Taking into account interaction between stereocenters in a graph kernel framework

Pierre-Anthony Grenier†, Luc Brun†, and Didier Villemin‡

†GREYC UMR CNRS 6072, ‡LCMT UMR CNRS 6507,
Caen, France
{pierre-anthony.grenier,luc.brun,didier.villemin}@ensicaen.fr,

Abstract. An important field of chemoinformatics consists in the prediction of molecule's properties, and within this field, graph kernels constitute a powerful framework thanks to their ability to combine a natural encoding of molecules by graphs, with classical statistical tools. 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. In this report, ordered graphs are introduced in order to represent those molecules. Then the stereoisomerism property of each atom of a molecule is encoded by a local ordered subgraph. Finally a graph of interactions between those local subgraphs is constructed.

Keywords: Graph kernel, Chemoinformatics, Stereoisomerism.

1 Introduction

The prediction of molecule's properties through Quantitative Structure Activity (resp. Property) Relationships are two active research subfields of chemoinformatics named QSAR and QSPR. Methods of those fields are based on the similarity principle: two structurally similar molecules should have similar properties.

One common method to predict chemical properties consists to design a vector of descriptors from a molecule and use statistical machine learning algorithms to predict molecule's properties. Such methods [4, 8], can use structural information, physical properties or biological activities in order to compute vectors of descriptors. However, such an approach requires to either select a random set of predefined descriptors (before a variable selection step) or to use an heuristic definition of appropriate descriptors by a chemical expert. In both cases, the transformation of the graph into a finite vector of features induces a loss of information.

Another approach consists to encode a molecule by a graph, and use it to predict properties. A molecular graph is a labeled simple graph $G = (V, E, \mu, \nu)$ representing a molecule. The unlabeled graph (V, E) encodes the structure of the molecule, each node $v \in V$ encoding an atom and each edge $e = (v, w) \in E$

a bond between two atoms. The labelling function μ associates to each vertex $v \in V$ a label $\mu(v)$ encoding the nature of the atom and the function ν associates to each edge e a type of bond $\nu(e)$ (single, double, triple or aromatic). Note that for the rest of this report, we denote the neighborhood of a vertex $v \in V$ by $N(v)$:

$$\forall v \in V,\ N(v) = \{u \in V \mid (v,u) \in E\}$$

Note that v does not belong to $N(v)$.

Several methods based on graph theory use this representation to predict molecular properties. One approach consists to search subgraphs with a large difference of frequencies between a set of positive and a set of negative examples [15]. Another approach consists to encode each class of molecules by a graph prototype and to measure the structural similarity between each prototype and an input molecule [2]. However, these methods can not be easily combined with machine learning algorithms. Conversely graph kernel methods can be coupled to machine learning algorithms provided that the kernel is definite positive. Let $\mathcal{G}$ be the set of all graphs. A definite positive kernel is a symmetric function $k : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ such that:

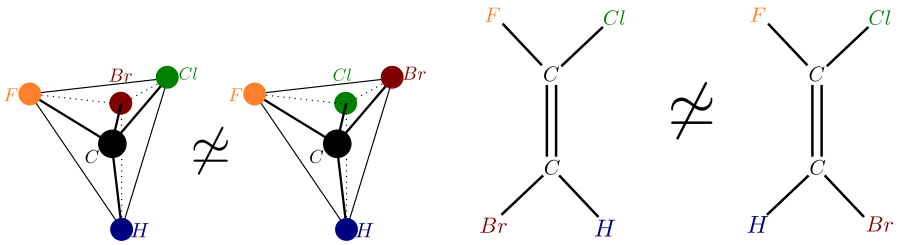
$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(G_i, G_j) \geq 0 \text{ where } n > 0,\ G_1, \ldots, G_n \in \mathcal{G},\ c_1, \ldots, c_n \in \mathbb{R}$$

Such a definite positive kernel corresponds to a scalar product between two vectors $\psi(G)$ and $\psi(G')$ in an Hilbert space induced by the kernel.

A large family of graph kernel methods, associate a bag of patterns to each graph, and define the kernel value from a measure of similarity between those bags [13, 14, 7]. In [13] a graph kernel is defined as a measure of similarity between sets of walks extracted from each graph. But those walks are linear features and thus have limited expressiveness. An infinite set of tree patterns is used in [14] to define kernels. However, the similarity between two graphs is based on an implicit enumeration of their common tree patterns which does not allow to readily analyze the influence of a pattern on the prediction. Finally [7] is based on an explicit enumeration of patterns. All subtrees of a labeled graph up to size 6, called treelets are enumerated.

However, some molecules may have a same molecular formula, a same molecular graph but a different relative positioning of their atoms. Such molecules are said to be stereoisomers. Different stereoisomers may be associated to different properties. However, usual graph kernels based on the molecular graph representation are not able to capture any dissimilarity between these molecules encoded by a same graph. From a more local point of view, an atom or two connected atoms are called stereocenters if a permutation of the positions of two atoms belonging to the union of their neighborhoods produces a different stereoisomer.

One example of stereocenter is an asymmetric carbon, which is a carbon atom with four different neighbors. We can represent each of its neighbors on a summit of a tetrahedron. If we permute two of the atoms, we obtain a different spatial configuration and hence an alternative stereoisomer (Figure 1a).



(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.

Two carbons, connected by a double bond, can also define stereoisomers (Figure 1b). Indeed, on the left side of Figure 1b fluorine (F) and iodine(I) atoms are located on the same side of the double bond while they are located on opposite sides on the stereoisomer represented on the right. In this case both carbon atoms of the double bond correspond to a stereocenter.

According to chemical experts [12], within molecules currently used in chemistry, 98% of stereocenters correspond either to asymmetric carbons (Figure 1a) or to couples of two carbons adjacent through a double bond (Figure 1b). We thus restrict the present report to such cases.

To distinguish those configurations, we introduce the two following subsets of the set of vertices V of a molecular graph:

Definition 1. Potential Asymmetric Carbons

Let us denote V_{PAC} the subset of V containing all vertices encoding atoms of carbon with four neighbors:

$$V_{PAC} = \{v \in V \mid \mu(v) = \text{'}C\text{'} \text{ and } |N(v)| = 4\}$$

Since being an atom with four neighbors is a necessary condition to define an asymmetric carbon, the set V_{PAC} contains all vertices which may encode such atoms.

Definition 2. Set of double-bonds connecting carbon atoms

The subset of V containing all atoms of carbon which share a double bond with another carbon is noted V_{DB}:

$$V_{DB} = \left\{ v \in V \mid \exists e(v,w) \in E, \nu(e) = 2, \begin{pmatrix} |N(v)| = |N(w)| = 3 \\ \text{and} \\ \mu(v) \quad = \mu(w) \quad = \text{'}C\text{'} \end{pmatrix} \right\}$$

An atom of carbon with two double bounds must have a degree equal to two. Hence, each vertex v belonging to V_{DB} is incident to a single double bound and we denote $n_=(v)$ the other carbon connected by this double bond. Note that $n_=(v) \in V_{DB}$.

Brown et al. [1] have proposed to incorporate this information through an extension of the tree-pattern kernel [14]. In this method, similarity between molecules are deduced from the number of common tree-patterns between two molecules. These patterns take into account the configuration around stereocenters. 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 which deteriorates the prediction. When several stereocenters are close to each other, one pattern may encode all of them. However the size of patterns are limited, so in some cases the influence of a permutation around stereocenters may not be detected by patterns containing them.

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 [11] to characterized locally a stereocenter by a subgraph, big enough to highlights the influence of each permutation of neighbors of this stereocenter. We then proposed a kernel based on those subgraphs.

One drawback of our previous approach is that each subgraph, and thus each stereocenter, are considered independently.

In the next section we present an encoding of molecules distinguishing stereoisomers, which was introduced in our previous report [9]. In Section 3 we present the construction of a subgraph, which allows to characterizes locally a stereocenter, introduced in [11]. Then in Section 4 we present a method which take into account interactions between the different subgraphs which characterize stereocenters. Finally, we demonstrate the validity of our kernel through experiments in Section 5.

