A Graph Kernel incorporating molecule's stereisomerism information

Pierre-Anthony Grenier, Luc Brun, Didier Villemin

A Graph Kernel incorporating molecule's stereisomerism information

Pierre-Anthony Grenier, Luc Brun
GREYC UMR CNRS 6072
Caen, France
{pierre-anthony.grenier, luc.brun}@ensicaen.fr

Didier Villemin
LCMT UMR CNRS 6507,
Caen, France
didier.villemin@ensicaen.fr

***Abstract*—The prediction of molecule's properties through Quantitative Structure Activity (resp. Property) Relationships are two active research fields named QSAR and QSPR. Within these frameworks Graph kernels allow to combine a natural encoding of a molecule by a graph with classical statistical tools such as SVM or kernel ridge regression. Unfortunately some molecules encoded by a same graph and differing only by the three dimensional orientations of their atoms in space have different properties. Such molecules are called stereoisomers. These latter properties can not be predicted by usual graph methods which do not encode stereoisomerism. In this paper we propose a new graph encoding of molecules taking explicitly into account stereoisomerism and propose a new kernel between these structures in order to predict properties related to stereoisomerism.**

I. Introduction

Most of QSAR and QSPR methods are based on a basic principle of the chemoinformatics framework which states that: "two similar molecules should have similar properties". Prediction of molecular properties thus involves the design of a model encoding molecules and a similarity measure between such models. We here implicitly assume that similarity between models corresponds to a similarity between molecules. However different models may encode different amount of information about molecules hence leading to different degree of relevance of their associated similarity measures.

Molecules can be represented by their molecular formula (e.g. CH_4). However, as this representation does not encode the bond connections between atoms, different molecules, called structural isomers, can have a same molecular formula. An usual way to overcome this limitation consists in the use of molecular graphs. A molecular graph is a simple graph $G = (V, E, \mu, \nu)$, where each node $v \in V$ encodes an atom, each edge $e \in E$ a bond between two atoms and the labelling functions μ and ν associate to each vertex and each edge a label encoding respectively the nature of the atom (carbon, oxygen,...) and the type of the bond (single, double, triple or aromatic). Molecular graphs, allow to encode neighborhood relationships between atoms, and thus allow to differentiate structural isomers.

However, molecular graphs have also a limitation: they do not encode the spatial configuration of atoms. Indeed, some molecules, called stereoisomers, are associated to a same molecular graph but differ by the relative positioning of their atoms. Hence, molecular graphs do not allow us to distinguish two stereoisomers with different properties. Most of stereoisomers are characterized by the three dimensional orientation of the direct neighbors of a single atom or two connected atoms. We can imagine for example, a carbon atom, with four neighbors, each of them located on a summit of a tetrahedron. If we permute two of the atoms, we obtain a different spatial configuration (Figure 1). An atom is called a stereocenter if a permutation of two atoms belonging to its neighborhood produces a different stereoisomer. We should stress here that, to a large extend, stereoisomerism is independent of a particular embedding of a molecule. Indeed, in Figure 1, any particular embedding keeping the same relative positioning of atoms H, Cl, Br and F according to the central carbon atom C, would correspond to a same stereoisomer. In the same way, two connected atoms form a stereocenter if a permutation of the positions of two atoms belonging to the union of their neighborhoods produces a different stereoisomer (Figure 2). According to chemical experts, within molecules currently used in chemistry, 98% of stereocenters correspond either to carbons with four neighbors, called asymmetric carbons (Figure 1) or to couples of two carbons adjacent through a double bond (Figure 2). We thus restrict the present paper to such cases.

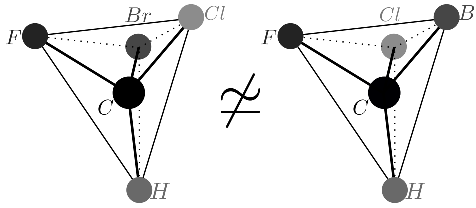


Fig. 1. Two different spatial configurations of the neighbors of a carbon

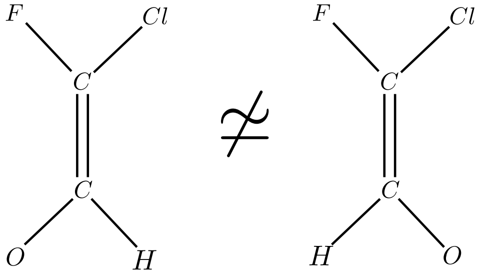


Fig. 2. Two different spatial configurations of two carbons linked by a double bond.

