Adaptive Virtual Screening of Drug-Like Molecules by Recursive Neural Networks for Undirected Graphs

Alessandro Lusci, Ian Walsh, and Gianluca Pollastri

***Abstract*—Virtual screening of drug-like molecules is a very important task in drug design, but it is generally time consuming and expensive. In fact, the common practice of virtual screening requires a feature selection process that is performed by experts. The time necessary to accomplish it is long and often unpredictable. Moreover, given that even experts can fail, this hand-crafted selection process can lead to the loss of molecular properties that may be essential to predict the desired target point. Therefore, providing a time efficient and cost-free virtual screening method for predicting molecular endpoints may be a key step in the development of new drug design approaches. Here we present a screening method based on a neural network model called Recursive Neural Networks for Undirected Graphs (UGRNN), where the feature selection process requires no human intervention. We test the method on a broad range of regression tasks (mostly concerning prediction of solubility and melting point for small molecules). The results we obtain generally match or surpass the state of the art.**

***Index Terms*—drug design, virtual screening, recursive neural network, undirected graph, regression, molecular endpoint.**

I. Introduction

OVER the last few decades numerous methods have been developed to perform virtual screening of chemical compounds. Most of these methods belong to the broad category of QSAR (Quantitative Structure-Activity Relationship). The aim of QSAR is to find an appropriate function $F()$, which, given a structured representation of a molecule, predicts its biological activity [1]. QSAR's most general form is:

$$Activity = F(structure) \tag{1}$$

The definition of function $F()$ is a complex task which can be factorized into two sub-problems: the *encoding problem* and the *mapping problem*. The former refers to the task of mapping a molecule, which is naturally described as an undirected graph representing its chemical structure, into an array of features. This step is necessary in order to obtain a representation which is suitable for standard regression/classification tools like Artificial Neural Networks (ANN) or Support Vector Machines (SVM). The latter consists in mapping the array of features into the property of interest and, as mentioned, is generally a regression or a classification task, which may be tackled by one of numerous algorithmic tools that are available. According to this view, $F()$ can be decomposed as follows [1]:

$$F() = g(t()) \tag{2}$$

where $t()$ is the encoding function and $g()$ is the mapping function. The way $t()$ is defined is rather open-ended and ultimately one could argue that the essence of the problem is precisely finding $t()$ or, equivalently, that once an informative $t()$ is found for the problem at hand, the following step is trivial. In most cases $t()$ is hand-crafted and requires the intervention of experts. If this is the case, finding $t()$ is usually time consuming and, given that even experts may fail, or overlook, may lead to the loss of important information to predict the desired target. In [2], [3] and [4] a similar approach is followed to predict acqueous solubility by a Multi Layer Perceptron (MLP) and SVM, respectively. In [5] a large set of molecular features, including physical and graphical properties, is compressed by Principal Component Analysis (PCA) to be the input to an ANN, with the aim of predicting melting points. In [6] numeric codes for alkanes are applied to provide an input for an MLP in order to predict melting points and in [7] a set of 2D and 3D molecular descriptors for each molecule is calculated, to predict melting points using a method based on partial least squares Projection to Latent Structures (PLS).

Among all current state-of-the-art automated methods (i.e., where the function $t()$ is defined by a fully automated computational process), one of special interest is represented by N-Dimensional Kernels as described in [8]. In particular, when the number of examples in the training set is large enough (greater than 1000), 2D spectral kernels proved to yield robust results, generally better than 3D kernels.

Here we compare the performances of a different automated method which we have developed, Recursive Neural Network for Undirected Graphs (UGRNN)[9], against state-of-the-art 2D kernels. In order to do so we select several benchmarks on which 2D kernels were assessed, and test the UGRNN model on them in the same experimental setting. We implement a simple stationary UGRNN model (stationary in that one single network/function processes all parts of the molecular structure) where the input is limited to the atom label and bond type. No other descriptors or hand-crafted features are input to the model, which therefore needs to find its own encoding and feature set without any human expert intervention. In brief, the feature selection process is fully automated in our approach.

