# Taking into account stereoisomerism in the prediction of molecular properties

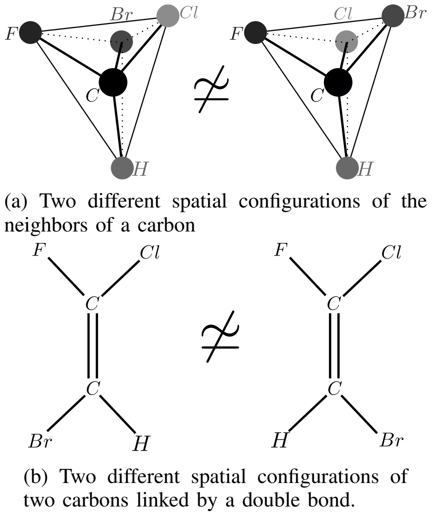
Pierre-Anthony Grenier, Luc Brun
Normandie Univ, ENSICAEN,
CNRS, GREYC, 14000 Caen, France
{pierre-anthony.grenier,luc.brun}@ensicaen.fr

Didier Villemin
Normandie Univ, ENSICAEN,
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 orientation 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 a previous paper, we proposed to encode the stereoisomerism property of each atom by a local subgraph, called minimal stereo subgraph, and we designed a kernel based on the comparison of bags of such subgraphs.**

**However, the encoding of a molecule by a bag of subgraphs induces an important loss of information. In this paper, we propose a new kernel based both on the spatial relationships between minimal stereo subgraphs and the local neighbourhood of each minimal stereo subgraph within its supergraph. Our experiments show the benefits of taking into account such information.**



(a) Two different spatial configurations of the neighbors of a carbon

(b) Two different spatial configurations of two carbons linked by a double bond.

Fig. 1: Two types of stereocenters.

## 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". An usual way to encode molecules is to use their 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 $e \in E$ encodes a bond between two atoms and the labeling functions $\mu$ and $\nu$ 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).

However, molecular graphs have a limitation: they do not encode the spatial configuration of atoms. Some molecules, called stereoisomers, are associated to a same molecular graph but differ by the relative positioning of their 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 1a). An atom is called a stereocenter if a permutation of two atoms belonging to its neighborhood produces a different stereoisomer. Two connected atoms also define a stereocenter if a permutation of the positions of two atoms belonging to the union of their neighborhoods produces a different stereoisomer (Figure 1b). According to chemical experts [1], within molecules currently used in chemistry, $98\%$ of stereocenters correspond either to carbons with four neighbors, called asymmetric carbons (Figure 1a) or to couples of two carbons adjacent through a double bond (Figure 1b). We thus restrict the present paper to such cases.

Graph kernels [2], [3], [4], allow to combine a graph encoding of molecules with usual machine learning methods. Brown et al. [5] have proposed to take into account stereoisomerism through an extension of the tree-pattern kernel [3]. In this method, similarity between molecules is deduced from the number of common tree-patterns between two molecules.

Intuitively, stereoisomerism property is related to the fact that permuting two neighbors of a stereocenter produces a different spatial configuration. If those two neighbors have a same label, the influence of the permutation should be searched beyond the direct neighborhood of this stereocenter. Based on this ascertainment, we have proposed in [6] to characterize a stereocenter by a subgraph, called a minimal stereo subgraph, big enough to highlight the influence of each permutation of the neighbors of this stereocenter but sufficiently small to provide a local characterization of it. We then proposed a kernel based on those subgraphs. One limitation of this approach, is that graph information is reduced to a bag of subgraphs without taking into account the possible interactions between these subgraphs nor the neighbourhood of each instance of a subgraph within the whole graph. Thus, in [7], we proposed to construct a graph, where each vertex represent a minimal stereo subgraph and each edge encodes an interaction between two subgraphs. By using a graph kernel on this graph we are able to take into account interactions between minimal stereo subgraphs. However we do not take into account the neighbourhood of a minimal stereo subgraph within the whole graph. In this paper we present a way to combine both information within a unified framework.