Graph kernels [1], [2], provide a measure of similarity between graphs. Under the assumption that a kernel k is symmetric and definite positive, the value $k(G, G')$, where G and G' encode two graphs, corresponds to a scalar product between two vectors $\Psi(G)$ and $\Psi(G')$ in an Hilbert space. This latter property allows us to combine graph kernels with usual machine learning methods such as SVM or kernel ridge regression by using the well known kernel trick, which consists in replacing the scalar product between $\Psi(G)$ and $\Psi(G')$ by $k(G, G')$ in these algorithms.

Up to now, only few methods have attempted to incorporate stereoisomerism within the graph kernel framework. Brown et al. [3] have proposed to incorporate this information through an extension of the tree-pattern kernel [1]. One drawback of this method is that, patterns which encode stereo information, and patterns which do not, are combined without any weighting in the final kernel value. So for a property only related to stereoisomerism, patterns that do not encode stereo information may be assimilated to noise. Intuitively, stereoisomerism property is related to the fact that permuting two neighbors of a given atom produces a different spatial configuration. Stereoisomerism property may be easily detected if all the direct neighbors of an atom have different labels (e.g. Figure 1). However, if two neighbors of a stereocenter have a same label, the influence of a permutation should be searched beyond the direct neighborhood of this stereocenter. Based on this ascertainment, Grenier et al. [4] have introduced the minimal subtree which characterizes a stereocenter within an acyclic molecule. They also proposed a kernel based on this minimal subtree which takes into account stereoisomerism. This kernel is however restricted to acyclic graphs.

Based on [4], we present in Section II an encoding of molecules distinguishing stereoisomers. Section III presents the construction of a subgraph, which allows to characterize locally a stereocenter. A molecule can then be described by the set of subgraphs describing each of its stereocenters. In Section IV, we propose a new graph kernel which compares the number of occurences of these subgraphs. This kernel is valid for cyclic as well as acyclic molecules and thus, overcomes the main limitation of [4]. We finally present in Section V results obtained using this new kernel and compare these results with state of the art methods.

II. Ordered Graph and Stereo Vertices

The spatial configuration of the neighbors of each atom may be encoded through an ordering of its neighborhood. For example, considering the left part of Figure 1, and looking at the central carbon from the hydrogen atom (H), the sequence of remaining neighbors of the carbon: Cl, Br and F may be considered as lying on a plane and are encountered clockwise. Thus, this spatial configuration is encoded by the sequence H, Cl, Br, F and the sequence H, Br, Cl, F encodes the second configuration. The configuration around a double bond can also be encoded by ordered sequences. Considering the left part of Figure 2 and assuming a clockwise orientation with the plane embedding provided by this figure, we encounter F and Cl when turning around the carbon at the top of the molecule, and H and O for the carbon at the bottom. Thus this configuration may be encoded by both sequences F, Cl and H, O respectively for the top and bottom carbon atoms. Sequences F, Cl and O, H encode the second configuration.

In order to encode this information in a graph, we introduce the notion of ordered graph. An ordered graph $G = (V, E, \mu, \nu, ord)$ is a molecular graph $G_m = (V, E, \mu, \nu)$ together with a function $ord : V \to V^*$ which maps each vertex to an ordered list of its neighbors. Two ordered graphs G and G' are isomorphic ($G \underset{o}{\simeq} G'$) if there exists an isomorphism f between their respective molecular graphs G_m and G'_m such that $ord'(f(v)) = (f(v_1) \dots f(v_n))$ with $ord(v) = (v_1 \dots v_n)$ (where $N(v) = \{v_1, \dots, v_n\}$ denotes the neighborhood of v).

However, different ordered graphs may encode a same molecule. In the example of the left part of Figure 1, if we look to the central carbon from a different neighbor, we can obtain a different sequence, for example F, Br, Cl, H, that represents the same configuration but now considered from the atom F. In the same way, considering the molecule on the left part of Figure 2 still with a clockwise orientation but now from the opposite side of its plane embedding we obtain the opposite sequences Cl, F and O, H. We thus have to define an equivalence relationship between ordered graphs, such that two ordered graphs are equivalent if they represent a same configuration.

To do so, we introduce the notion of re-ordering function σ, which associates to each vertex $v \in V$ of degree n a permutation $\sigma(v)$ on $\{1, \dots, n\}$, which allows to re-order its neighborhood. The graph with re-ordered neighborhoods $\sigma(G)$ is obtained by mapping for each vertex v its order $ord(v) = v_1 \dots . v_n$ onto the sequence $v_{\sigma(v)(1)} \dots . v_{\sigma(v)(n)}$ where $\sigma(v)$ is the permutation applied on v.

