}_{ch^n_{[v,k]}}\right) \\
&\ldots \\
\mathbf{G}^{T}_{v,k} &= \mathcal{M}^{(T)}\left(i_v, \mathbf{G}^{u(1)}_{ch^1_{[v,k]}}, \ldots, \mathbf{G}^{u(n)}_{ch^n_{[v,k]}}\right) \\
\mathbf{G}^{general}_{v,k} &= \mathcal{M}^{(general)}\left(i_v, \mathbf{G}^{u(1)}_{ch^1_{[v,k]}}, \ldots, \mathbf{G}^{u(n)}_{ch^n_{[v,k]}}\right)
\end{aligned} \tag{4}$$

where $u(c) \in \{1, \ldots, T, general\}$ is the index of the transition function applied to $ch^c_{[v,k]}$ based on its identity and neighbours.

2.2 Atomic Label

We keep the label i_v attached to an atom v as simple as possible in order to make the architecture portable. It should be noted, though, that any feature of an atom or of its neighbourhood can be included into i_v. Each atom is labelled as follows:

- The element type.
- The atom charge.
- Using openbabel version 2.1.1 [27] we calculate the Smallest Set of Smallest Rings (SSSR), hybridization and aromaticity of an atom.

2.3 Training Procedure

When training for regression all target endpoints are normalised between [0,1] by finding the maximum and minimum target values in the training set of each fold and normalising all targets in the training and testing fold to $\frac{target-max}{max-min}$. We use a sigmoid activation function for the final output neuron and perform gradient descent on a squared error. For classification softmax activation function is used at the output neurons with relative cross entropy as the cost function. All inner neurons have a $\tanh()$ activation function irrespective of regression or classification.

Weights are randomly initialised. We update the weights once per epoch (batch learning). A gradient component dw is applied directly when its absolute value is in the [0.1,1] range, but set to sign(dw) when greater than 1 and to $0.1 \times$sign(dw) when smaller than 0.1. We train 5 distinct models with different random initial weights and number of units and ensemble them to produce the final output. Each model is trained for 2000 epochs. We can learn on a set of 3000-4000 molecules in one day on one core of a modern PC. The final systems can predict millions of molecules per day on a small cluster of machines, making them particularly suitable for high throughput screening. Classification models estimate the probability of the endpoint given the inputs. This is also advantageous for screening since strict criteria can be imposed on the probability in order to increase the confidence of a prediction.

3 Results

For all the regression problems we only report the results on the models where stationarity is relaxed. In the mutagencity classification results we show that this type of model significantly outperforms the stationary model without the dedicated processing units. We also test other slight variations to the model architecture when predicting mutagenicity (see below).

3.1 Regression

For all the regression tasks described below three measures are reported. The squared correlation coefficient (r^2, or squared Pearson correlation coefficient) where the correlation coefficient is:

$$r = \frac{\sum_{i=1}^{N} t_i p_i - N\hat{t}\hat{p}}{(N-1)s_t s_p} \tag{5}$$

where $\hat{t}$ and $\hat{p}$ are the mean of the targets and predicted values respectively, N is the total number of examples and s_t and s_p are the standard deviations of the target and prediction respectively. We also report the root mean squared error (RMSE) and the average absolute error (AAE) which are $\frac{1}{N}\sqrt{\sum_{i=1}^{N}(t_i - p_i)(t_i - p_i)}$ and $|\frac{1}{N}\sum_{i=1}^{N}(t_i - p_i)|$ respectively.