In Section II we remind the points of [6] and [7], the encoding of molecules by ordered graphs, the construction of minimal stereo subgraphs which characterize stereocenters and the construction of graphs of interactions. Then in Section III we present a model which allows to take into account the neighbourhood of a minimal stereo subgraph and a way to integrate this model in our previous framework. Results obtained with this new method are provided in Section IV.

## II. Minimal stereo subgraphs and graphs of interactions

### A. Encoding of molecules by ordered graphs

The spatial configuration of the neighbors of each atom may be encoded through an ordering of its neighborhood [6]. In order to encode this information, we introduce the notion of ordered graph. An ordered graph $G = (V, E, \mu, \nu, ord)$ is a molecular graph $\widehat{G} = (V, E, \mu, \nu)$ 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 it exists an isomorphism $f$ between their respective molecular graphs $\widehat{G}$ and $\widehat{G'}$ 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 $v$). In this case $f$ is called an ordered isomorphism between $G$ and $G'$.

However, different ordered graphs may encode a same molecule. We thus have to define an equivalence relationship between ordered graphs, such that two ordered graphs are equivalent if they represent a same molecular configuration.

To do so, we introduce the notion of re-ordering function $\sigma$, 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 $v$.

The set of re-ordering functions, transforming an ordered graph into another one representing the same configuration is called a valid family of re-ordering functions $\Sigma$ [8]. We say that it exists an equivalent ordered isomorphism $f$ between $G$ and $G'$ according to $\Sigma$ if it exists $\sigma \in \Sigma$ such that $f$ is an ordered isomorphism between $\sigma(G)$ and $G'$ ($\sigma(G) \underset{o}{\simeq} G'$). The equivalent order relationship defines an equivalence relationship [8] and two different stereoisomers are encoded by non equivalent ordered graphs. We denote by $\text{IsomEqOrd}(G, G')$ the set of equivalent ordered isomorphism between $G$ and $G'$.

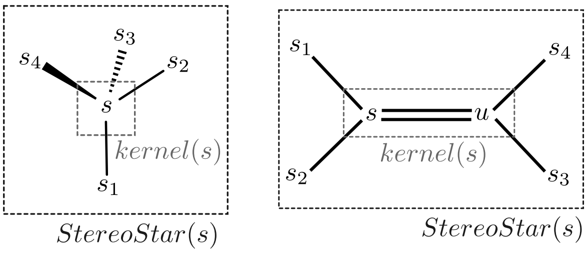


Fig. 2: Stereocenters and their neighbourhoods.

Combinatorial map are a special case of ordered graph where reordering functions are cyclic permutation. In general, ordered graphs can be used for any application of pattern recognition where data can be represented by graph and where the ordering of the vertices is important.

Carbons with four neighbors, and double bonds between carbons, are not necessarily stereocenters. If they are not stereocenters, any permutation in their neighbourhood would lead to an equivalent ordered graph. We thus define for an ordered graph $G = (V, E, \mu, \nu, ord)$ and one of its vertex $v \in V$ a set of ordered isomorphism $\mathcal{F}_G^v$:

$$\mathcal{F}_G^v = \bigcup_{\substack{(i,j)\in\{1,\dots,|N(v)|\}^2 \\ i\neq j}} \{f \mid f \in \text{IsomEqOrd}(G, \tau_{i,j}^v(G)) \quad \text{with } f(v) = v\}$$

where $\tau_{i,j}^v$ is a re-ordering function equals to the identity on all vertices except $v$ for which it permutes the vertices of index $i$ and $j$ in $ord(v)$. Intuitively, isomorphisms in $\mathcal{F}_G^v$ correspond to a symmetry of the neighbors of $v$.

We then 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 $\mathcal{F}_G^v = \varnothing$.

Two carbons linked by a double bond form a stereocenter and we have proved in [8] that if a carbon of a double bond is a stereo vertex then the other one is also a stereo vertex. Therefore we denote by $kernel(s)$ the set of stereo vertices corresponding to a stereocenter ($kernel(s) = \{s\}$ if $s$ is an asymmetric carbon and $kernel(s) = \{s, u\}$ if $s$ is a carbon of a double bond, where $u$ is the other carbon of the double bond). We further denote by $StereoStar(s)$ the set composed of a stereocenter and its neighbourhood: $StereoStar(s) = kernel(s) \cup N(kernel(s))$ (Figure 2) where $N(kernel(s))$ is the neighbourhood of the vertices of $kernel(s)$.