In order to define a permutation $\sigma(v)$ for each vertex of a graph, we first introduce the notion of potential asymmetric carbon which corresponds to a carbon with four neighbors. Such a vertex corresponds to a stereocenter if one permutation of two of its neighbors provides a different stereoisomer (Section I). Permutations associated to a potential asymmetric carbon correspond to all even permutations of its four neighbors [5]. We can easily check that the different orders obtained by these permutations, encode a same configuration but either seen from a different neighbor or with a same view point but with a different encoding of the cyclic order of the three remaining neighbors. For example, even permutations $(1, 4)(2, 3)$ and $(2, 3)(3, 4)$ applied on the order $H.Cl.Br.F$ of the central carbon in Figure 1 produce respectively the orders $F.Br.Cl.H$ and $H.Br.F.Cl$ which both encode the same configuration. For a double bond between two carbons, permutations associated to each carbon of the double bound must have a same parity. In the same way, we can check that these permutations correspond to different representations of a same configuration. Finally, for any vertex which does not correspond to a potential asymmetric carbon nor to a carbon of a double bond, we do not search to characterize its spatial configuration. So these vertices are associated to all possible permutations of their neighbors.

The set of re-ordering functions, transforming an ordered graph into another one representing a same configuration is called a valid family of re-ordering functions Σ [6]. Two ordered graphs G and G' are said to be equivalent according to Σ ($G \underset{\Sigma}{\simeq} G'$) if it exists a re-ordering function $\sigma \in \Sigma$ such that

$\sigma(G) \underset{o}{\simeq} G'$. This relationship defines an equivalence relationship [6] and two different stereoisomers are encoded by non equivalent ordered graphs. We denote by IsomEqOrd(G, G') the set of equivalent ordered isomorphism between G and G'.

Potentials asymmetric carbons, and double bonds between carbons, are not necessarily stereocenters. For example if the label of vertex Br of Figure 1 is replaced by Cl, both left and right molecules of Figure 1 would be identical. In the same way, if the label of the vertex F in Figure 2 is replaced by Cl, the left and right molecules of this figure would also become identical. For those cases, any permutation in the ordered list of the carbons would lead to an equivalent ordered graph. We thus define a stereo vertex as a vertex for which any permutation of two of its neighbors produces a non-equivalent ordered graph:

Definition 1 (Stereo vertex)**.** Let $G = (V, E, \mu, \nu, ord)$ be an ordered graph. A vertex $v \in V$ is called a stereo vertex iff:

$$\begin{aligned} \forall (i,j) \ &\in \{1, \ldots, |N(v)|\}^2, i \neq j, \\ &\nexists f \in \text{IsomEqOrd}(G, \tau^v_{i,j}(G)) \text{ with } f(v) = v. \quad (1) \end{aligned}$$

where $\tau^v_{i,j}(G)$ corresponds to an ordered graph deduced from G by permuting nodes of index i and j in $ord(v)$.

III. Minimal Stereo SubGraph

Definition 1 is based on the whole graph G to test if a vertex v is a stereo vertex. However, given a stereo vertex s, one can observe that on some configurations, the removal of some vertices far from s should not change its stereo property. In order to obtain a more local characterization of a stereo vertex, we should thus determine a vertex induced subgraph H of G, including s, big enough to characterize the stereo property of s, but sufficiently small to encode only the relevant information characterizing the stereo vertex s. Such a subgraph is called a minimal stereo subgraph of s.

We now present an heuristic, used to compute a minimal stereo subgraph of a stereo vertex. We first focus our attention on asymmetric carbons. Let H be a subgraph of G containing a stereo vertex s corresponding to an asymmetric carbon. We say that the stereo property of s is not captured by H if (Definition 1):

$$\begin{aligned} &\exists (i,j) \in \{1, \ldots, |N(s)|\}^2, i \neq j, \\ &\qquad \exists f \in \text{IsomEqOrd}(H, \tau^s_{i,j}(H)) \text{ with } f(s) = s \quad (2) \end{aligned}$$

To define a minimal stereo subgraph of s, we consider a finite sequence $(H_s^k)_{k=1}^n$ of vertex induced subgraphs of G. The first element of this sequence H_s^1 is the smaller vertex induced subgraph for which we can test (2) :

$$V(H_s^1) = \{s\} \cup N(s)$$

where $V(H_s^1)$ and $N(s)$ denote respectively the set of vertices of H_s^1 and the set of neighbors of s.