Aqueous Solubility. The biological activity of potential drugs is influenced by their aqueous solubility - effectiveness of the drug may depend on this property for delivery. Hence computational methods to evaluate solubility of compounds are important, and many approaches to tackle this task have being described. A review of computational methods for the early phase of drug development can be found in [28]. In [29] the octanol-water partition coefficient (which can be predicted somewhat accurately from the molecular structure, see section 3.1) and 51 2D descriptors are input to a multiple linear regression model, yielding a squared correlation coefficient of 0.74 and an average absolute error of 0.68 on a dataset of 2688 training compounds and 640 test compounds. Delaney [2] uses the water partition coefficient and three other parameters (molecular weight, number of rotatable bonds and the aromatic proportion in aromatic rings) as inputs to a simple linear model. Although the model is simple it outperforms

Table 1. Prediction performance for Aqueous Solubility in 10 fold cross validation on the 1144 compounds in the Delaney "Small" dataset

	r^2	RMSE	AAE
UG-RNN	0.914	0.613	0.437
Delaney [2]	-	-	0.75
GSE [30]	-	-	0.47
2D kernel (param d=2) [12]	0.91	0.61	0.44

the General Solubility Equation [30] which is based on the melting point and octanol-water partition coefficient. In [12] various kernels are designed showing state of the art results on the "Small" Delaney dataset for the 2D kernel based on path lengths of two.

Table 1 shows our results obtained in 10 fold cross validation on the Delaney "Small" dataset. Comparisons are made with the kernel method in [12], Delaney's own method [2] and the GSE equation [30]. The dataset contains 1144 compounds of ranging types with solubility values (measured as a $\log S$) ranging from -11.6 to 1.58 Mol/litre.

Our results on this dataset are state of the art with a r^2 of 0.91, 0.61 RMSE and an AAE of 0.44 which are identical to the best kernel from the work in [12]. The only number available for comparison with Delaney's work and the GSE equation is an AAE of 0.75 and 0.47 on this "Small" dataset.

Another common dataset is the one in Huuskonen [3], consisting of 1297 compounds of ranging $\log S$ values from -11.62 to 1.58 Mol/Litre. Huuskonen's method relies on Molecular connectivity, shape, and atom-type electrotopological indices, input to a multi layer neural network, yielding a r^2 of 0.92 and standard deviation of 0.6. However no cross validation is performed in order to assess the true generalisation ability of the method. In [4] a support vector machine is used to learn from derived descriptors and the final model achieves a r^2 value of 0.90 in 8-fold cross validation on the same set. Again the kernel methods of [12] produce state of the art performances. However the best results are now on a different (3D) kernel. Our method remains the same, indicating general applicability. Although the Huuskonen dataset consists of 1297 compounds, Azencott et al. report results on 1026 compounds without mention of redundancy reduction. Table 2 shows a comparison of the methods on the Huuskonen dataset. On this set we achieve a r^2 of 0.92, slightly above the squared correlation for the kernel method in Azencott et al. and the method in Frohlich et. al., and a somewhat worse, if still nearly perfect, AAE (0.43 on a range of 12 log Mol/Litre units).

In figure 3 we show the correlation graphs for the Delaney dataset. We observe a very similar trend on the Huuskonen set.

Table 2. Prediction performance for Aqueous Solubility in 10 fold cross validation on the 1297 compounds in the Huuskonen dataset

	r^2	RMSE	AAE
UG-RNN	0.92	0.35	0.43
Frohlich [4]	0.90	-	-
3D kernel [12]	0.91	0.15	0.11

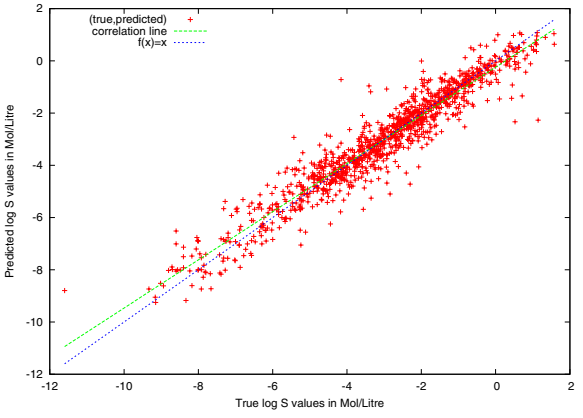

Fig. 3. Plot of Experimental Log S values against Predicted Log S values for the Delaney Small dataset. S is measured in Mol/Litre.