### B. Minimal stereo subgraphs

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 $s$, large enough to characterize the stereo property of $s$, but sufficiently small to encode only the relevant information characterizing the stereo property of $s$. Such a subgraph is called a minimal stereo subgraph of $s$.

We now present a constructive definition of a minimal stereo subgraph of a stereo vertex. Let $s$ denotes a stereo vertex and let $H_s$ be a subgraph of $G$ containing $kernel(s)$. We say that the stereo property of $s$ is not captured by $H_s$ if (Definition 1):

$$\mathcal{F}^s_{H_s} \neq \varnothing \qquad (1)$$

To define a minimal stereo subgraph of $s$, we consider a finite sequence $(H^k_s)^n_{k=1}$ of vertex induced subgraphs of $G$. The first element of this sequence $H^1_s$ is the smallest vertex induced subgraph for which we can test (1): $V(H^1_s) = StereoStar(s)$.

If the current vertex induced subgraph $H^k_s$ does not capture the stereo property of $s$, we know by (1), that it exists some isomorphisms $f \in \mathcal{F}^s_{H^k_s}$. We denote by $\mathcal{E}^k_f$ the set of vertices of $H^k_s$ inducing the isomorphism $f$ in $H^k_s$:

$$\begin{aligned}\mathcal{E}^k_f = \{&v \in V(H^k_s) \mid \exists p = (v_0, \ldots, v_q) \in H^k_s \\ &\text{with } v_0 \in kernel(s),\, v_q = v \text{ and } f(v_1) \neq v_1\} \qquad (2)\end{aligned}$$

In [8], we show that for any $f$ in $\mathcal{F}^s_{H^k_s}$, $\mathcal{E}^k_f$ is not empty. A vertex $v$ belongs to $\mathcal{E}^k_f$ if neither its label nor its neighborhood in $H^k_s$ allow to differentiate it from $f(v)$. The basic idea of our algorithm consists in enforcing constraints on each $v \in \mathcal{E}^k_f$ at iteration $k+1$ by adding to $H^k_s$ the neighborhood of $v$ in $G$. The set of vertices of the vertex induced subgraph $H^{k+1}_s$ is thus defined by:

$$V(H^{k+1}_s) = V(H^k_s) \cup \bigcup_{f \in \mathcal{F}^s_{H^k_s}} N(\mathcal{E}^k_f) \qquad (3)$$

where $N(\mathcal{E}^k_f)$ denote the neighborhood of $\mathcal{E}^k_f$.

The algorithm stops when the set $f \in \mathcal{F}^s_{H^k_s}$ becomes empty. We proved in [8] that the subgraph obtained by this algorithm captures the stereo property of $s$. Figure 3 illustrates our algorithm. Remarks that the computation of the minimal stereo subgraph requires the computation of graph isomorphisms and is thus nearly NP-complete. However, minimal stereo subgraphs correspond to a local characteristic of a vertex and have consequently a limited size [6].

Thus for each stereo vertex we can construct its minimal stereo subgraph to characterize it. We consider two stereo vertices as similar if they have a same minimal stereo subgraph, and to test it efficiently, we transform each minimal stereo subgraph $S$ into a code $c_S$ thanks to the method described in [9]. The set of minimal stereo subgraphs of a graph $G$ is denoted by $\mathcal{H}(G)$.