If the current vertex induced subgraph H_s^k does not characterize the stereo property of s, we know by (2), that it exists some isomorphisms f of equivalent ordered graphs between H_s^k and $\tau^s_{i,j}(H_s^k)$ with $i \neq j$ and $f(s) = s$. Let us consider such an isomorphism f. By definition of equivalent ordered isomorphism, it exists $\sigma \in \Sigma$ such that f is an ordered isomorphism between H_s^k and $\sigma\left(\tau^s_{i,j}(H_s^k)\right)$. By definition of ordered isomorphisms, and since $f(s) = s$, we have:

$$\forall l \in \{1, \ldots, |N(s)|\}, f(v_l) = v_{\sigma(s) \circ \tau^s_{i,j}(l)}.$$

with $ord(s) = v_1, \ldots, v_n$.

As $\sigma(s)$ is an even permutation, $\sigma(s) \circ \tau^s_{i,j}$ is an odd one. Hence it exists l in $\{1, \ldots, |N(s)|\}$ such that $l \neq \sigma(s) \circ \tau^s_{i,j}(l)$.

In other words, any equivalent ordered isomorphism corresponding to equation (2) maps at least one vertex in the neighborhood of s in H_s^k onto a different vertex in the same neighborhood. Let us denote by $\mathcal{E}_f^k$ the set of vertices of H_s^k connected to s by a path whose all vertices are mapped onto other vertices by f:

$$\begin{aligned} \mathcal{E}_f^k = \{v \in V(H_s^k) \mid \exists c = (v_0, \ldots, v_q) \in H_s^k \\ \text{with } v_0 = s \text{ and } v_q = v \ s.t. \\ \forall r \in \{1, \ldots, q\}, f(v_r) \neq v_r\} \quad (3) \end{aligned}$$

For any equivalent ordered isomorphism f satisfying (2), the set $\mathcal{E}_f^k$ is not empty since it contains at least 2 vertices. A vertex v belongs to $\mathcal{E}_f^k$ if neither its label nor its neighborhood in H_s^k allows to differentiate it from $f(v)$. The basic idea of our algorithm consists in enforcing the constraints on each $v \in \mathcal{E}_f^k$ at iteration $k+1$ by adding to H_s^k the neighborhood in G of all vertices belonging to one $\mathcal{E}_f^k$, with f satisfying (2). The set of vertices of the vertex induced subgraph H_s^{k+1} is thus defined by:

$$V(H_s^{k+1}) = V(H_s^k) \cup \bigcup_{f \in \mathcal{F}_s^k} N(\mathcal{E}_f^k) \quad (4)$$

where $\mathcal{F}_s^k$ denotes all equivalent ordered isomorphisms satisfying (2).

Since $f \in \mathcal{F}_s^k$ implies that $\mathcal{E}_f^k$ is not empty, adding iteratively constraints on the existence of vertices in $\mathcal{E}_f^k$ removes f from $\mathcal{F}_s^k$. The algorithm stops when the set $\mathcal{F}_s^k$ becomes empty. Note that such a condition must be satisfied since s is a stereocenter and hence the whole molecule does not satisfies (2).

Algorithm 1 Construction of a minimal stereo subgraph

```
Input: a stereo vertex $s$ and an ordered molecular graph $G$
Output: a minimal stereo sub graph
  $H_s^1 \leftarrow \{s\} \cup N(s)$
  $(\mathcal{F}_s^1, \mathcal{E}_f^1) \leftarrow getIsomorphism(H_s^1)$
  $k \leftarrow 1$
  while $\mathcal{F}_s^k \neq \varnothing$ do
    $k \leftarrow k + 1$
    $V(H_s^k) \leftarrow V(H_s^{k-1}) \cup N(\mathcal{E}_f^{k-1})$
    $(\mathcal{F}_s^k, \mathcal{E}_f^k) \leftarrow getIsomorphism(H_s^k, \mathcal{F}_s^{k-1})$
  end while
```

The main steps of our method are summed up in Algorithm 1. The function getIsomorphism uses a fast isomorphism algorithm [7] to compute the isomorphisms f between H_s^k and $\tau^s_{i,j}(H_s^k)$ and the sets $\mathcal{E}_f^k$ for each $(i,j) \in \{1, \ldots, |N(s)|\}^2$.