2 Encoding of stereoisomers

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 1a, 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 1b 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 have introduced in [9] ordered graphs:

Definition 3. Ordered Graphs

An ordered graph $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ is a molecular graph $\widehat{G}$ and a function ord which maps each vertex v belonging to a subset V_{ord} of V to an ordered list of its neighbors:

$$ord \begin{cases} V_{ord} \rightarrow & V^* \\ v & \mapsto v_1 \dots v_n \end{cases}$$

where $N(v) = \{v_1, \dots, v_n\}$ denotes the neighborhood of v.

V_{ord} is defined as $V_{PAC} \cup V_{DB}$. The function ord is defined as follows for each vertex $v \in V_{ord}$:

- If $v \in V_{PAC}$:
  We set randomly one of its neighbor v_1 at the first position. The three other neighbors of v are ordered such that if we look at v from v_1, the three remaining neighbors are ordered clockwise. One of the three orders (defined up to circular permutations) fulfilling this condition is chosen randomly (Figure 2a).
- If $v \in V_{DB}$:
  Let us consider $w = n_{=}(v)$ and the two neighborhoods $N(v) = \{w, a, b\}$ and $N(w) = \{v, c, d\}$. The order on the neighborhood of v is set as $ord(v) = w, a, b$ and the order on w's neighborhood is set as $ord(w) = v, c, d$, whereby a, b, c, d are traversed clockwise when turning around the double bond for a given plane embedding (Figure 2b).

The set of ordered graph is denoted $\mathcal{OG}$.

Lemma 1. *Let* $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ *and* $G' = (\widehat{G'} = (V', E', \mu', \nu'), ord')$ *be two ordered graph such that it exists* f *in Isom*$(\widehat{G}, \widehat{G'})$.

Then $f(V_{ord}) = V'_{ord}$

Proof. By Definition 1, $v \in V_{PAC} \Leftrightarrow \mu(v) =$ 'C' and $|V(v)| = 4$.

As f is an isomorphism we have $\mu(v) =$ 'C' and $|V(v)| = 4 \Leftrightarrow \mu(f(v)) =$ 'C' and $|V(f(v))| = 4$.

Then again by Definition 1, $v \in V_{PAC} \Leftrightarrow f(v) \in V'_{PAC}$.

By Definition 2, $v \in V_{DB} \Leftrightarrow \exists e(v, w) \in E$, $\nu(e) = 2$, $|V(v)| = |V(w)| = 3$, and $\mu(v) = \mu(w) =$ 'C'.

As f is an isomorphism we have $\exists e(v, w) \in E$, $\nu(e) = 2$, $|V(v)| = |V(w)| = 3$, and $\mu(v) = \mu(w) =$ 'C' $\Leftrightarrow \exists e'(f(v), f(w)) \in E$, $\nu(e') = 2$, $|V(f(v))| = |V(f(w))| = 3$, and $\mu(f(v)) = \mu(f(w)) =$ 'C'.

Then again by Definition 2, $v \in V_{DB} \Leftrightarrow f(v) \in V'_{DB}$.

By Definition 3, we have $V_{ord} = V_{PAC} \cup V_{DB}$, so $v \in V_{ord} \Leftrightarrow f(v) \in V'_{ord}$.

Thus $f(V_{ord}) = V'_{ord}$.

□

Definition 4. Isomorphism between Ordered Graphs

Two ordered graphs G and G' are isomorphic ($G \underset{o}{\simeq} G'$) if there exists an isomorphism f between their respective molecular graphs $\widehat{G}$ and $\widehat{G'}$ which respect the order around each vertex:

$$G \underset{o}{\simeq} G' \Leftrightarrow \exists f \in \mathrm{Isom}(\widehat{G}, \widehat{G'}) \ s.t. \\ \forall v \in V_{ord} \text{ with } ord(v) = v_1 \dots v_n,\ ord'(f(v)) = f(v_1) \dots f(v_n)$$

where $N(v) = \{v_1, \dots, v_n\}$ denotes the neighborhood of v. Notice that by Lemma 1, if $v \in V_{ord}$ then $f(v) \in V'_{ord}$, so $ord'(f(v))$ is always defined.

In this case, f is called an ordered isomorphism between G and G', and we denote $\mathrm{IsomOrd}(G, G') \subset \mathrm{Isom}(\widehat{G}, \widehat{G'})$ the set of ordered isomorphism between G and G'.

As we have to make some arbitrary choice to define an order (Definition 3), a spatial configuration of atoms may be encoded by several equivalent orders. We thus have introduced in [9] the notion of re-ordering function, which associates to each vertex of an ordered structured object a permutation on its neighborhood.

Definition 5. Re-ordering functions

A re-ordering function σ on an ordered graph $G = (\widehat{G} = (V, E, \mu, \nu), ord)$, associates to each vertex $v \in V_{ord}$ a permutation φ_v on $\{1, \dots, |N(v)|\}$.

$$\sigma \begin{cases} V_{ord} \rightarrow & \mathcal{P} \\ v \ \mapsto & \varphi_v \in \Pi_{|N(v)|} \end{cases}$$

where Π_n is the group of permutations of n elements and $\mathcal{P}$ is the union of Π_n for all $n \in \mathbb{N}$.

Application of a re-ordering function on an ordered graph provides a new ordered structured graph defined as follows:

Definition 6. Re-ordered structured objects

Let $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ denotes an ordered graph, $\sigma(G) = (\widehat{G}, ord_\sigma)$ is defined as the ordered graph obtained after applying the re-ordering function σ on the ordered list of neighbours of each vertex:

$$\forall v \in V_{ord} \ s.t. \ \begin{pmatrix} ord(v) = v_1, \dots, v_n \\ \text{and} \\ \sigma(v) = \varphi_v, \end{pmatrix} \ ord_\sigma(v) = v_{\varphi_v(1)}, \dots, v_{\varphi_v(n)}$$

Re-ordering functions previously defined may apply any re-ordering on a structured object hence removing the notion of order on these objects. In order to obtain a useful notion of re-ordering, we define a set of specific re-ordering functions.

Definition 7. Set of re-ordering functions

The set Σ of re-ordering functions contains all the re-ordering functions σ such that:

- For each v in V_{PAC}, $\sigma(v)$ is an even permutation:

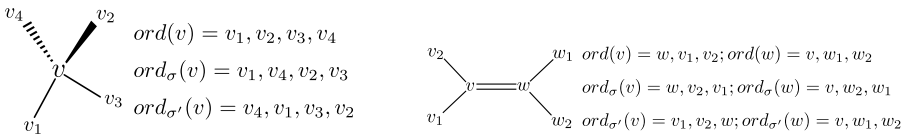
$$\forall v \in V_{PAC},\ \epsilon(\sigma(v)) = 1.$$

- For each v in V_{DB}, $\sigma(v)$ and $\sigma(n_{=}(v))$ have the same parity:

$$\forall v \in V_{DB},\ \epsilon(\sigma(v)) = \epsilon(\sigma(n_{=}(v)))$$

where ϵ denotes the signature of a permutation.

With this set we can define a notion of equivalent orders, such that two identical molecules will be represented by ordered graphs of equivalent orders.



(a) Element of V_{PAC} (b) Two elements of V_{DB}

Fig. 2: Example of elements of V_{PAC} and of V_{DB} with their ordered list (top) and the ordered lists obtained using two permutations σ and σ'

Definition 8. Equivalent orders

Let us consider two ordered graphs $G = (\widehat{G}, ord)$ and $G' = (\widehat{G'}, ord')$. These graphs are said to be equivalent $G \underset{\Sigma}{\simeq} G'$ according to the set of re-ordering functions Σ if:

$$\exists \sigma \in \Sigma\ s.t\ \sigma(G) \underset{o}{\simeq} G' \tag{1}$$

In other word, we consider that two ordered graphs are equivalent if, up to a valid re-ordering σ we can establish an ordered graph isomorphism f between them. In that case the ordered isomorphism f is called an equivalent ordered isomorphism through σ between G and G'. We denote by $\text{IsomEqOrd}(G, G')$ the set of equivalent ordered isomorphism between G and G':

$$\text{IsomEqOrd}(G, G') = \bigcup_{\sigma \in \Sigma} \text{IsomOrd}(\sigma(G), G')$$

As the set defined in Definition 7 is a valid family of re-ordering functions [9], the relationships defined in Definition 8 is an equivalence relationship [9]. We have thus now a way to encode stereoisomers.