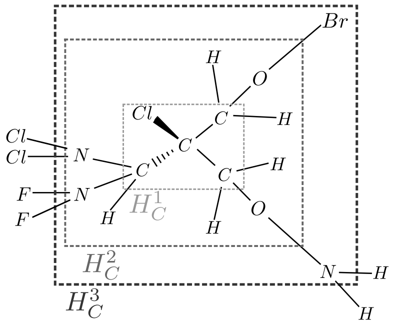


Fig. 3: An asymmetric carbon and its associated sequence $(H^k_C)^3_{k=1}$

### C. Graph of interactions

We now propose to encode interactions between minimal stereo subgraphs. To do so, we define a function of interactions $F$ between minimal stereo subgraphs. This function of interactions is defined according to a sequence of conditions $(cond_1, \ldots, cond_n)$. These conditions are increasingly constraining:

$$\forall i \in \{1, \ldots, n-1\}\, cond_{i+1} \Rightarrow cond_i$$

Let $S_1$ and $S_2$ be two minimal stereo subgraphs of a same ordered graph, such that $s_1$ is the stereo vertex of $S_1$ and $s_2$ is the stereo vertex of $S_2$. We propose the following set of conditions:

$$\begin{aligned} cond_0 &: StereoStar(s_1) \not\subset S_2 \\ cond_1 &: StereoStar(s_1) \subset S_2 \\ cond_2 &: S_1 \subset S_2 \end{aligned} \qquad (4)$$

The value $F(S_1, S_2)$ is obtained by taking the index $j$ of the condition $cond_j$ which represents the strongest interaction between $S_1$ and $S_2$:

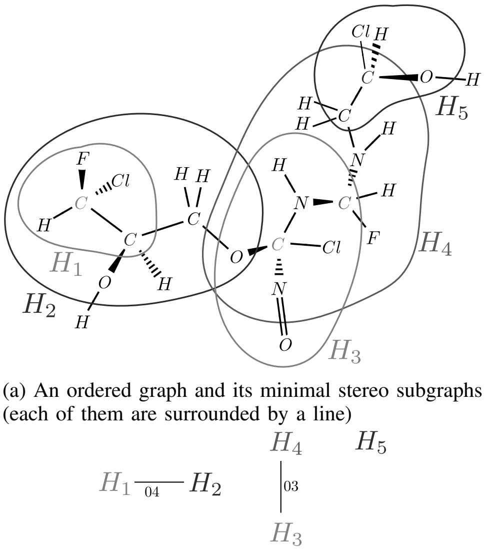
$$F(S_1, S_2) = \max\{j \in \{0, 1, 2\} \mid cond_j\}$$

$F(S_1, S_2)$ equals to zero means that we consider that $S_1$ does not interact with $S_2$. Note that $F$ is a non symmetric function.

We define thanks to this function, a graph of interactions $G_i$ where each vertex $v \in V_i$ represents a minimal stereo subgraph and each edge encodes an interaction between two minimal stereo subgraphs deduced from $F$ :

**Definition 2** (Graph of interactions)**.** Let $G = (V, E, \mu, \nu, ord)$ denotes an ordered graph, and $\mathcal{H}(G) = \{S_1, \ldots, S_n\}$ its set of minimal stereo subgraphs. A graph of interactions $G_i = (V_i, E_i, \mu_i, \nu_i)$ is a graph built from $G$ and the function of interaction $F$. Each vertex $u_j$ of $V_i$ corresponds to a minimal stereo subgraphs $S_j$ of $G$ and $(u_j, u_k) \in E_i$ only if $F(S_j, S_k)$ or $F(S_k, S_j)$ is not null. The labels of the graph of interactions are defined by :

- $\forall u_j \in V_i,\ \mu_i(u) = c_{S_j}$.
- $\forall e = (u_j, u_k) \in E_i,\ \nu_i(e) = \min(F(S_j, S_k), F(S_k, S_j)) \odot \max(F(S_j, S_k), F(S_k, S_j))$.



(a) An ordered graph and its minimal stereo subgraphs (each of them are surrounded by a line)

(b) Its graph of interaction

Fig. 4: One ordered graph and its graph of interactions $G_i$

where $\odot$ denotes the concatenation and $c_S$ is the code describing $S$ and defined in [9].

Figure 4 shows a graph of interactions obtained from an ordered graph. We can see that the edge between $H_1$ and $H_4$ is labeled by 04 because $F(H_1, H_2) = 4$ and $F(H_2, H_1) = 0$. In practice, a molecular ordered graph have few identical minimal stereo subgraphs. Thus few vertices have identical labels in a graph of interaction and an edge is almost defined by the labels of its vertices. Therefore using directed graphs as graph of interaction would not give much more information. Moreover there exists a lot of graph kernels (e.g. [2], [3], [4]) which can be used to measure similarities of undirected graphs. By using one of those graph kernel on the graph of interactions, we obtain a kernel which takes into account stereoisomerism and interactions between minimal stereo subgraphs.