Melting point. Melting point can be used for the rapid determination of the purity of a substance and it is often a core property in QSAR/QSPR analysis for determining solubility and boiling point [31, 30]. The General solubility equation [30] is $\log S = 0.5 - \log P_{ow} - 0.01(t_m - 25)$ where $\log P_{ow}$ is the octanol water partition coefficient and t_m is the melting point. $\log P_{ow}$ can be predicted with high accuracy (see next section) while the automatic prediction of melting points still remains difficult. The above equation generally works well (RMSE 0.7-0.8 log units) so long as the melting point can be determined accurately.

We test our method on a melting point dataset extracted from the literature [5], containing 4173 compounds from the Molecular Diversity Preservation International database (MDPI) [32]. The melting points range between $14°C$ to $392.5°C$.

Our results (table 3) on this set compare favourably with two other methods [5, 12], with a correlation coefficient of 0.753 (r^2 of 0.57). Karthikeyan [5] uses a large set of 2D and 3D features which are dimensionally reduced using PCA, then input to a feed forward neural network. The model producing the best results in Azencott et al. [12] is again the 2D kernel using a minmax similarity measure but the path length used to determine the similarity is now 10.

Table 3. Prediction performance for Melting point in 10 fold cross validation on the 4173 compounds in the Karthikeyan dataset. Other methods based on the same dataset.

	r^2	RMSE	AAE
UG-RNN	0.57	$42.5°C$	$32.6°C$
Karthikeyan [5]	0.42	$52.0°C$	$41.3°C$
2D kernel (param d=10) [12]	0.56	$42.71°C$	$32.58°C$

Octanol water partition coefficient. Accurate determination of octanol water partition coefficient, i.e. the ratio of the concentrations of a compound in a mixture of water and octanol at equilibrium (normally measured as a logarithm of the ratio, or $\log P_{ow}$), is central to QSAR/QSPR. The magnitude of $\log P_{ow}$ is useful in estimating the distribution of drugs within the body. It is therefore an important factor of a candidate drug. Moreover, $\log P_{ow}$ and melting point can be used to accurately determine the solubility of a compound.

Table 4 shows the results for the prediction $\log P_{ow}$ on the dataset used in [6] and in [12]. The results in [6] are nominally accurate but are measured on the training set, hence meaningless and not reported in the table. The 2D kernel (the best performing is again different, this time with $d = 5$) is marginally more accurate. Our method is unchanged from the previous tests.

Table 4. Prediction performance for octanol/water partition coefficient (as $logP_{ow}$) by 10 fold cross validation on the dataset in [6, 12]. Other methods based on the same dataset.

	r^2	RMSE	AAE
This work	0.934	0.279	0.394
2D kernel (param d=5) [12]	0.94	0.25	0.38

3.2 Classification

Finally, we apply UG-RNN to the problem of mutagenicity classification. In this case we also gauge whether contextual information, and relaxing stationarity have an effect on the results. We also test two different variants of the output network $\mathcal{M}^{(O)}()$ (in Eqn. 2), one in which only the global state of the structure $\mathbf{G}_{structure}$ is passed as an argument, and one in which the average of the labels i_v over the molecule is also input. We term the last two models, respectively, Moore, and Mealy UG-RNN.

If TP, FP, TN, and FN are the true positives, false positives, true negatives, and false negatives respectively, the performance measures we use are precision $\frac{TP}{TP+FP}$ and recall $\frac{TP}{TP+FN}$ for the classes. Matthews Correlation Coefficient (MCC) is also computed for the sake of comparison with other works that report recall and MCC and leave out precision. MCC is defined as $\frac{(TP.TN)-(FN.FP)}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}}$.

Mutagenicity. Mutagenicity is the ability of a compound to cause mutations in DNA with these mutations often causing the onset of cancerous tumors. It has being previously shown that a positive experimental mutagenicity test (known as the Ames test) results in carcinogenicity as high as 77% to 90% in rodents [33]. Screening of drug candidates for mutagenicity is a regulatory requirement for drug approval since mutagenic properties pose risks to humans. At present, a variety of toxicological tests have to be conducted by a drug manufacturer. Although mutagencity testing has a relatively simple experimental procedure, these tests are generally low-throughput and hence cannot be applied to large scale screening. A fast effective method for the prediction of mutagenicity could therefore serve as an initial estimate of carcinogenicity and greatly aid in the manufacture of novel drugs.