Potentials asymmetric carbons, and double bonds between carbons, are not necessarily stereocenters. For example if the label of vertex Br of Figure 1a is replaced by Cl, both left and right molecules of Figure 1a would be identical. In the same way, if the label of the vertex F in Figure 1b 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, i.e it exists an equivalent ordered isomorphism between the graph and a permuted graph :

Definition 9. Set of Isomorphism of non-stereocenter
Let $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ be an ordered graph. Let $v \in V_{ord}$.
We denote by $\mathcal{F}_G^v$ the set of ordered isomorphism f such that :

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

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 $ord(v)$.

We define a stereo vertex as a vertex for which any permutation of two of its neighbors produces a non-equivalent ordered graph:

Definition 10. Stereo vertices
Let $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ be an ordered graph. A vertex $v \in V_{ord}$ is called a stereo vertex iff:

$$\mathcal{F}_G^v = \varnothing$$

We denotes $\mathcal{SV}(G)$ the set of stereo vertices of G.

Lemma 2. *Let us consider a graph $\widehat{G}$ and one of its automorphism f, $f \in Isom(\widehat{G}, \widehat{G})$.*
$\forall v \in V_{DB}$ s.t $f(v) = v$, we have $f(n_=(v)) = n_=(v)$.

Proof. Let $v \in V_{DB}$ with $f(v) = v$.
By Definition 2 $\exists! w \in V$ *s.t* $e = (v, w) \in E$ with $\nu(e) = 2$ and $w = n_=(v)$. Since f is an isomorphism $(f(v), f(w)) = (v, f(w)) \in E$ with $\nu((f(v), f(w))) = 2$. By Definition 2, w is the unique neighbor of v incident to an edge with a label 2, so we have $f(w) = w$. □

Proposition 1. $v \in V_{DB}$ is a stereo vertex iff $n_=(v)$ is a stereo vertex.

Proof. We consider an ordered graph $G = (\widehat{G} = (V, E, \mu, \nu), ord)$. Let us consider $v \in V_{DB}$. We denote $w = n_=(v) \in V_{DB}$.
Let us suppose that $v \notin \mathcal{SV}(G)$.
Then $\mathcal{F}_G^v \neq \varnothing$ and so, $\exists (i, j) \in \{1, 2, 3\}^2$, with $i \neq j$,
$\exists f \in \text{IsomEqOrd}(G, \tau_{i,j}^v(G))$ 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 $ord(v)$.

As $f \in \text{IsomEqOrd}(G, \tau^v_{i,j}(G)) \subset \text{Isom}(\widehat{G}, \widehat{G})$, by Lemma 2 we have $f(w) = w$.

We define $\sigma = \tau^w_{i',j'} \circ \tau^v_{i,j}$ where $(i', j') \in \{1, 2, 3\}^2$ $s.t$ $i' \neq j'$.

By Definition 7 and since $\epsilon(\sigma(v)) = \epsilon(\tau^v_{i,j}) = -1$ and $\epsilon(\sigma(w)) = \epsilon(\tau^w_{i',j'}) = -1$, σ is a valid re-ordering function.

By Definition 8, $\exists \sigma' \in \Sigma$ $s.t$ f is an ordered isomorphism between $\sigma'(G)$ and $\tau^v_{i,j}(G)$. Thus, by [9](Lemma 1), f is an ordered isomorphism between $\sigma \circ \sigma'(G)$ and $\sigma \circ \tau^v_{i,j}(G) = \tau^w_{i',j'}(G)$.

As the valid family of re-ordering function is a group [9] $\sigma \circ \sigma' \in \Sigma$, and thus $f \in \text{IsomEqOrd}(G, \tau^w_{i',j'}(G))$ with $f(w) = w$. So $\mathcal{F}^w_G \neq \varnothing$

So $v \notin \mathcal{SV}(G)$ implies that $w = n_=(v) \notin \mathcal{SV}(G)$.

Since $\forall u \in V_{DB}$, $n_=(n_=(u)) = u$, this last statement is equivalent to :

$$n_=(u) \notin \mathcal{SV}(G) \Rightarrow u \notin \mathcal{SV}(G)$$

The contrapositive of the above implication provides the expected implication. Furthermore, an additional use of the relationship $n_=(n_=(u)) = u$ provides the expected equivalence relationship.

□

As Proposition 1 shows, two carbons linked by a double bond are simultaneously stereo vertices or non-stereo vertices. So we have to consider their stereo property together. We thus introduce the following notations :

Definition 11. Set of bounded Stereo vertices

For $s \in \mathcal{SV}(G)$ we define its set $kernel(s)$ of bounded stereo vertices as :

$$kernel(s) = \begin{cases} \{s\} \text{ if } s \in V_{PAC} \\ \{s, n_=(s)\} \text{ if } s \in V_{DB} \end{cases}$$

Definition 12. Star of Stereo vertices

For $s \in \mathcal{SV}(G)$ we define its set $StereoStar(s)$:

$$StereoStar(s) = \begin{cases} N(s) \cup \{s\} \text{ if } s \in V_{PAC} \\ N(s) \cup N(n_=(s)) \text{ if } s \in V_{DB} \end{cases}$$

Definition 13. Set of neighbours of Stereo vertices

For $s \in \mathcal{SV}(G)$ we define its set of neighbour $StereoStar^*(s)$:

$$StereoStar^*(s) = \begin{cases} N(s) \text{ if } s \in V_{PAC} \\ N(s) \cup N(n_=(s) - \{s, n_=(s)\} \text{ if } s \in V_{DB} \end{cases}$$

Note that $StereoStar^*(s) = StereoStar(s) - kernel(s)$.

3 From a global to a local characterization of stereo information

Definition 10 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.

Definition 14. Subgraph characterizing a Stereo vertex

Let $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ be an ordered graph. Let S be a subgraph of G. We say that the stereo property of $s \in \mathcal{SV}(G)$ is captured by S if:

- $StereoStar(s) \subset S$.
- $\mathcal{F}_S^s = \varnothing$.

Remark 1. With the same argument than for the proof of Proposition 1, we can show that if the stereo property of $s \in \mathcal{SV}(G)$ is captured by S and $s \in V_{DB}$ then the stereo property of $n_=(s)$ is captured by S.

Thus a couple of carbons linked by a double bond only needs one subgraph to characterizes both of them.

Definition 15. Set of vertices inducing isomorphism

Let $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ be an ordered graph. Let s be a stereo vertex. Let H be a subgraph of G that do not capture the stereo property of s and such that $StereoStar(s) \subset H$.

By Definition 14, $\mathcal{F}_H^s$ is not empty. Let $f \in \mathcal{F}_H^s$.

We define $\mathcal{E}_f^H$ as the set of vertices inducing the isomorphism f in H:

$$\mathcal{E}_f^H = \{v \in V(H) \mid \exists p = (v_0, \ldots, v_q) \in H \text{ with } v_0 \in kernel(s) \text{ and } v_q = v \quad s.t.\ f(v_1) \neq v_1\} \quad (2)$$

where $(v_0, \ldots, v_q)$ denotes a path in H (Hence $v_0 \neq v_1$).

Proposition 2. For H and f defined as in Definition 15, $\mathcal{E}_f^H$ is not empty.

Proof. Let $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ be an ordered graph. Let s be a stereo vertex. Let H be a subgraph of G that do not capture the stereo property of s and such that $StereoStar(s) \subset H$.

Let us consider $f \in \mathcal{F}_H^s$.

By definition of equivalent ordered isomorphism, it exists $\sigma \in \Sigma$ such that f is an ordered isomorphism between H and $\sigma\left(\tau_{i,j}^s(H)\right)$.