However, from an intuitive point of view, a stereo subgraph partially fixes the geometry of a part of a molecule. Remaining parts of the graph attached to the different extremities of a stereo subgraph should thus play different roles in the property to predict according to the extremity to which they are attached. We have thus to take into account the neighbourhood of each minimal stereo subgraph into our final kernel.

## III. Neighbourhood of minimal stereo subgraphs

In this section, we present a method to take into account the direct neighbourhood of minimal stereo subgraphs. We first construct a kernel between minimal stereo subgraphs, which compares their direct neighbourhood.

### A. *Kernel between minimal stereo subgraphs*

For a stereo subgraph $S$, we denote $\delta_{in}(S)$ the set of vertices on the boundaries of $S$ :

$$\delta_{in}(S) = \{v \in S \mid N(v) \not\subset S\}$$

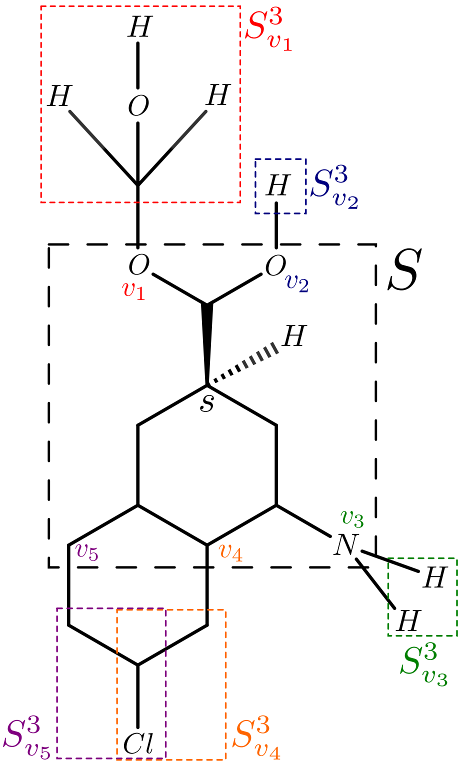


Fig. 5: A minimal stereo subgraph $S$ with the vertex of its boundaries $\{v_1, v_2, v_3, v_4, v_5\}$ and their 3-neighborhood.

For each vertex $v$ on the boundary of a minimal stereo subgraphs $S$ we define a subgraph $S_v^k$ called the $k$-neighborhood of $v$ :

**Definition 3** ( $k$-neighborhood )**.** Let $G = (V, E, \mu, \nu, ord)$ be an ordered graph. We denote $s$ a stereo vertex of $G$ and $S$ its minimal stereo subgraph. The $k$-neighborhood of $v$, a vertex of $\delta_{in}(S)$, is the induced subgraph $S_v^k$ of $G$ such that:

$$V_{S_v^k} = \left\{ u \in G - S \;\middle|\; \begin{array}{l} d(u,v) \leq k \\ \forall v' \in \delta_{in}(S),\ d(u,v) \leq d(u,v') \end{array} \right\}$$

Figure 5 shows an example of $k$-neighborhoods associated to vertices of the boundary of a minimal stereo subgraph. We can notice that a $k$-neighborhood can be disconnected ($S^3_{v_3}$) and that two $k$-neighborhoods can have a non empty intersection ($S^3_{v_4}$ and $S^3_{v_5}$).

We want to compare two minimal stereo subgraphs located in different graphs, such that there is an equivalent ordered isomorphism $f$ between them. As they can have different surroundings, their $k$-neighborhoods are compared in order to have a local measure of similarity. However, in order to respect the orientation provided by stereocenters, we do not compare all the pairs of $k$-neighborhood but only the ones associated to vertices $u$ and $v$ which can be matched by an equivalent ordered isomorphism. As we compare a subset of pairs of $k$-neighborhoods we define our kernel as a matching kernel [11].