In the tests we present in this article, simple stationary

A.Lusci and G.Pollastri are with the School of Computer Science and Informatics and Complex and Adaptive Systems Laboratory, University College Dublin, Belfield, Dublin 4, Ireland.

I.Walsh is with the Department of Biology, University of Padua, Padova 35131, Italy.

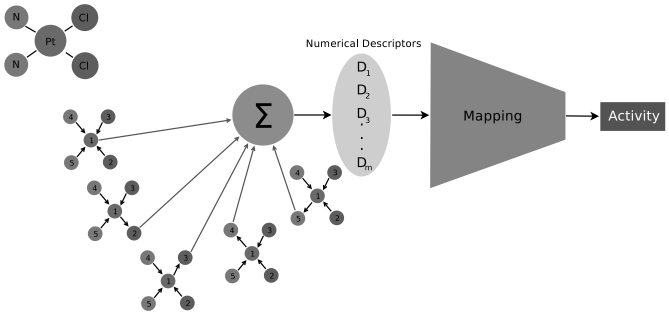

Fig. 1. UGRNN model of Cisplatin molecule

UGRNN models outperform or match state-of-the-art 2D kernels. According to the protocol we followed, a stationary UGRNN generally matches and in some cases outperforms a more complex non-stationary model (see[9]) which has in addition a more informative atom label. From our tests we observe that the stationary UGRNN is able to extract, from the graph strucuture defining a chemical compound, the molecular feautures which are necessary to predict a given target point. We are currently researching whether these features, which are a sub-product of training and come as a fixed-width array or real numbers, may be informative descriptors for the molecular space as a whole, and may be used to map this space and compute similarity between molecules.

II. Approach

The function $F()$ mapping a molecule onto a property can be factorized into two functions: the encoding function $t()$ and the mapping function $g()$. $t()$ is a function from the domain of the molecular structure S to the domain of molecular descriptors D:

$$t : M \rightarrow D, D \in \mathbb{R}^m, \tag{3}$$

$g()$ is a function from the domain of the molecular descriptors D to the domain of the target properties T:

$$g : D \rightarrow T, D \in \mathbb{R}^m, T \in \mathbb{R}^n. \tag{4}$$

In order to create an automated algorithm that can extract the molecular descriptors from the graph describing the molecule, we choose to approximate function $t()$ by a Feed Forward Neural Network (FFNN). In order to train the network by gradient descent through the backpropagation algorithm, it is necessary to represent the input domain as a Directed Acyclic Graph (DAG)[10], [11]. However a molecule is naturally described as an Undirected Graph (UG), where atoms represent nodes and bonds represent edges (as in the Cisplatin molecule in the top-left corner of Figure 1). We solve the problem by factorising the UG into as many DAGs as the number of atoms in the molecule, and retrieve contextual information from the root node of each DAG to obtain a single vector of molecular descriptors, as represented in Figure 1. In the following section we describe UGRNN in detail.

III. Method

The process for generating the UGRNN model of a molecule is here described. First we factorise a UG into as many Directed Acyclic Graphs (DAGs) as there are atoms in the molecule, so that each node/atom is the root of one DAG. The k^{th} DAG is obtained from the UG representing the molecule, by directing all its edges in the graph towards its k^{th} atom v_k, along the shortest path.

At each node in the k^{th} DAG a state, or contextual vector, is stored, which is a function $M^{(G)}$ of the node label (atom identity in this case, but potentially anything else we know about an atom), and the states of its parents (the nodes upstream of the vertex, along the directed edges). More formally, the state of each vertex v in the k^{th} DAG is a hidden vector $G_{v,k}$ describing the contextual information upstream of vertex v as:

$$G_{v,k} = M^{(G)}(i_v, G_{pa^1_{[v,k]}}, ..., G_{pa^n_{[v,k]}}), \tag{5}$$

where $i_v \in \mathbb{R}^l$ is the label associated to vertex v (i.e., the input information about atom v) and $pa^1_{[v,k]}, ..., pa^n_{[v,k]}$ are the parents of vertex v in the k^{th} DAG. The function $M^{(G)}$ is the same for each DAG and each vertex, resulting in a stationarity hypothesis which can be regarded as a form of weight sharing to keep the number of free parameters in the model low. Notice how, in order for this representation to be possible, there must be an upper bound n to the number of parents a node can have (in the application presented in this work, typically $n = 4$). If a node has m parent nodes with $m < n$, blank vectors (all zeroes) are passed to function $M^{(G)}$ as its last $n - m$ arguments. If a node has no parents, only i_v is input to the function and all the inputs reserved for contextual information are blank.

As there is a path from each atom in a DAG to the root of the DAG, the state of the root receives a contribution from each of the nodes in the DAG. To obtain a representation for the whole graph, we add up the states of the root nodes of all DAGs. That is, the overall description of the molecule is obtained as the sum of descriptions of the molecule "as seen" from each of its nodes/atoms. More formally, $G_{structure}$ is defined as:

$$G_{structure} = \sum_{k=1}^{N} G_{v_k,k} \tag{6}$$

$G_{structure}$ is a global description of the molecule (a feature vector), which we map into the property of interest via a mapping function $M^{(O)}$, as follows:

$$o = M^{(O)}(G_{structure}) \tag{7}$$

It is important to notice that: both the encoding function $M^{(G)}$ and the mapping function $M^{(O)}$ are approximated by FFNN so that the whole model can be trained by gradient descent, because it is itself an FFNN; the process of extraction of the molecular descriptors from the molecular structure depends on the minimization of the error between predicted and target values (a sum of squares for regression tasks, a relative entropy for classification), resulting in an adaptative form of compression of the molecular structure, *property driven and fully automated*. That is, if training is successful, the $G_{structure}$ vector will be the description of the molecule that yields the best prediction according to the errors adopted.

TABLE I
LEAVE-ONE-OUT SQUARED CORRELATION COEFFICIENT RESULTS FOR THE ALKANES AND BENZODIAZEPINES (BZD) AND BERGSTRÖM DATASETS

	BZD	alkanes	Bergström
UGRNN	0.52	**0.98**	**0.40**
2D Kernels[8]	**0.69**	0.94	0.36

While in preliminary tests in [9] we implemented a model where the stationarity hypotesis is relaxed, meaning that dedicated transition functions $M^{(G)}$ are associated to the most common bonding patterns for an atom, here we decide not to relax the stationarity constraint with the aim of avoiding lack of generality and measuring the baseline performances of a UGRNN.

In [9] each atomic label i_v includes: the element type, the atom charge, the Smallest Set of Smallest Rings, hybridisation state and aromaticity for an atom. Here we limit our input to atom and bond type, in order to discover what the model can learn from the molecular structure itself and to obtain a more reliable comparison with the perfomances of 2D Kernels described in [8].

Since the model presented in this work can be considered as a Feed Forward Neural Network with weight sharing, we train the UGRNN by gradient descent via the backpropagation algorithm. In particular, given the large amount of weight sharing, we need to modify gradient descent similarly to [10], [11]. Thus, the gradient of the error wrt a weight dw is applied if $|dw| \in$ [0.1,1], otherwise it is set to $sign(dw)$ if $|dw| > 1$ or to $0.1 * sign(dw)$ if $|dw| < 0.1$.

IV. RESULTS

A. Small Datasets

Benzodiazepines QSAR. The dataset[12] consists of 72 1,4-benzodiazepine-2-ones. The target point for each molecule is represented by its measured affinity toward the γ-aminobutyric acid. In this dataset we obtain a Leave One Out Pearson's squared correlation coefficient of 0.52 which is lower than the one declared with 2D Kernels (0.69). However if we split the dataset in training and test set as in [1] we obtain a squared correlation coefficient of 0.98 which is the same value obtained with 2D Kernels in [8]. This vast difference in performances puts into question the stability of the results obtained by other authors and in this work, as deviations of over 40% are observed simply by splitting the dataset in different training and testing subsets.