– If $s \in V_{PAC}$:
By definition of ordered isomorphisms, and since $f(s) = s$, we have:

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

with $ord(s) = v_1, \dots, v_n$.
As $\sigma(s)$ is an even permutation, $\sigma(s) \circ \tau^s_{i,j}$ is an odd one. Hence it exists l in $\{1, \dots, |N(s)|\}$ such that $l \neq \sigma(s) \circ \tau^s_{i,j}(l)$.
Thus $f(v_l) \neq v_l$, and $\mathcal{E}^H_f$ is not empty as it contain at least v_l.
– If $s \in V_{DB}$:
Let us consider $w = n_=(s)$. Since $f \in \text{IsomEqOrd}(H, \tau^s_{i,j}(H)) \subset \text{Isom}(\widehat{H}, \widehat{H})$, by Lemma 2 we have $f(w) = w$.
We denote by $N(s) = \{s_1, s_2, w\}$ the neighbourhood of s and $N(w) = \{w_1, w_2, s\}$ the neighbourhood of w.
Let us suppose that $f(w_1) = w_1$.
As $f(s) = s$, we have $f(w_2) = w_2$ and since f is an ordered isomorphism between H and $\sigma\left(\tau^s_{i,j}(H)\right)$ we have $\sigma(w) = Id$ where Id is the identity permutation. Thus $\epsilon(\sigma(s)) = \epsilon(\sigma(w)) = 1$ and $\sigma(s) \circ \tau^s_{i,j}$ is odd.
Hence f defines an odd permutation on the ordered neighbour of s and $f(w) = w$, thus by definition of ordered isomorphisms, we have $f(s_1) = s_2 \neq s_1$.
Thus we have either $f(w_1) \neq w_1$ or $f(s_1) \neq s_1$, so $\mathcal{E}^H_f$ is not empty as it contains at least w_1 or s_1.

□

Definition 16. Minimal stereo subgraph

Let $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ be an ordered graph. Let $s \in \mathcal{SV}(G)$. We consider a sequence $(H^k_s)_{k \in \mathbb{N}}$ of vertex induced subgraphs of G defined such that:

– $V(H^0_s) = StereoStar(s)$
– $V(H^{k+1}_s) = V(H^k_s) \cup \bigcup_{f \in \mathcal{F}_{H^k_s}} N(\mathcal{E}^{H^k_s}_f)$.

We define S_s as:

$$S_s = \lim_{k \to +\infty} H^k_s$$

The vertex induced subgraph S_s is called the minimal stereo subraph of s. We say that s is the stereo vertex of S_s (if $s \in V_{DB}$, the stereo vertex of S_s is arbitrarily chosen between s and $n_=(s)$ as they have a same role). We denote $\mathcal{H}(G)$ the set of minimal stereo subraphs of G. Figure 3 shows one exemple of minimal stereo subraph.

Proposition 3. The sequence $(H^k_s)_{k \in \mathbb{N}}$ converges.

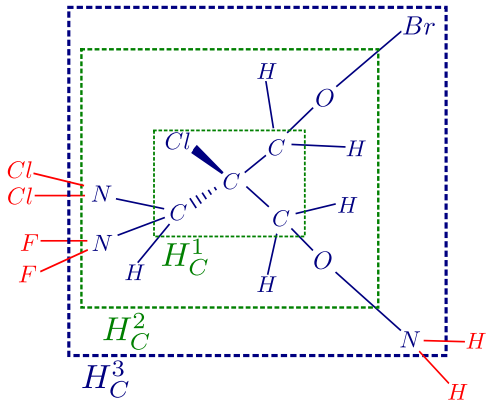


Fig. 3: An asymmetric carbon and its associated sequence $(H^k_C)_{k\in\mathbb{N}}$. Its minimal stereo subgraph is $S_C = H^3_C$.

Proof. As $(H^k_s)_{k\in\mathbb{N}}$ is a sequence of vertex induced subgraphs of G, we know that $\forall k \in \mathbb{N}\ H^k_s \trianglelefteq G$. The sequence is thus upper bounded.

By Definition 16, we have $\forall k \in \mathbb{N}\ V(H^{k+1}_s) = V(H^k_s) \cup \bigcup_{f\in\mathcal{F}_{H^k_s}} N(\mathcal{E}^{H^k_s}_f)$. Thus $(H^k_s)_{k\in\mathbb{N}}$ is an increasing sequence.

As $(H^k_s)_{k\in\mathbb{N}}$ is upper bounded and increasing, $(H^k_s)_{k\in\mathbb{N}}$ converges.

□

Remark 2. $(H^k_s)_{k\in\mathbb{N}}$ is a sequence of vertex induced subgraphs of G, so it is a discrete sequence, and thus the limit is reached: $\exists n \in \mathbb{N}\ s.t\ \forall k > n,\ H^k_s = S_s$.

Proposition 4. For any $k \in \mathbb{N}$ such that $V(H^k_s) \neq V(S_s)$, the stereo property of s is not captured by H^k_s.

Proof. We consider $k \in \mathbb{N}$ such that $V(H^k_s) \neq V(S_s)$.

As H^{k+1}_s is only constructed from H^k_s $(V(H^{k+1}_s) = V(H^k_s) \cup \bigcup_{f\in\mathcal{F}_{H^k_s}} N(\mathcal{E}^{H^k_s}_f))$, and $V(H^k_s) \neq V(S_s)$, we have $V(H^k_s) \neq V(H^{k+1}_s)$.

Thus $\bigcup_{f\in\mathcal{F}_{H^k_s}} N(\mathcal{E}^k_f) \neq \varnothing$. So it exists at least one $f \in \mathcal{F}_{H^k_s}$ such that $\mathcal{E}^k_f \neq \varnothing$ and thus $\mathcal{F}_{H^k_s} \neq \varnothing$.

Then by Definition 14 and 9 the stereo property of s is not captured by H^k_s. □

Lemma 3. *Let $G = (\widehat{G} = (V, E, \mu, \nu), ord)$ be an ordered graph. Let s be a stereo vertex. Let H be a connected vertex induced subgraph of G that do not capture the stereo property of s and such that $StereoStar(s) \subset H$. Let $f \in \mathcal{F}^s_H$.*

We denotes $cc_1 \dots cc_n$ the set of connected components obtained by removing $kernel(s)$ from H.

We define the sets of indices :

- $\mathcal{I} = \{i \mid cc_i \cap \mathcal{E}_f^H \neq \varnothing\}$.
- $\mathcal{J} = \{j \mid cc_j \cap \mathcal{E}_f^H = \varnothing\}$.

We denote $cc_{\mathcal{I}} = \bigcup_{i \in \mathcal{I}} cc_i$ *and* $cc_{\mathcal{J}} = \bigcup_{j \in \mathcal{J}} cc_j$.

Then we have :

$$cc_{\mathcal{I}} \subset \mathcal{E}_f^H. \tag{3}$$

$$\forall j \in \mathcal{J},\ cc_j \cap StereoStar^*(s) \neq \varnothing \text{ and } \forall v \in cc_j \cap StereoStar^*(s) \text{ we have } f(v) = v. \tag{4}$$

Proof. We first show that:

$$cc_{\mathcal{I}} \subset \mathcal{E}_f^H.$$

Let $i \in \mathcal{I}$. By definition of $\mathcal{I}$ we have $cc_i \cap \mathcal{E}_f^H \neq \varnothing$. We denote u one vertex such that $u \in cc_i$ and $u \in \mathcal{E}_f^H$. By definition of $\mathcal{E}_f^H$, it exists a path $p_u = (u_1, \ldots, u_q) \in H$ with $u_q = u$, $u_1 \in StereoStar^*(s)$ and $f(u_1) \neq u_1$. As cc_i is a connected component of $H - kernel(s)$, we have that $p_u \in cc_i$ and so $u_1 \in cc_i$.

As cc_i is a connected component $\forall v \in cc_i$, $\exists p = (u_1, v_1, \ldots, v) \in cc_i$. So $v \in \mathcal{E}_f^H$ and $cc_i \subset \mathcal{E}_f^H$.