For a minimal stereo subgraph $S$ we denote by $(v_1, \ldots, v_n)$ an ordering of $\delta_{in}(S)$. We also denote $se(S)$ the ordered sequence of the $k$-neighborhoods associated to the sequence $(v_1, \ldots, v_n)$:

$$se(S) = (S^k_{v_1}, \ldots, S^k_{v_n})$$

The mapping between two minimal stereo subgraphs $S$ and $S'$ is defined as :

$$M_{S,S'} = \left\{ \text{(se(S),se(S'))} \;\middle|\; \begin{array}{l} \exists f \in \text{IsomEqOrd}(S, S') \\ s.t\ \forall i \in \{1, \ldots, n\}, \\ f(v_i) = v'_i \end{array} \right\}$$

where $se(S) = (S^k_{v_1}, \ldots, S^k_{v_n})$ and $se(S') = (S'^k_{v'_1}, \ldots, S'^k_{v'_n})$.

The kernel between those minimal stereo subgraphs is then defined by:

$$k_{inf}(S, S') = \frac{\sum\limits_{(se(S),se(S')) \in M_{S,S'}} \prod\limits_{i=1}^{n} k_t(S^k_{v_i}, S'^k_{v'_i})}{\sqrt{|\delta_{in}(S)|!|\delta_{in}(S')|!}} \quad (5)$$

where $k_t$ is a kernel between graphs. Note that, the normalization by $\sqrt{|\delta_{in}(S)|!|\delta_{in}(S')|!}$ allows to discard the influence of the arbitrary ordering of $\delta_{in}(S)$. If $M_{S,S'}$ is empty, there is no equivalent ordered isomorphism between $S$ and $S'$, and thus $k_{inf}(S, S')$ is equal to zero.

The kernel $k_t$ is a "weight" kernel defined by :

$$k_t(G, G') = e^{\frac{-(w-w')^2}{d}} \quad (6)$$

where $d$ is a parameter and $w$ the weight of a molecular graph (defined as the sum of the weights of atoms encoded by vertices of the graph). In practice, we also tested usual graph kernels [3], [4], which do not provide significantly better results than this kernel.

We can define a kernel between ordered graphs by comparing their sets of minimal stereo subgraphs:

$$k_{infG}(G, G') = \sum_{S \in \mathcal{H}(G)} \sum_{S' \in \mathcal{H}(G')} k_{inf}(S, S') \quad (7)$$

We propose in the next section a method to use the kernel of equation 5, within the framework of graph of interactions (Section II-C).

### *B. Integration within the graph of interactions*

In [7], we used several graph kernels [2], [3], [4] to compute a measure of similarity between graphs of interactions.

Treelet are all the labeled subtrees with six or less vertices of a graph. The authors of [4] define how to compute those treelets, and a code formed with their labels, which allows to test efficiently if two treelets are isomorphic. The treelet kernel applied to the graph of interactions [7] is defined by :

$$k_T(G, G') = \sum_{t \in \mathcal{T}(G_i) \cap \mathcal{T}(G'_i)} K(f_t(G_i), f_t(G'_i)) \quad (8)$$

where $G_i$ is the graph of interactions of $G$, $\mathcal{T}(G_i)$ is the set of treelets of $G_i$, $f_t(G_i)$ is the number of occurrence of the treelet $t$ in the graph $G_i$ and $K$ is a definite positive kernel between real numbers.

The treelet kernel compare the number of occurrences of each treelet present in both graph. Two treelets are identical if they have the same structure and the same labels. In graph of interactions, labels of vertices are the code defined in [9] which describes minimal stereo subgraph. However, if we use the kernel between minimal stereo subgraphs (equation 5), two minimal stereo subgraphs with identical code are no longer considered as identical since they may have different neighbourhoods. Thus we can no longer count the identical patterns within graphs of interactions. Let us first remark that if $K$ is a scalar product (8) may be rewritten as follows :

$$k_T(G, G') = \sum_{t \in \mathrm{T}(G_i)} \sum_{t' \in \mathrm{T}(G'_i)} \delta(t, t') \quad (9)$$

where $\delta(t, t')$ a function equal to 1 if there is an isomorphism between $t$ and $t'$ and 0 otherwise and $\mathrm{T}(G_i)$ is the bag of treelets of $G_i$. Hence unlike in $\mathcal{T}(G_i)$ (equation 8), an element $t$ of $\mathrm{T}(G_i)$ may appear several times.