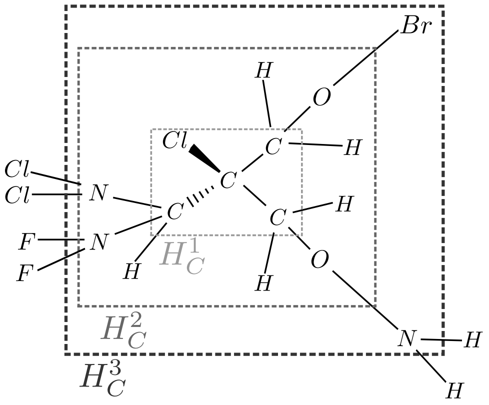


Fig. 3. An asymmetric carbon and its associated sequence $(H_C^k)_{k=1}^3$

Moreover, in order to improve execution times, isomorphisms $\mathcal{F}_s^{k-1}$ found during a previous iteration between H_s^{k-1} and $\tau_{i,j}^s(H_s^{k-1})$ are used to initialize the isomorphism algorithm at step k.

The intermediate vertex induced subgraphs found by our algorithm are illustrated in Fig. 3. Note that at iteration 2, it exists an equivalent ordered isomorphism $f \in \mathcal{F}_C^2$ mapping the path CCO (bottom right of the figure) onto the same path located on the top right part of Fig 3. In this case $\mathcal{E}_f^2$ contains the three carbons of these two paths and both oxygen atoms. The oxygen atoms belong to $\mathcal{E}_f^2$ since their neighborhoods in H_C^2 does not allow to differentiate them (Fig. 3). At iteration 3, the neighborhood in G of these oxygen atoms are added to H_C^3, hence adding N and Br which allows to differentiate both paths and thus removes the equivalent ordered isomorphism f from $\mathcal{F}_C^3$. Note that the neighborhoods of the three carbons atoms are also added but without any influence on H_C^3 since these neighborhoods already belong to H_C^2.

Note that smaller subgraphs than H_s^n which also capture the stereo property of s may be found. For example, in Fig. 3, the subgraph H_{Br}, obtained from H_C^3 by removing the bromine atom (Br) also captures the stereo property of the central carbon C since no equivalent ordered isomorphism between H_{Br} and $\tau_{i,j}^C(H_{Br})$, $i,j \in \{1,\ldots,|N(C)|\}, i \neq j$ may be found. In the same way, the subgraph H_N obtained by removing the nitrogen atom (N) from H_C^3 also captures the stereo property of the central carbon C. However these graphs capture the stereo property of the central carbon atom only thanks to the absence of atoms N in H_{Br} and Br in H_N which forbids to map their incident oxygen atoms one onto the other. Hence these graphs only encode the stereo property of the central carbon atom thanks to a lack of encoding of an information present in the graph. Moreover, there is no easy way to decide which of these graphs should represent the stereo property of the central carbon. This last point may induce a bias if one wish to compare two graphs through their sets of minimal stereo subgraphs. We hence decide to keep the subgraph H_C^3 produced by our algorithm which in Fig. 3 encodes the fact that the central carbon is a stereocenter from the difference of labels between the vertex encoding nitrogen (N) and the one encoding bromine (Br).

For double bond between two carbons, v and w, we have to compute a single minimal stereo subgraph since the stereo property of each carbon of a double bond is connected to the stereo property of the other [6]. A minimal stereo subgraph of v and w is defined the same way as for an asymmetric carbon, the only difference being in the initialisation of the sequence $(H_{v,w}^k)_{k=1}^n$. Indeed to test (2) the smaller vertex induced subgraph for a double bond between carbons is defined by the set of vertices:

$$V(H_{v,w}^1) = \{v,w\} \cup N(v) \cup N(w)$$

IV. Stereo Kernel

The method described in Section III allows to associate a minimal stereo subgraph to each stereocenter of a molecule. We can thus characterize the stereo properties of a molecule through its set of minimal stereo subgraphs. However, a same stereo subgraph may be present more than once in a given molecule. In order to compute efficiently the frequency of a given stereo subgraph within a molecule, we need to associate a unique code to each such subgraph so that the existence of an equivalent ordered isomorphism between two stereo subgraphs may be tested efficiently. We perform such a transformation from stereo subgraphs to codes thanks to [8] which associates each stereo subgraph to an unique code. Note that this code takes explicitly into account the stereoisomerism. Moreover, unlike [7] which allows to find efficiently all isomorphisms between two graphs, [8] associates to each molecule a unique code which allows to test the existence of an equivalent ordered isomorphism between two stereo subgraphs.