We now show that:

$$\forall j \in \mathcal{J},\ cc_j \cap StereoStar^*(s) \neq \varnothing \text{ and } \forall v \in cc_j \cap StereoStar^*(s) \text{ we have } f(v) = v.$$

H is connected, thus for each $u \in H$, $\exists p = (u_0, \ldots, u_q) \in H$ with $u_q = u$ and $u_0 \in kernel(s)$. So $\forall u \in cc_j$, $\exists p = (u_1, \ldots, u_q) \in cc_j$ with $u_q = u$ and $u_1 \in StereoStar^*(s)$. Hence $u_1 \in cc_j$ and $cc_j \cap StereoStar^*(s) \neq \varnothing$

Let us consider $v \in cc_j \cap StereoStar^*(s)$, if we suppose that $f(v) \neq v$ then $v \in \mathcal{E}_f^H$ and thus $cc_j \cap \mathcal{E}_f^H = \varnothing$ is false. So $\forall v \in cc_j$ *s.t* $v \in StereoStar^*(s)$ we have $f(v) = v$.

□

Remark 3. By definition of $cc_{\mathcal{I}}$ and $cc_{\mathcal{J}}$ we have :

- $V(H) = cc_{\mathcal{I}} \cup cc_{\mathcal{J}} \cup kernel(s)$
- $V = V(G - H) \cup cc_{\mathcal{I}} \cup cc_{\mathcal{J}} \cup kernel(s)$

Lemma 4. *With the same hypothesis and notations than in Lemma 3 we have:*

$$f(cc_{\mathcal{I}}) = cc_{\mathcal{I}}$$

Proof. Let $v \in cc_i$ with $i \in \mathcal{I}$.

Thus $\exists p = (v_1, \ldots, v_q) \in H$ with $v_q = v$, $v_1 \in StereoStar^*(s)$ and $f(v_1) \neq v_1$. So the sequence $(f(v_1), \ldots, f(v_q))$ is also a path of H.

We denote $\tilde{v_1} = f(v_1)$.

As $v_1 \in StereoStar^*(s)$, $\tilde{v_1} \in StereoStar^*(s)$. As $f(v_1) \neq v_1$ and f is bijective, $f(\tilde{v_1}) = f(f(v_1)) \neq f(v_1) = \tilde{v_1}$.

In conclusion we have $\exists p' = (\tilde{v_1}, \ldots, f(v_q)) \in H$ with $v_q = v$, $\tilde{v_1} \in StereoStar^*(s)$ and $f(\tilde{v_1}) \neq \tilde{v_1}$, so $f(v) \in cc_{i'}$ with $i' \in \mathcal{I}$.

So $f(cc_{\mathcal{I}}) \subset cc_{\mathcal{I}}$.

As f is bijective we have $|f(cc_{\mathcal{I}})| = |cc_{\mathcal{I}}|$. So $f(cc_{\mathcal{I}}) = cc_{\mathcal{I}}$.

□

Lemma 5. *Using the same hypothesis and notations than in Lemma 3,* $\forall f \in \mathcal{F}_H^s$ *we have* $N(\mathcal{E}_f^H) \not\subset V(H)$.

Proof. Let us supposed $N(\mathcal{E}_f^H) \subset V(H)$.

To obtain a contradiction we will construct an equivalent ordered isomorphism between G and $\tau_{i,j}^s(G)$. To prove that the function we will construct is an ordered isomorphism between G and $\tau_{i,j}^s(G)$ we will need some properties about the cc_i and cc_j defined previously which are consequences of our assumption $N(\mathcal{E}_f^H) \subset V(H)$.

We need to prove that we have :

$$\forall v \in V(G-H),\ \forall p = (v_0, \ldots, v_q) \text{ with } v_0 = v,\ v_q \in H \text{ and } kernel(s) \cap p = \varnothing, \text{ we have } v_q \in cc_{\mathcal{J}}. \quad (5)$$

Let us suppose that $\exists v \in V(G-H)$, $\exists p = (v_0, \ldots, v_q)$ with $v_0 = v$, $v_q \in H$ and $kernel(s) \cap p = \varnothing$ such that $v_q \in cc_i$ with $i \in \mathcal{I}$.

Let us denote $v_r \in p$ with $r \in \{0, \ldots, q-1\}$ one vertex such that $v_r \notin V(H)$ and $v_{r+1} \in cc_i$. Such a vertex exists as $v_0 \notin V(H)$ and $v_q \in cc_i$.

As $v_{r+1} \in cc_i$, $v_{r+1} \in \mathcal{E}_f^H$ by (3) of Lemma 3. Thus $v_r \in N(\mathcal{E}_f^H)$ and $v_r \notin V(H)$. As we have supposed $N(\mathcal{E}_f^H) \subset V(H)$ we have a contradiction, so $v_q \in cc_{\mathcal{J}}$.

Equation (5) and Remark 3 implies that:

$$\forall v \in V(G-H),\ N(v) \subset V(G-H) \cup cc_{\mathcal{J}} \quad (6)$$

$$\forall v \in cc_{\mathcal{J}},\ N(v) \subset V(G-H) \cup cc_{\mathcal{J}} \cup kernel(s) \quad (7)$$

$$\forall v \in cc_{\mathcal{I}},\ N(v) \subset cc_{\mathcal{I}} \cup kernel(s) \quad (8)$$

Let us now define a function g such that:

$$\forall v \in V, g(v) = \begin{cases} f(v) \text{ if } v \in cc_{\mathcal{I}} \\ v \text{ if } v \in cc_{\mathcal{J}} \cup V(G-H) \\ v = f(v) \text{ if } v \in kernel(s) \end{cases}$$

We want to prove that $g \in \text{IsomEqOrd}(G, \tau_{i,j}^s(G))$.

First, we have to prove that g is bijective.

By Lemma 4 we have $f(cc_{\mathcal{I}}) = cc_{\mathcal{I}}$ so $g(cc_{\mathcal{I}}) = cc_{\mathcal{I}}$.

We have $g_{|cc_{\mathcal{I}}} = f$ is bijective, $g_{|G-cc_{\mathcal{I}}} = Id$ is bijective and $g(cc_{\mathcal{I}}) = cc_{\mathcal{I}}$, so g is bijective.

We now prove that $g \in \mathrm{Isom}(\widehat{G}, \widehat{G}) = \mathrm{Isom}(\widehat{G}, \widehat{\tau^s_{i,j}(G)})$.

Let $e = (u, v) \in E$:

- If $u \in cc_{\mathcal{I}}$.
  We have $g(u) = f(u)$ and $v \in cc_{\mathcal{I}}$ or $v \in kernel(s)$ by (8). In both cases we have $g(v) = f(v)$. As $f \in \mathrm{IsomEqOrd}(H, \tau^s_{i,j}(H))$, $(g(u), g(v)) = (f(u), f(v)) \in E$.
- If $u \in cc_{\mathcal{J}}$ or $u \in V(G - H)$.
  We have $g(u) = u$ and $v \in V(G - H)$, $v \in cc_{\mathcal{J}}$ or $v \in kernel(s)$ by (6) and (7). In each case we have $g(v) = v$. Thus $(g(u), g(v)) = (u, v) \in E$.
- If $u \in kernel(s)$.
  We have $g(u) = f(u)$ and $v \in StereoStar^*(s)$ or $v \in kernel(s)$. If $v \in cc_{\mathcal{I}}$, then $f(v) = g(v)$. However if $v \in cc_{\mathcal{J}}$, then $g(v) = v$. But by (4) of Lemma 3 we have $v = f(v)$, so $g(v) = v = f(v)$. Finally if $v \in kernel(s)$, $g(v) = v = f(v)$.
  So, in each case we have $g(v) = f(v)$. As $f \in \mathrm{IsomEqOrd}(H, \tau^s_{i,j}(H))$, $(g(u), g(v)) = (f(u), f(v)) \in E$.

In each case we have either $(g(u), g(v)) = (f(u), f(v))$ or $(g(u), g(v)) = (u, v)$, thus :

$$e = (u, v) \in E \Rightarrow$$
$$e' = (g(u), g(v)) \in E,\ \nu(e) = \nu(e'),\ \mu(g(u)) = \mu(u) \text{ and } \mu(g(v)) = \mu(v). \quad (9)$$

We define $\tilde{g}$ such that:

$$\tilde{g} \begin{cases} V \times V \to & V \times V \\ (u, v) \to & (g(u), g(v)) \end{cases}$$

As g is bijective, $\tilde{g}$ is bijective.