We use a dataset collected from the literature [7]. The set consists of 4337 diverse chemical compounds along with information indicating whether they have mutagenicity in Salmonella Typhimurium strains TA98, TA100, TA1535 and either TA1537 or TA97, which are the standard Ames test strains required for regulatory evaluation of drug approval. In this dataset a compound is considered a mutagen if at least one Ames test result was positive, negative otherwise. This results in 54% of the data to be mutagenic making it a well balanced set.

In table 5 we compare UG-RNN with the substructure/structural alert method of [8] and a method based on a novel molecular electrophilicity descriptor input to an SVM [25]. The dataset used in the substructure mining method is a slightly reduced version of the dataset used here since mixtures, counter ions, molecules above 500 moleculular weight and stereoisomers were removed. We have not removed these noisy examples, which is expected to yield slightly worse results. The automatic substructure mining method [8] known as the elaborate chemical representation (ECR) is an extension of a similar manual method in [7], and simply assigns a chemical to a particular class based on substructures previously identified. The method in [25] constructs a novel feature vector based on the atomic electrophilicty which is a highly domain specific vector for mutagenicity prediction. MOLFEA [34] generates molecular fragments by a mining algorithm then uses three types of machine learning systems: decision trees, rule learner and support vector machines. The authors prove that the substructure based approach improves over machine learning with fixed length molecular property vectors, and obtain the highest predictive accuracy by optimised structures and a linear SVM.

We test four distinct models:

– *Multi Layer Perceptron (MLP)* The input is defined as:

$$I = \frac{1}{N} \sum_{v \in \{V\}}^{N} i_v \tag{6}$$

where N is the number of atoms, i_v is the input label at atom v and $\{V\}$ is the set of all vertices.

– *Stationary UG-RNN* A UG-RNN with only one transition function and therefore only one transition network to encode the molecule.

– *Moore UG-RNN* A UG-RNN with specialised transition functions for the most frequent patterns, and output obtained as $o = \mathcal{M}^{(O)}(\mathbf{G}_{structure})$.

– *Mealy UG-RNN* A UG-RNN with specialised transition functions for the most frequent patterns, but output obtained as $o = \mathcal{M}^{(O)}(I, \mathbf{G}_{structure})$, where I is the average label over the molecule, as in Eqn.6.

Table 5 shows the 10-fold cross validation results of all the models we tested, compared with those of the other methods described above (which are all also tested in 10-fold cross validation, although MOLFEA is tested on a different set).

It is clear from the table that the SVM+electrophilicity method [25] performs best, however: the feature vector is highly task-dependent; there is no evidence of test vs. validation separation in the choice of the feature vector; no precision values are reported. Our two best methods (Mealy UG-RNN and Moore UG-RNN) perform better than both

Table 5. Performance of the different mutagenicity models in a 10-fold cross validation. (P) is precision and (R) is recall. (*) different dataset of 684 compounds used for training and evaluation.

	Q all	mutagen (R)	non-mutagen (R)	mutagen (P)	non-mutagen (P)	mcc
MLP	77.8%	79.8%	75.2%	80.0%	75.0%	55.0%
Stationary UG-RNN	78.8%	75.9%	81.1%	76.5%	80.6%	57.1%
Moore UG-RNN	81.7%	84.7%	77.9%	82.5%	80.5%	62.7%
Mealy UG-RNN	82.4%	86.0%	78.0%	82.8%	81.8%	64.3%
SVM+electrophilicity [25]	90.1%	87.7%	92.1%	–	–	80.0%
Substructure mining [8]	80.6%	83.0%	74.5%	80.8%	77.2%	57.4%
MOLFEA [34]*	78.5%	77.5%	79.4%	–	–	56.9%

Substructure Mining [8] and MOLFEA [34]. Moreover, all UG-RNN's perform better than the static MLP, thus contextual information appears to be incorporated effectively. The non-stationary models perform better than the stationary one, and the Mealy model (in which both contextual information and average input labels are input to the output network) is the best performing of all with an overall accuracy of 82.4% and an MCC of 64.33%.