As we want to compare treelets with identical labeling, we can replace $\delta(t, t')$ by a kernel between $t$ and $t'$ in (9):

$$k_T(G, G') = \sum_{t \in \mathrm{T}(G_i)} \sum_{t' \in \mathrm{T}(G'_i)} \sum_{\phi \in \text{Isom}(t,t')} k(t, \phi(t)) \quad (10)$$

where $\text{Isom}(t, t')$ is the set of isomorphism between $t$ and $t'$. Note that if $\text{Isom}(t, t')$ is empty, $\sum\limits_{\phi \in \text{Isom}(t,t')} k(t, \phi(t))$ is equal to zero.

Each vertex of $t$ (respectively $t'$) is associated to a minimal stereo subgraph of $G$ (respectively $G'$). To compare the neighbourhood of each minimal stereo subgraph associated to a treelet $t$, the subkernel $k$ is defined by :

$$k(t, \phi(t)) = \prod_{v \in t} k_{inf}(S(v), S(\phi(v))) \quad (11)$$

where $k_{inf}$ is the kernel between minimal stereo subgraphs defined in equation (5), and $S(v)$ is the minimal stereo subgraph associated to the vertex $v$ of $t$.

By using the kernel of equation (10) on graph of interaction, with the subkernel of equation (11), we obtain a measure of similarity between molecules, which takes into account both the direct neighbourhoods of minimal stereo subgraphs and the interactions between minimal stereo subgraphs.

Moreover, some treelets may have more influence on a property than others. Based on this assumption, [12] proposed to combine the treelet kernel with a multiple kernel learning method [13], by learning a weight $w_t$ for each treelet $t$. The formulation of equation (8) becomes :

$$k_T(G, G') = \sum_{t \in \mathcal{T}(G) \cap \mathcal{T}(G')} w_t K(f_t(G), f_t(G')) \quad (12)$$

where $w_t$ is the weight associated to the kernel $t$.

TABLE I: Time of computation of the gram matrices

| | Method | Time in seconds |
|---|---|---|
| 1 - | Stereo Kernel [6] | 0.7 |
| 2 - | Graph of interactions [7] with [4] | 0.8 |
| 3 - | Kernel between minimal stereo subgraphs (7) | 3 |
| 4 - | Graph of interactions and neighbourhood (10) | 11 |

TABLE II: Prediction of the biological activity of synthetic vitamin D derivatives.

| | Method | RMSE |
|---|---|---|
| 1 - | Tree pattern kernel [3] | 0.251 |
| 2 - | Treelet kernel [4] | 0.271 |
| 3 - | Tree pattern kernel with stereo [5] | 0.184 |
| 4 - | Stereo Kernel [6] | 0.194 |
| 5 - | Graph of interactions [7] with [4] | 0.171 |
| 6 - | Graph of interactions [7] with [3] | 0.161 |
| 7 - | Graph of interactions [7] with [4] and MKL | 0.172 |
| 8 - | Kernel between minimal stereo subgraphs (7) | 0.177 |
| 9 - | Graph of interactions and neighbourhood (10) | 0.177 |
| 10 - | Graph of interactions, neighbourhood and MKL (13) | **0.154** |

In the same way, we can change the formulation of equation (10) to integrate a weight for each treelet extracted from the graph of interactions:

$$k_T(G,G') = \sum_{t\in \mathrm{T}(G)} \sum_{t'\in \mathrm{T}(G')} w_t \sum_{\phi\in \mathrm{Isom}(t,t')} k(t,\phi(t)) \quad (13)$$

With this formulation, we can use a multiple kernel learning algorithm to weight the influence of each treelet of the graph of interactions.