Alkanes Boiling Point. The dataset [6] consists of the first 150 noncyclic alkanes (C_nH_{2n+2} with $n < 11$). The target point for each molecule is representend by its boiling point. In this dataset we obtain a Leave One Out Pearson's squared correlation coefficient of 0.98 which is higher than the one obtained with 2D Kernels (0.94).

Bergström. The dataset[7] consists of 277 druglike compounds. The target point for each molecule is representend by its melting point. In this dataset we obtain a Leave One Out Pearson's squared correlation coefficient of 0.41 which is higher than the one obtained with 2D Kernels (0.36).

TABLE II
PREDICTION PERFORMANCE FOR ACQUEOUS SOLUBILITY IN 10 FOLD CROSS VALIDATION ON DELANEY DATASET

	r^2	RMSE	AAE
UGRNN	**0.92**	**0.59**	**0.43**
Non Stationary UGRNN[9]	0.91	0.61	0.44
Delaney[2]	-	-	0.75
GSE[13]	-	-	0.47
2D Kernel (param d=2)[8]	0.91	0.61	0.44

TABLE III
PREDICTION PERFORMANCE FOR ACQUEOUS SOLUBILITY IN 10 FOLD CROSS VALIDATION ON HUUSKONEN DATASET

	r^2	RMSE	AAE
UGRNN	**0.93**	0.55	0.35
Non Stationary UGRNN[9]	0.91	0.43	0.35
Frolich[14]	0.90	-	-
2D Kernel (param d=2)[8]	0.91	**0.15**	**0.11**

B. Large Datatasets

Aqueous Solubility (Delaney). The dataset[2] consists of 1144 low molecular weight compounds. The target point for each molecule is representend by its acqueous solubility in $logM/L$. Results in table 2 show that our model outperforms by all metrics both 2D kernels and the the non stationary UGRNN model with a more complex atom label.

Aqueous Solubility (Huuskonen). The dataset[3] consists of 1026 compounds. The target for each molecule is represented by its acqueous solubility. Table 3 shows that our model outperforms in correlation both 2D kernels and the non stationary UGRNN with a more complex atom label, while the results are slightly lower (albeit nearly perfect) by the two other metrics.

Melting Point (Karthikeyan). The dataset consists of 4173 compounds annotated with melting points and a wide range of additional properties. In our tests we limit the molecular target to the melting point. Results in table 4 show that UGRNN, non stationary UGRNN[9] and 2D kernels[8] achieve nearly identical results, with the best performance by a slight margin being achieved by non stationary UGRNN.

V. CONCLUSIONS

In this article we have presented a broad assessment of UGRNN, a model of recursive neural network we have developed, which is capable of dealing with undirected graphs, on a set of benchmarks composed of chemical compounds. No feature set has to be designed for UGRNN, as the model is capable of obtaining a fixed-width representation of a molecule ($G_{structure}$) by design, without human intervention in the process. Because of this, prediction of properties or activities

TABLE IV
PREDICTION PERFORMANCE FOR MELTING POINT USING 10-FOLD CROSS-VALIDATION ON THE 4173 COMPOUNDS IN THE KARTHIKEYAN DATASET

	r^2	RMSE	AAE
UGRNN	0.56	42.6	33.2
Non Stationary UGRNN[9]	**0.57**	**42.5**	**32.6**
Karthikeyan[5]	0.42	52.0	41.3
2D Kernel (param d=10)[8]	0.56	42.7	32.6

of chemical compounds may be automated in full, as the only requirement to build a predictor is the availability of a set of examples for which the property of interest has been determined experimentally. This eliminates the need for expert knowledge and a potentially time-consuming and fault-prone initial stage in which features that are informative for the task at hand have to be indentified. In our tests UGRNN generally match or outperform state-of-the-art models which operate on the 2D representation of a molecule, including systems in which more complex, and theoretically more informative, atom labels are devised. It should also be noted that UGRNN may easily accommodate more informative descriptors, both at the atom level (by simply extending the input corresponding to the label of an atom), and at a molecular level (by providing supplementary information alongside the $G_{structure}$ as inputs to the output network), i.e. although UGRNN are capable of working with only minimal human intervention, if expert knowledge is available it may be included into the prediction process.