It should also be noted that the average intra-laboratory reproducibility of a series of Ames test data from the National Toxicological Program (NTP) was determined to be 85% [35], hence it is unclear what an accuracy of 90% [25] might mean and our best result is close to the experimental accuracy of the test.

4 Conclusions

We have developed a general class of machine learning models (UG-RNN) that map graphs to global labels, and have tested it in four separate problems in QSAR/QSPR. The models discussed in this work have remained essentially the same throughout all of these tasks. In all cases we obtained results close or above the state of the art, which depending on the task is represented by algorithms based on kernels, substructure mining/structural alert and manual or semi-automatic feature extraction, algorithms which are usually domain- and task-specific. The input features we have used are not domain-specific, are very simple and may be expanded, and the method is highly portable to other tasks in QSAR/QSPR, in molecular biology and elsewhere, so long as the input instances are naturally represented as Undirected Graphs.

In the future we plan to expand our research in a number of directions, including: testing whether the feature vector automatically generated by UG-RNN (the $\mathbf{G}_{structure}$ in Eqn.2), which is effectively a task-dependent encoding of the molecule, could be used as input to classifiers other than MLP, for instance SVM; whether UG-RNN can be expanded to include 3D information, alternatively or alongside 2D, for instance by representing the interaction between two atoms closer than a certain threshold as an edge, and introducing the distance as a label on the edges of the model; incorporating

information about stereoisomers in the model - this is currently overlooked and is likely to have hindered our results in the case of mutagenicity classification, where stereoisomers are present in the set.

Acknowledgments

This work is supported by grant RP/2005/219 from the Health Research Board of Ireland.

References

1. Hansch, C., Muir, R.M., Fujita, T., Maloney, P., Geiger, E., Streich, M.: The correlation of biological activity of plant growth regulators and chloromycetin derivatives with hammett constants and partition coefficients. J. Am. Chem. Soc. 85, 2817 (1963)
2. Delaney, J.: Esol: Estimating aqueous solubility directly from molecular structure. J. Chem. Inf. Comput. Sci. 44(3), 1000–1005 (2004)
3. Huuskonen, J.: Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. J. Chem. Inf. Comput. Sci. 40(3), 773–777 (2000)
4. Fröhlich, H., Wegner, J., Zell, A.: Towards optimal descriptor subset selection with support vector machines in classification and regression. J. Chem. Inf. Comput. Sci. 45(3), 581–590 (2005)
5. Karthikeyan, M.: General melting point prediction based on a diverse compound data set and artificial neural networks. J. Chem. Inf. Comput. Sci. 45(3), 581–590 (2005)
6. Wang, R., Fu, Y., Lai, L.: Towards optimal descriptor subset selection with support vector machines in classification and regression. J. Chem. Inf. Comput. Sci. 37(3), 615–621 (1997)
7. Kazius, J., McGuire, R., Bursi, R.: Derivation and validation of toxicophores for mutagenicity prediction. J. Med. Chem. 48(1), 312–320 (2005)
8. Kazius, J., Nijssen, S., Kok, J., Bäck, T., Ijzerman, A.: Substructure mining using elaborate chemical representation. J. Chem. Inf. Model. 46(2), 597–605 (2006)
9. Deshpande, M., Kuramochi, M., Wale, N., Karypis, G.: Frequent substructure-based approaches for classifying chemical compounds. IEEE Transactions on Knowledge and Data Engineering 17(8), 1036–1050 (2005)
10. Benigni, R., Giuliani, A.: Putting the predictive toxicology challenge into perspective: reflections on the results. Bioinformatics 19(10), 1194–1200 (2003)
11. Mahé, P., Ueda, N., Akutsu, T., Perret, J., Vert, J.: Graph kernels for molecular structure-activity relationship analysis with support vector machines. Journal of Chemical Information and Modeling 45, 939–951 (2005)
12. Azencott, C., Ksikes, A., Swamidass, A., Chen, J., Ralaivola, L., Baldi, P.: One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties. J. Chem. Inf. Comput. Sci. 47(3), 965–974 (2007)
13. Ceroni, A., Costa, F., Frasconi, P.: Classification of small molecules by two- and three-dimensional decomposition kernels. Bioinformatics 23(16), 2038–2045 (2007)
14. Swamidass, S., Chen, J., Bruand, J., Phung, P., Ralaivola, L., Baldi, P.: Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. Bioinformatics 21(suppl. 1), 359–368 (2005)
15. Micheli, A., Sperduti, A., Starita, A.: An introduction to recursive neural networks and kernel methods for cheminformatics. Current Pharmaceutical Design 13(14), 1469–1495 (2007)