Given an ordered graph G, we can thus compute its set of minimal stereo subgraphs $\mathcal{H}(G)$ together with its spectrum $spec(G) = (S_H(G))_{H\in\mathcal{H}(G)}$ which encodes the frequency $S_H(G)$ of each $H \in \mathcal{H}(G)$. The set $\mathcal{H}(G)$ and the spectrum $spec$ provide a characterisation of each stereo center of G and hence describe the stereoisomerism of G.

The comparison of the spectrum of two ordered graphs, is then used to define a kernel between two molecules taking into account the stereoisomerism:

$$k(G,G') = \sum_{H\in\mathcal{H}(G)\cap\mathcal{H}(G')} K(S_H(G), S_H(G')). \quad (5)$$

where K denotes a kernel between real values (e.g. Gaussian, intersection or polynomial). For example, if K is the intersection kernel, the kernel value k of two ordered graphs with non intersecting sets $H(G)$ of stereo subgraphs is zero. The kernel value of two identical ordered graphs is equal to the total number of occurences of their stereo subgraphs. The choice of a particular kernel, together with its parameters is performed through cross-validation.

V. Experiments

We have evaluated our kernel on two datasets connected with stereoisomerism properties. Both datasets correspond to regression problems, one connected to the prediction of a physical molecular property, the other one being related to a biological property. We use the standard SVM regression method [9] for all kernels evaluated in this section.

The first dataset [10] is composed of 90 molecules together with their optical rotations. Optical rotation of a molecule is

TABLE I. ROOT MEAN SQUARED ERROR (RMSE) OBTAINED BY 4 KERNELS IN PREDICTING THE OPTICAL ROTATIONS OF MOLECULES.

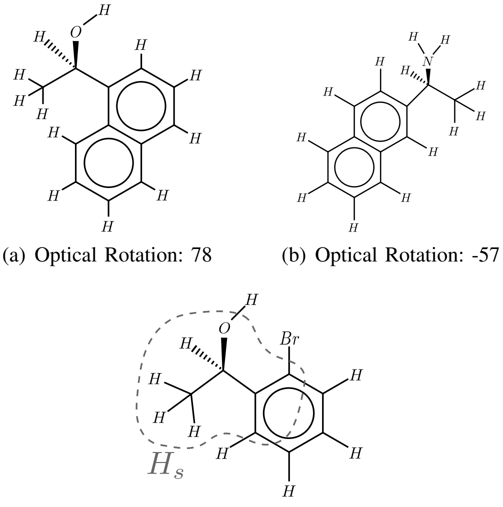
Method	RMSE
Tree patterns Kernel [1]	27.9
Treelet Kernel [2]	26.2
Tree patterns Kernel incorporating stereo information [3]	*25.8*
Stereo Kernel	**17.9**

a physical property measuring the deviation angle of a plane-polarized light passing through a solution of this molecule. In practice, we only select 35 molecules, since almost all molecules have only one stereocenter, and for 55 molecules this stereocenter is unique in the dataset. Such molecules correspond thus to a property (a rotation angle) present only once in the dataset which can not be accurately predicted. The standard deviation of the optical rotation angle is equal to 38.25 for the 35 selected molecules. Due to the limited number of molecules of this dataset we evaluate the rotation angle of each molecule using a leave one out procedure. We used a grid search to choose the different parameters: C of the SVM, the type of sub kernel used in (5) and the parameters of [1], [3]. As we do not have a validation set, the selected parameters for each kernel are the ones which obtain the lowest Root Mean Squared Error (RMSE).

Table I shows the RMSE obtained by our method and three other kernels [1], [2], [3]. The Tree pattern kernel [1] and the treelet kernel [2] do not incorporate any information related to stereoisomerism and obtain consequently the highest errors. The adaptation of the Tree pattern kernel to stereoisomerism [3] obtains better results than these two latter methods. However, in this experiment devoted to the prediction of a property only related to stereoisomerism, the inclusion by this kernel of information not related to stereoisomerism forbids an important decrease of the mean error. Our method represented on the last line of Table I obtains the lowest root mean squared error. We can notice that this error is significantly lower than both the error obtained by [3] (line 3) and the standard deviation of the dataset. As shown in Table II, the sizes of the minimal stereo subgraphs computed by our method on this dataset remain small but are usually larger than the size of the direct neighborhood of an asymmetric carbon (5) or a double bound (6).