By (9) we have $\tilde{g}(E) \subset E$, and as $\tilde{g}$ is bijective we have $\tilde{g}(E) = E$.

So $\forall (u, v) \in V \times V$, such that $\tilde{g}(u, v) = (g(u), g(v)) \in E$ we have $(u, v) \in E$. With (9) we can conclude that :

$$e = (g(u), g(v)) \in E \Rightarrow$$
$$e' = (u, v) \in E,\ \nu(e) = \nu(e'),\ \mu(g(u)) = \mu(u) \text{ and } \mu(g(v)) = \mu(v). \quad (10)$$

By (9) and (10) we have $g \in \mathrm{Isom}(\widehat{G}, \widehat{G}) = \mathrm{Isom}(\widehat{G}, \widehat{\tau^s_{i,j}(G)})$.

We finally have to prove that $\exists \sigma \in \Sigma$ such that $g \in \mathrm{IsomOrd}(\sigma(G), \tau^s_{i,j}(G))$.

$f \in \mathrm{IsomEqOrd}(H, \tau^s_{i,j}(H))$ so by Definition 8 $\exists \sigma \in \Sigma,\ \sigma(H) \underset{o}{\simeq} \tau^s_{i,j}(H)$.

We denote σ' the re-ordering function such that:

$$\forall v \in V_{ord}, \sigma'(v) = \begin{cases} \sigma(v) \text{ if } v \in cc_{\mathcal{I}} \cup kernel(s) \\ Id_n \text{ if } v \in cc_{\mathcal{J}} \cup V(G-H) \end{cases}$$

where Id_n is the identity permutation of n elements and n is the degree of v.

Let $v \in V_{PAC}$.

If $v \in kernel(s) \cup cc_{\mathcal{I}}$ then $\sigma'(v) = \sigma(v)$. $\sigma \in \Sigma$ so by Definition 7 $\sigma(v)$ is even. If $v \in V(G-H) \cup cc_{\mathcal{J}}$ then $\sigma'(v) = Id_n$. The identity is even, so $\forall v \in V_{PAC}$, $\sigma'(v)$ is even.

Let $v \in V_{DB}$.

- By Definition 11 $v \in kernel(s)$ iff $n_=(v) \in kernel(s)$.
  So, if $v \in kernel(s)$, then $\sigma'(v) = \sigma(v)$ and $\sigma'(n_=(v)) = \sigma(n_=(v))$.
  $\sigma \in \Sigma$ so by Definition 7 $\sigma(v)$ and $\sigma(n_=(v))$ have a same parity.
- If $v \in cc_{\mathcal{I}}$ then by (8) $n_=(v) \in cc_{\mathcal{I}} \cup kernel(s)$.
  As $n_=(v) \in kernel(s)$ iff $v \in kernel(s)$ (Definition 11) and $v \in cc_{\mathcal{I}}$, $n_=(v) \notin kernel(s)$ and so $n_=(v) \in cc_{\mathcal{I}}$.
  Thus $\sigma'(v) = \sigma(v)$ and $\sigma'(n_=(v)) = \sigma(n_=(v))$.
  $\sigma \in \Sigma$ so by Definition 7 $\sigma(v)$ and $\sigma(n_=(v))$ have a same parity.
- Finally if $v \in V(G-H) \cup cc_{\mathcal{J}}$ then by (6) and (7) $n_=(v) \in V(G-H) \cup cc_{\mathcal{J}} \cup kernel(s)$.
  As $n_=(v) \in kernel(s)$ iff $v \in kernel(s)$ (Definition 11) and $v \in V(G-H) \cup cc_{\mathcal{J}}$, $n_=(v) \notin kernel(s)$ and so $n_=(v) \in V(G-H) \cup cc_{\mathcal{J}}$.
  So $\sigma'(v) = Id_n$ and $\sigma'(n_=(v)) = Id_n$ have a same parity.

Using Definition 7, we deduce from the previous considerations that $\sigma' \in \Sigma$.

Let us now show that $g \in \text{IsomOrd}(\sigma'(G), \tau^s_{i,j}(G))$. We denote by ord' the order in $\tau^s_{i,j}(G)$. We stress here that $\forall v \in V_{ord} - \{s\}$, $ord'(v) = ord(v)$.

Let $v \in V_{ord}$ with $ord_{\sigma'}(v) = v_1, \ldots, v_n$:

- If $v \in cc_{\mathcal{I}}$.
  We have $g(v) = f(v)$. We know that $\sigma(H) \underset{o}{\simeq} \tau^s_{i,j}(H)$ and $v \in H$, so $ord'(g(v)) = ord'(f(v)) = f(v_1), \ldots, f(v_n)$.
  By (8) we know that $\forall k \in \{1, \ldots, n\}$ we have $v_k \in cc_{\mathcal{I}}$ or $v_k \in kernel(s)$. In both cases we have $f(v_k) = g(v_k)$ and thus $ord'(g(v)) = g(v_1), \ldots, g(v_n)$.
- If $v \in cc_{\mathcal{J}}$ or $v \in V(G-H)$.
  We have $g(v) = v$. So $ord'(g(v)) = ord'(v)$. As $v \in V_{ord} - \{s\}$, $ord'(v) = ord(v)$, and thus $ord'(g(v)) = ord(v)$. As $\sigma'(v) = Id_n$, we have $ord(v) = ord_{\sigma'}(v) = v_1, \ldots, v_n$. So $ord'(g(v)) = v_1, \ldots, v_n$.
  By (7) and (8) we know that $\forall k \in \{1, \ldots, n\}$ we have $v_k \in V(G-H)$, $v_k \in cc_{\mathcal{J}}$ or $v_k \in kernel(s)$. So $v_k = g(v_k)$ and thus $ord'(g(v)) = g(v_1), \ldots, g(v_n)$.
- If $v \in kernel(s)$.
  We have $g(v) = f(v)$. We know that $\sigma(H) \underset{o}{\simeq} \tau^s_{i,j}(H)$ and $v \in H$, so $ord'(g(v)) = ord'(f(v)) = f(v_1), \ldots, f(v_n)$.
  Let $k \in \{1, \ldots, n\}$. If $v_k \in cc_{\mathcal{I}}$, then $f(v_k) = g(v_k)$. However if $v_k \in cc_{\mathcal{J}}$, then $g(v_k) = v_k$. But by (4) we have $v_k = f(v_k)$, so $g(v_k) = v_k = f(v_k)$. Thus $ord'(g(v)) = g(v_1), \ldots, g(v_n)$

In each case we have $ord'(g(v)) = g(v_1), \ldots, g(v_n)$, so $\sigma'(G) \underset{o}{\simeq} \tau^s_{i,j}(G)$.

Thus $g \in \text{IsomEqOrd}(G, \tau^s_{i,j}(G))$ and $g(s) = s$. This is not possible since $s \in \mathcal{SV}(G)$, thus $N(\mathcal{E}^H_f) \not\subset V(H)$.

□

Theorem 1. *The stereo property of s is captured by S_s.*

Proof. We denote $n \in \mathbb{N}$ an integer such that $H^n_s = S_s$ (Remark 2).

The sequence is initialized by $V(H^0_s) = StereoStar(s)$. As the sequence is increasing we have $V(H^0_s) \subset V(H^n_s)$, so the first condition of Definition 14 is true.

We thus have to prove that $\mathcal{F}^s_{H^n_s} = \varnothing$.

Let us suppose that $\exists (i,j) \in \{1, \ldots, |N(s)|\}^2$ with $i \neq j$,
$\exists f \in \text{IsomEqOrd}(H^n_s, \tau^s_{i,j}(H^n_s))$ with $f(s) = s$.

Thus by Lemma 5 we have $N(\mathcal{E}^{H^n_s}_f) \not\subset V(H^n_s)$. So $V(H^{n+1}_s) = V(H^n_s) \cup \bigcup_{f \in \mathcal{F}_{H^k_s}} N(\mathcal{E}^{H^k_s}_f) \neq V(H^n_s)$.