We are currently extending our research in three concurrent directions: testing UGRNN with more informative input descriptors; evaluating the ability of $G_{structure}$ to describe the space of molecules; including 3D information by an extension of UGRNN. The second direction is of particular interest, as UGRNN provide a fixed-width representation of a molecule which may be used to gauge similarity between compounds without resorting to complex graph matching techniques.

VI. Funding

AL is funded through a GREP Ph.D. Scholarship by the Irish Research Council for Science, Engineering and Technology. GP's research is partly funded through Science Foundation Ireland RFP grant 10/RFP/GEN2749.

References

[1] A. Starita, A. Micheli, A. Sperduti. *Analysis of the Internal Representations Developed by Neural Networks for Structures Applied to Quantitative Structure-Activity Relationship Studies of Benzodiazepines*, J. Chem. Inf. Comput. Sci., 41, 202-218, 2000.

[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] Fróhlich 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), 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] Cherqaoui D., Villemin D., *Use of Neural Network to Determine the Boiling Point of Alkanes* J. Chem. Soc., Faraday Trans., 90, 97-102, 1994.

[7] Bergström C. Norinder U. Luthman K. Artursson, *Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs*, J. Chem. Inf. Model., 2003, 43, 1177-1185

[8] Chloé Agathe Anzecott, Alexandre Ksikes, S. Joshua Swamidass, Jonathan H. Chen., Liva Ralaivola and Pierre Baldi *One- to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties*, J. Chem. Inf. Model., 2007, 47, 965-974

[9] Ian Walsh, Alessandro Vullo and Gianluca Pollastri *Recursive Neural Networks for Undirected Graphs for learning molecular endpoints*, in Pattern Recognition for Bioinformatics, Lecture Notes in Computer Science, 5780/2009, 391-403, 2009.

[10] Gianluca Pollastri, Pierre Baldi *Prediction of Contact Maps by Recursive Neural Network Architectures and Hidden Context Propagation from All Four Cardinal Corners*, Bioinformatics, 18(S1), S62-S70, 2002.

[11] Pierre Baldi, Gianluca Pollastri *The principled Design of Large-Scale Recursive Neural Network Architectures-DAG-RNNs and the Protein Structure Prediction Problem*, Journal of Machine Learning Research, 4, 575-602, 2003.

[12] Hadjipavlou-Litina D. Hansch C. *Quantitative Structure-Activity Relationship of the Benzodiazepines. A Review and Reevaluation.* Chem. ReV., 94, 1483-1505, 1994.

[13] Jain N. and Yalkowsky S., *Estimation of the acqueous solubility: Application to organic non-electrolytes*, Journal of pharmaceutical science, 90:311-316, 2001.

[14] Fröhlich H. Wegner J. K. Zell A., *Towards Optimal Descriptor Subset Selection with Support Vector Machines in Classification and Regression*, QSAR Comb. Sci., 23, 313-318, 2004.

Alessandro Lusci received his M.Sc. degree from the Department of Electrical and Electronic Engineering, Cagliari University, Cagliari, Italy, in 2009. He is currently a Ph.D. student in the School of Computer Science and Informatics, University College Dublin, Dublin, Ireland. His current research interests include recursive neural networks and chemoinformatics.

Ian Walsh holds a B.Sc. and a Ph.D. in Computer Science from University College Dublin, Ireland. He is a postdoctoral fellow at the University of Padua, Italy, since 2010. He is interested in machine learning, protein bioinformatics and chemioinformatics.

Gianluca Pollastri holds an M.Sc. degree in Telecommunication Engineering from the University of Florence, Italy, and a Ph.D. in Computer Science from University of California at Irvine. He is a Principal Investigator at University College Dublin, Ireland, since 2003. He is interested in designing machine learning models for structured data, and in their application to bioinformatics and chemioinformatics.