## IV. Experiments

We have tested our method on a dataset of synthetic vitamin D derivatives, used in [5]. This dataset is composed of 69 molecules, with an average of 8.55 stereocenters per molecule. This dataset is associated to a regression problem, which consists in predicting the biological activity of each molecule.

Table I shows the computing time of the gram matrices for this dataset for each of our methods. We can see that the extension induce more computing time.

For all the experiments we use the same protocol: a nested cross-validation which selects parameters and estimates the performance. The outer cross-validation is a leave-one-out procedure, used to compute an error of prediction 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. We use standard SVM methods for regression of molecules.

We can see in Table II, that methods which do not encode stereoisomerism information [3], [4] obtain poor results (lines 1 and 2). Adding stereoisomerism information allows to obtain better results as the tree pattern kernel with stereo [5] (line 3) and the stereo kernel [6] (line 4) obtain better results than the two previous ones. The adaptation of the tree pattern kernel to stereoisomerism [5] is however better than the stereo kernel [6] because it can implicitly take into account the neighbourhood of minimal stereo subgraphs. By taking this neighbourhood into account explicitly (line 8) or by taking into account the interactions between minimal stereo subgraphs (lines 5-7) we are able to obtain better results than [5] (line 3).

By combining the graph of interactions, with the kernel between minimal stereo subgraphs (line 9) we do not obtain better results than with the kernel between minimal stereo subgraphs alone (line 8) or the graph of interactions alone (lines 7-5). However, by using a multiple kernel learning algorithm on the treelets of the graph of interactions and using the kernel between minimal stereo subgraphs (line 10) we improve significantly the best result obtained so far. Note that the multiple kernel method applied on the graph of interactions (with a dirac kernel between stereo subgraphs) does not provide any improvement (lines 5 and 7). Our kernel between stereo subgraphs taking into account their embedding into their original graphs (equation 10 and 11) provide thus an additional information put in evidence by the multiple kernel method.

## V. Conclusion

The stereo kernel [6] which compares the minimal stereo subgraphs of two ordered graphs has one drawback : graph information are reduced to a bag of subgraphs without taking into account possible interactions between these subgraphs nor the neighbourhood of each instance of a subgraph within the whole graph. The graph of interactions introduced in [7] allows to take into account the interactions between minimal stereo subgraphs, but not the direct neighbourhood of those subgraphs.

In this paper we have presented a kernel which take into account the neighbourhood of minimal stereo subgraphs. Moreover, we have shown how to integrate this kernel into the framework of the graph of interactions. This combination provides a significant decrease of the prediction error.

## VI. Acknowledgements

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

## References

[1] J. Jacques, A. Collet, and S. H. Wilen, *Enantiomers, racemates, and resolutions*. Wiley, 1991.

[2] H. Kashima, K. Tsuda, and A. Inokuchi, "Marginalized kernels between labeled graphs," in *ICML*, vol. 3, 2003, pp. 321–328.

[3] 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.

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

[5] 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.

[6] P.-A. Grenier, L. Brun, and D. Villemin, "A graph kernel incorporating molecule's stereisomerism information," *Proceedings of 22nd International Conference on Pattern Recognition (ICPR)*, pp. 631–636, 2014.

[7] ——, "From bags to graphs of stereo subgraphs in order to predict molecules properties," *Graph-Based Representations in Pattern Recognition*, pp. 305–314, 2015.

[8] P.-A. Grenier, "Modélisation de la stéréochimie: une application à la chémoinformatique," Ph.D. dissertation, Université de Caen Normandie, 2015.

[9] 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.

[10] S. Bernstein, W. J. Kauzmann, and E. S. Wallis, “The relationship between optical rotatory power and constitution of the sterols,” *The Journal of Organic Chemistry*, vol. 6, no. 2, pp. 319–330, 1941.

[11] K. Shin and T. Kuboyama, “A generalization of haussler’s convolution kernel: mapping kernel,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 944–951.

[12] B. Gaüzère, “Application des méthodes à noyaux sur graphes pour la prédiction des propriétés des molécules.” Ph.D. dissertation, Université de Caen, 2013.

[13] M. Varma and B. R. Babu, “More generality in efficient multiple kernel learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1065–1072.