This is in contradiction with the fact that $H^n_s = S_s = \lim_{k \to +\infty} H^k_s$, so the stereo property of s is captured by $H^n_s = S_s$.

□

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 subgraphs, and to test it efficiently, we transform our minimal stereo subgraphs S into codes c_S thanks to the method described in [17].

4 Interactions between stereo vertices

In the previous section we have defined a way to construct an oriented subgraph which characterizes a stereocenter. We may use the set of subgraphs, associated to each stereocenter of a molecule, to compare molecules. However two stereocenter may not have the same influence on a property if they are close from each other or far from each other in a molecule. In the same way, two same minimal stereo subgraph may not have a same influence on a property if they have different surroundings. We now propose to construct some new graphs, based on the set of minimal stereo subgraphs, to encode more information about those subgraphs.

4.1 Graphs of interaction

To represent the interactions between stereo vertices we define different functions of interactions, which encode different degrees of information about the interactions between stereo vertices:

Definition 17. Functions of interactions

Let $G = (G_m = (V, E, \mu, \nu), ord)$ an ordered graph and $\mathcal{H}(G)$ its set of minimal stereo subgraphs.

Functions of interactions are defined according to a sequence of conditions $(c_0, \ldots, c_n)$. These conditions are increasingly constraining:

$$\forall i \in \{1, \ldots, n-1\}\, c_{i+1} \Rightarrow c_i \text{ and } c_0 = \neg c_1$$

Let H_1 and H_2 be two minimal stereo subgraphs, such that s_1 is the stereo vertex of H_1 and s_2 is the stereo vertex of H_2. The value $F(H_1, H_2)$ is obtained by taking the maximum index j of conditions c_j which represents the strongest interaction between H_1 and H_2:

$$F(H_1, H_2) = \max\{j \in \{0, \ldots, n\} \mid c_j\}$$

We consider 4 sequences of conditions defining 4 functions of interactions F_i :

- F_1 is defined by using $\begin{cases} c_1 : H_1 \bigcap H_2 \neq \emptyset \\ c_2 : kernel(s_1) \subset H_2 \\ c_3 : StereoStar(s_1) \subset H_2 \\ c_4 : H_1 \subset H_2 \end{cases}$
- F_2 is defined by using $\begin{cases} c_1 : H_1 \bigcap H_2 \neq \emptyset \\ c_2 : StereoStar(s_1) \subset H_2 \\ c_3 : H_1 \subset H_2 \end{cases}$
- F_3 is defined by using $\begin{cases} c_1 : kernel(s_1) \subset H_2 \\ c_2 : StereoStar(s_1) \subset H_2 \\ c_3 : H_1 \subset H_2 \end{cases}$
- F_4 is defined by using $\begin{cases} c_1 : StereoStar(s_1) \subset H_2 \\ c_2 : H_1 \subset H_2 \end{cases}$

Note that the F_i is not symmetric.

We define thanks to those functions, 4 graphs of interactions G_i where each vertex $v \in V_i$ represent a minimal stereo subgraph and each edge encode the interaction between two minimal stereo subgraphs.

Definition 18. Directed Graph of interactions

A directed graph of interactions $G_i = (V_i, A_i, \mu_i, \nu_i)$ is a graph constructed from an ordered graph $G = (G_m = (V, E, \mu, \nu), ord)$ such that :

- $\forall u \in V_i$, $\exists! H(u) \in \mathcal{H}(G)$.
- $\forall u \in V_i$, $\mu_i(u) = c_{H(u)}$, where c_H is the code defined in [17] (Section 3).
- $\exists a = (u_1, u_2) \in A_i \iff F_i(H(u_1), H(u_2)) \neq 0$.
- $\forall a = (u_1, u_2) \in A_i$, $\nu_i(a) = F_i(H(u_1), H(u_2))$.

where F_i is one of the function defined in Definition 17.

As very few nodes have the same label in the directed graph of interactions, the direction of edges does not provide a lot of information. We thus define undirected graphs of interactions, which encode not a lot less infromation than directed graphs of interactions, but which is simpler.

Definition 19. Graph of interactions

A graph of interactions $G_i = (V_i, E_i, \mu_i, \nu_i)$ is a graph constructed from an ordered graph $G = (G_m = (V, E, \mu, \nu), ord)$ such that :

- $\forall u \in V_i$, $\exists! H(u) \in \mathcal{H}(G)$.
- $\forall u \in V_i$, $\mu_i(u) = c_{H(u)}$, where c_H is the code defined in [17] (Section 3).
- $\exists e = (u_1, u_2) \in E_i \iff F_i(H(u_1), H(u_2)) \neq 0$ or $F_i(H(u_2), H(u_1)) \neq 0$.
- $\forall e = (u_1, u_2) \in E_i$, $\nu_i(e) = \min(F_i(H(u_1), H(u_2)), F_i(H(u_2), H(u_1))) \odot \max(F_i(H(u_1), H(u_2)), F_i(H(u_2), H(u_1)))$.

where $\odot$ denotes the concatenation and F_i is one of the function defined in Definition 17.

Figure 4 show all graphs of interactions we can construct from an ordered graph by taking the four different functions of interactions.

The first graph G_1 is constructed by taking 4 different types of interaction. However we may suppose that some of those type of interaction are less relevant than the other.

Indeed, a vertex s_1 is a stereo vertex because of the relative positioning of its neighbour. So we may suppose that, if a stereo vertex is present in a stereo subgraph ($kernel(s_1) \subset H_2$), but not its neighbourhood ($StereoStar(s_1) \not\subset H_2$), the stereo vertex may have a similar influence in H_2 than a non-stereo vertex. G_2 is thus constructed without taking $kernel(s_1) \subset H_2$ as a type of interaction.

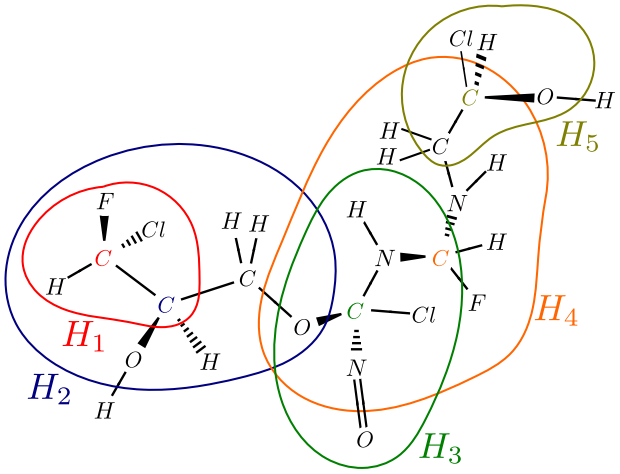
We also may suppose that an intersection between two minimal stereo subgraphs may not be a sufficiently relevant information. Thus the graph G_3 is constructed with 3 different type of interaction, by considering that two stereo vertex are related if we have at least $kernel(s_1) \subset H_2$ or $kernel(s_2) \subset H_1$.

Finally G_4 is constructed by taking the two previous assumptions together.

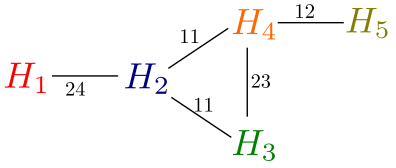
As graphs of interaction are graphs without order, we may apply any graph kernel (for the experiment in Section 5 we apply the treelet kernel [7]) to measure their similarity.

5 Experiments

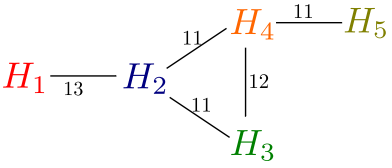
We have tested our method on two datasets. For both of them we use the same protocol, a nested cross-validation, to choose the parameter and estimate the performance. The outer cross-validation is a leave-one-out procedure, used to compute an error for each molecule of the dataset. For each fold, we use another leave-one-out procedure on the remaining molecules, to compute a validation error. We use standard SVM method [5, 6] for classification and regression of molecules.