16. Sperduti, A., Starita, A.: Supervised neural networks for the classification of structures. IEEE Transactions on Neural Networks 8(3), 714–735 (1997)
17. Frasconi, P.: An introduction to learning structured information. J. Chem. Inf. Comput. Sci. 1387/1998, 99 (2004)
18. Frasconi, P., Gori, M., Sperduti, A.: A general framework for adaptive processing of data structures. IEEE Transactions on Neural Networks 9(5), 768–786 (1998)
19. Bernazzani, L., Duce, C., Micheli, A., Mollica, V., Sperduti, A., Starita, A., Tiné, M.: Predicting physical-chemical properties of compounds from molecular structures by recursive neural networks. Applied Intelligence 19(1-2), 9–25 (2003)
20. Micheli, A., Portera, F., Sperduti, A.: QSAR/QSPR studies by kernel machines, recursive neural networks and their integration. In: Apolloni, B., Marinaro, M., Tagliaferri, R. (eds.) WIRN 2003. LNCS, vol. 2859, pp. 308–315. Springer, Heidelberg (2003)
21. Bianucci, A., Micheli, A., Sperduti, A., Starita, A.: Application of cascade correlation networks for structures to chemistry. Applied Intelligence 12(1-2), 117–147 (2000)
22. Siu-Yeung, C., Zheru, C.: Genetic evolution processing of data structures for image classification. IEEE Transactions on Knowledge and Data Engineering 17(2), 216–231 (2005)
23. Costa, F., Frasconi, P., Lombardo, V., Soda, G.: Towards incremental parsing of natural language using recursive neural networks. Applied Intelligence 19(1-2), 9–25 (2003)
24. Bianchini, M., Maggini, M., Sarti, L., Scarselli, F.: Recursive neural networks learn to localize faces. Pattern Recognition Letters 26(12), 1885–1895 (2005)
25. Zheng, M., Liu, Z., Xue, C., Zhu, W., Chen, K., Luo, X., Jiang, H.: Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. Bioinformatics 22(17), 2099–2106 (2006)
26. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5(2), 157–166 (1994)
27. The open babel package version 2.1.1, `http://www.openbabel.org/`
28. Huuskonen, J.: Estimation of aqueous solubility in drug design. Combinatorial Chemistry and High Throughput Screening 4(3), 311–316 (2000)
29. Butina, D., Gola, J.: Modeling aqueous solubility. J. Chem. Inf. Comput. Sci. 43, 837–841 (2003)
30. Jain, N., Yalkowsky, S.: Estimation of the aqueous solubility i: Application to organic nonelectrolytes. Journal of Pharmaceutical Sciences 90(2), 234–252 (2001)
31. Abramowitz, R., Yalkowsky, S.: Melting point, boiling point, and symmetry. Pharmaceutical Research 7(9), 942–947 (1990)
32. Molecular diversity preservation international database, `http://www.mdpi.org/`
33. Mortelmans, K., Zeiger, E.: The ames salmonella/microsome mutagenicity assay. Mutat. Res. 455(1-2), 29–60 (2000)
34. Helma, C., Cramer, T., Kramer, S., De Raedt, L.: Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. J. Chem. Inf. Comput. Sci. 44(4), 1402–1411 (2004)
35. Piegorsch, W., Zeiger, E.: Measuring intra-assay agreement for the ames salmonella assay. Statistical Methods in Toxicology. Lect. Notes Med. Informatics 43, 35–41 (1991)