Figure 4 shows three molecules of this dataset. The closest molecule from Figure 4(a) according to the treelet kernel is shown in Figure 4(b). Indeed, the only difference between those two molecules, besides the configuration around asymmetric carbons, is an oxygen atom replaced by a nitrogen atom. Thus all treelets, apart those which include the oxygen for the first molecule and the nitrogen for the second molecule, are identical. The third molecule (Figure 4(c)), is the most similar to the first molecule according to our stereo kernel, as they have a same minimal stereo subgraph. Since the treelet kernel considers as similar, molecules with very different optical rotation, it can not accurately predict this property. Unlike the treelet kernel, the stereo kernel provides a suitable measure of similarity when considering properties involving stereoisomerism, and thus allows to obtain a better prediction of optical rotation.

The second dataset is a dataset of synthetic vitamin D derivatives, used in [3]. This dataset is composed of 69 molecules containing cycles, with an average of 9 stereocenters per molecule. The value to predict is their biological activity. This dataset has been selected since, like in many biological properties, stereoisomerism is an important feature of molecules. However, other molecular properties, non related to stereoisomerism may also partially determine biological properties of this dataset.



(a) Optical Rotation: 78

(b) Optical Rotation: -57

(c) Optical Rotation: 54. The minimal stereo subgraph of the central carbon is surrounded by a grey dotted line.

Fig. 4. Three molecules of the first dataset with their optical rotations. The label of an atom is C if it is not specified.

TABLE II. SIZE OF GRAPHS ($|G|$) AND SIZE OF MINIMAL STEREO SUBGRAPHS ($|H_s|$) IN THE TWO DATASETS

	Dataset 1		Dataset 2	
	$\|G\|$	$\|H_s\|$	$\|G\|$	$\|H_s\|$
Minimum size	14	6	68	10
Maximum size	32	13	88	24
Average size	21.3	10.4	76.9	13.7

After normalizing the values of the dataset, the standard deviation of biological activities is equal to 0.258. To choose the different parameters and estimate the performance of each kernel on this dataset we use a nested cross-validation. The outer cross-validation is a leave-one-out procedure, used to compute an error for each molecule of the dataset. For each fold, we use another leave-one-out procedure on the remaining molecules, to compute a validation error. Parameters which provide the lowest root mean squared error on the validation are selected. We obtain for each molecule an error, and report in Table III, the mean of this distribution of errors together with the confidence interval at 95 % of this distribution.

As in our previous experiment, greatest errors in Table III are obtained by methods [1], [2] which do not include stereo

TABLE III. PREDICTION OF THE BIOLOGICAL ACTIVITY OF SYNTHETIC VITAMIN D DERIVATIVES.

Method	Mean Error	Confidence interval at 95%
Tree patterns Kernel [1]	0.193	± 0.060
Treelet Kernel [2]	0.207	± 0.064
Tree patterns Kernel with stereo information [3]	**0.138**	± 0.043
Stereo Kernel	0.141	± 0.047
Stereo Kernel x Treelet Kernel	**0.138**	± 0.044

TABLE IV. Execution times required to compute the 69×69 Gram matrices of our second dataset.

Method	Gram's matrices computations (s)
Tree patterns Kernel [1]	230
Treelet Kernel [2]	7
Stereo Kernel	86

information. The adaptation of the tree pattern kernel to stereoisomerism [3] and our method improve the results over the two previous methods hence showing the insight of adding stereoisomerism information. For this dataset, the inclusion by [3] of information not related to stereoisomerism allows to obtain slightly better results than our method alone, since the biological property to predict is not only dependent of stereosiomerism. By combining our method with the treelet kernel [2], we obtain results as good as those obtained by method [3]. To combine the two kernel, we multiply each subkernel of the treelet kernel by the stereo kernel. Hence, two molecules will be similar according to this new kernel if they have a similar set of stereo subgraphs and a similar set of treelets. We can note that the mean size of molecules on this dataset is about 3 times larger than on the previous one (Table II). However, the mean size of the minimal stereo subgraphs only slightly increases from 10.4 to 13.7. This last point illustrates the fact that the size of a stereo subgraph mainly depends on the variability of vertices and edges labels around its stereocenters and not on the size of the molecule. Nevertheless, a greater number of larger molecules increases the possibility of getting larger minimal stereo subgraphs as observed in Table II.

Execution times required to compute Gram matrices of the second dataset are displayed in Table IV. We do not discuss the execution times required to compute the Gram matrices of the first dataset, since this dataset is composed of a reduced number of molecules having only one stereocenter. The execution times required by our method on this dataset would thus be low but not sufficiently significant.