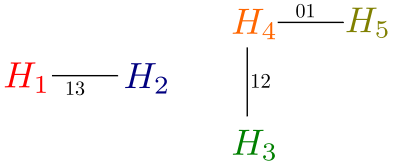
(a) An ordered graph and its minimal stereo subgraphs



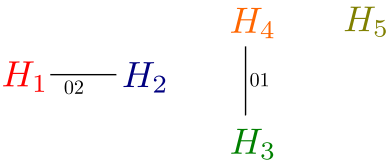
(b) G_1



(c) G_2



(d) G_3



(e) G_4

Fig. 4: One ordered graph and its different graph of interactions, obtained by taking different function of interactions.

Table 1: Averages values of numbers of vertices ($\overline{|V|}$), edges ($\overline{|E|}$), different labels($\overline{|\mathcal{L}_V|}$,$\overline{|\mathcal{L}_E|}$), and mean degree ($\overline{d}$) of graph of interactions.

	$\overline{\lvert V\rvert}$	$\overline{\lvert E\rvert}$	$\overline{\lvert\mathcal{L}_V\rvert}$	$\overline{\lvert\mathcal{L}_E\rvert}$	$\overline{d}$
Graph 1	5	7	4.5	3	2.8
Graph 2	5	7	4.5	3	2.8
Graph 3	5	2	4.5	2	0.8
Graph 4	5	1	4.5	1	0.4

(a) ACE dataset

	$\overline{\lvert V\rvert}$	$\overline{\lvert E\rvert}$	$\overline{\lvert\mathcal{L}_V\rvert}$	$\overline{\lvert\mathcal{L}_E\rvert}$	$\overline{d}$
Graph 1	8.55	17.4	8.38	5.71	4.07
Graph 2	8.55	17.4	8.38	3.71	4.07
Graph 3	8.55	11.3	8.38	4.71	2.62
Graph 4	8.55	6.14	8.38	2.71	1.43

(b) Vitamin dataset

Our first experiment is based on a dataset composed of all the stereoisomers of the perindoprilate [3]. As this molecule has 5 stereocenters, the dataset is composed of $2^5 = 32$ molecules. In this dataset, we try to predict if a molecule inhibit the angiotensin-converting enzyme (ACE). Basic statisistics about the graphs of interactions $G_i = (V_i, E_i, \mu_i, \nu_i)$ deduced from this dataset is displayed in Table 2a.

Table 2: Classification of the ACE inhibitory activity of perindopirilates stereoisomers

Method	Accuracy
Brown [1]	*96.875*
Stereo Kernel [11]	87.5
Stereo + Extended subgraphs [10]	*96.875*
Graph of interaction 1	93.75
Graph of interaction 2	93.75
Graph of interaction 3	93.75
Graph of interaction 4	84.375
Graph of interaction 1 with MKL	**100**
Graph of interaction 2 with MKL	**100**
Graph of interaction 3 with MKL	87.5
Graph of interaction 4 with MKL	90.625

For this first experiment we have not included results of method which do not include stereoisomerism information [14, 7]. Indeed all molecules of the dataset are stereoisomers of each other, so those methods cannot differentiate any molecule of this dataset and are consequently unable to predict the considered property. Moreover, information not related to stereoisomerism included in kernel [1] consists of the same patterns for all molecules. This leads to add a constant shift to all values of the kernel and hence does not deteriorate the prediction for this dataset. In this dataset two stereocenters have a same minimal stereo subgraph, but different surrounding. The stereo kernel [11] and one of the graph of interactions (G_4), can not differentiate those two stereocenters, which

have different influence on the property, this explains why other method ([1, 10] and the three other graphs of interactions) obtain a better accuracy. However, for our graphs of interactions, treelet of size one have a negative effect on the classification, this explains why we do not obtain better results than [1, 10]. By using a multiple kernel learning algorithm [16], we can learn a weight for each treelet, that allow us to discard treelet of size 1 and to obtain the best results with the first and second graph of interactions. The third graph of interactions have very few edges and a low degree (Table 2a) which explains why the treelet kernel with multiple kernel learning obtains poor results with this graph.

The second dataset is a dataset of synthetic vitamin D derivatives, used in [1]. This dataset is composed of 69 molecules, with an average of 9 stereocenters per molecule. This dataset is associated to a regression problem, which consists in predicting the biological activities of each molecules. As for the previous dataset, statistics about the graphs of interactions deduced from this dataset can be found in Table 2b.

Table 3: Prediction of the biological activity of synthetic vitamin D derivatives.

Method	RMSE
1 - Tree patterns Kernel [14]	0.251
2 - Treelet Kernel [7]	0.271
3 - Brown [1]	0.184
4 - Stereo Kernel [11]	0.194
5 - Stereo + Extended subgraphs [10]	0.180
6 - Graph of interaction 1	0.177
7 - Graph of interaction 2	0.177
8 - Graph of interaction 3	**0.169**
9 - Graph of interaction 4	0.172

Methods which do not encode stereoisomerism information [14, 7] obtain poor results as we can see in Table 3 (lines 1-2). The adaptation of the tree pattern kernel to stereoisomerism [1] and our previous kernels [11, 10] (lines 3-5) improves the results over the two previous methods hence showing the insight of adding stereoisomerism information. Taking into account relationships between minimal stereo subgraphs (lines 6-9) allows us to obtain better results than our previous method [10].

6 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 presented previously 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. In this report we have proposed an extension of this previous methods which consists in construct a new graph, where each nodes represent a stereo subgraph and each edge encode the interaction between stereo subgraphs. This graph allows us to take into account relationships between stereo subgraphs. The relevance of this approach is demonstrated by our experiments on two datasets.

Acknowledgements

This work was in part made using the computing resources funded under the CPER.

References

1. 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*, 8(1):63–81, 2010.
2. L. Brun, D. Conte, P. Foggia, M. Vento, and D. Villemin. Symbolic learning vs. graph kernels: An experimental comparison in a chemical application. In *ADBIS (Local Proceedings)*, pages 31–40, 2010.
3. J. A. Castillo-Garit, Y. Marrero-Ponce, F. Torrens, and R. Rotondo. Atom-based stochastic and non-stochastic 3d-chiral bilinear indices and their applications to central chirality codification. *Journal of Molecular Graphics and Modelling*, 26(1):32–47, 2007.
4. D. Cherqaoui and D. Villemin. Use of a neural network to determine the boiling point of alkanes. *J. Chem. Soc., Faraday Trans.*, 90(1):97–102, 1994.
5. C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
6. H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *NIPS*, pages 155–161, 1996.
7. B. Gaüzère, L. Brun, and D. Villemin. Two New Graphs Kernels in Chemoinformatics. *Pattern Recognition Letters*, 33(15):2038–2047, 2012.
8. A. Golbraikh and A. Tropsha. Qsar modeling using chirality descriptors derived from molecular topology. *Journal of chemical information and computer sciences*, 43(1):144–154, 2003.
9. P.-A. Grenier, L. Brun, and D. Villemin. Incorporating stereo information within the graph kernel framework. Technical report, CNRS UMR 6072 GREYC, 2013. http://hal.archives-ouvertes.fr/hal-00809066/.
10. P.-A. Grenier, L. Brun, and D. Villemin. Incorporating molecules stereisomerism within the machine learning framework. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 12–21. Springer, 2014.

11. P.-A. Grenier, L. Brun, D. Villemin, et al. A graph kernel incorporating molecule's stereisomerism information. *Proceedings of ICPR 2014*, 2014.
12. J. Jacques, A. Collet, and S. Wilen. *Enantiomers, racemates, and resolutions*. Krieger Pub. Co., 1991.
13. H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *ICML*, volume 3, pages 321–328, 2003.
14. P. Mahé and J.-P. Vert. Graph kernels based on tree patterns for molecules. *Machine Learning*, 75(1):3–35, Oct. 2008.
15. G. Poezevara, B. Cuissart, and B. Crémilleux. Discovering emerging graph patterns from chemicals. In *Foundations of Intelligent Systems*, pages 45–55. Springer, 2009.
16. M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1065–1072. ACM, 2009.
17. W. T. Wipke and T. M. Dyott. Stereochemically unique naming algorithm. *Journal of the American Chemical Society*, 96(15):4834–4842, 1974.