The first line of Table IV shows that the Tree pattern kernel [1] and its adaptation to stereoisomerism [3] take about 4 minutes to compute the whole Gram matrix. This important execution time may be partially explained by the polynomial complexity of this kernel. An additional and certainly more explicative reason of this important execution time comes from the use by this kernel of implicit bags of patterns. Such bags should be computed for each evaluation of a value of the kernel. On this particular dataset, the bag of each molecule is computed implicitly 69 times. On the contrary, the treelet kernel [2] is based on a bag extraction algorithm with a linear complexity and the bag of treelets attached to each molecule is stored explicitly. Such a bag is thus computed only once for each molecule during the computation of the Gram matrix. The execution time required to compute the Gram matrix of the treelet kernel is consequently the lowest one. Finally our stereo kernel obtains an intermediate result of 83 seconds. This latter execution time, may be explained by the fact that although our kernel uses a subgraph isomorphism algorithm, these subgraphs are in practice of small size (Table II) hence leading to small execution times. Moreover, like the treelet kernel, our kernel is based on an explicit enumeration of patterns. The bag of stereo subgraphs describing each molecule is thus computed only once for each molecule during the computation of the Gram matrix.

VI. Conclusion

The study and the definition of new stereoisomers constitutes an important subfield of chemistry and thus a major challenge in chemoinformatics. Indeed, stereoisomers of some common drugs may be considered as violent poisons. For example, a molecule called thalidomide was sold in the late fifties as an anti nausea for pregnant women. However, it turns out that one of the stereoisomer of this molecule could cause fetal malformation. Up to now, only few methods have proposed pattern recognition methods taking explicitly into account stereoisomerism.

We have proposed in this paper, a graph kernel based on an explicit enumeration of all the stereo subgraphs of a molecule. Each stereo subgraph is associated to a stereo vertex and encodes the part of the graph which provides the stereo property to this vertex. Based on the notion of stereo subgraph we propose to describe a molecule by its bag of stereo subgraphs. The similarity between two molecules is then encoded through a graph kernel based on the similarity of both bags. Experiments on two datasets related to stereoisomerism properties demonstrate the relevance of our approach. In a future work we plan to investigate the insight provided by the addition to our kernel of larger subgraphs encoding relationships between each stereo subgraph and the remaining parts of a molecule.

VII. Acknowledgements

The authors wish to thank the association CRIHAN for their computing resource.

References

[1] P. Mahé and J.-P. Vert, "Graph kernels based on tree patterns for molecules," *Machine Learning*, vol. 75, no. 1, pp. 3–35, Oct. 2008.

[2] B. Gaüzère, L. Brun, and D. Villemin, "Two New Graphs Kernels in Chemoinformatics," *Pattern Recognition Letters*, vol. 33, no. 15, pp. 2038–2047, 2012.

[3] J. Brown, T. Urata, T. Tamura, M. A. Arai, T. Kawabata, and T. Akutsu, "Compound analysis via graph kernels incorporating chirality," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. 1, pp. 63–81, 2010.

[4] P.-A. Grenier, L. Brun, and D. Villemin, "Treelet kernel incorporating chiral information," in *Graph-Based Representations in Pattern Recognition*. Springer, 2013, pp. 132–141.

[5] M. Petitjean, "Chirality in metric spaces," *Symmetry, Culture and Science*, vol. 21, pp. 27–36, 2010.

[6] P.-A. Grenier, L. Brun, and D. Villemin, "Incorporating stereo information within the graph kernel framework," CNRS UMR 6072 GREYC, Tech. Rep., 2013, http://hal.archives-ouvertes.fr/hal-00809066/.

[7] V. Bonnici, R. Giugno, A. Pulvirenti, D. Shasha, and A. Ferro, "A subgraph isomorphism algorithm and its application to biochemical data," *BMC Bioinformatics*, vol. 14, no. Suppl 7, p. S13, 2013.

[8] W. T. Wipke and T. M. Dyott, "Stereochemically unique naming algorithm," *Journal of the American Chemical Society*, vol. 96, no. 15, pp. 4834–4842, 1974.

[9] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *NIPS*, 1996, pp. 155–161.

[10] H.-J. Zhu, J. Ren, and C. U. Pittman Jr, "Matrix model to predict specific optical rotations of acyclic chiral molecules," *Tetrahedron*, vol. 63, no. 10, pp. 2292–2314, 2007.