# Decomposing Smiles into words : a molecular embedding

**Hamza Tajmouati[1]; Philippe Gendreau[1,2]; Quentin Perron[1]; Nicolas Do Huu[1]**
[1]Iktos, 61 rue Blomet 75015 Paris; [2]Mines ParisTech, 60 Boulevard Saint-Michel 75006 Paris

## Introduction

Natural language and Smiles are both sequential information that represent an upper level of understanding. Words in natural language transcribe some meaning that has been leveraged in continuous embedding representations (Word2Vec)[1].

| | Language | Chemistry |
|---|---|---|
| **Object** | Sentence | Smiles |
| **Building blocks** | letters: a, b, c … | symbols: C, (, @, 1 … |
| **Information level** | Words | C=O? ? c1ccccc1? |

- ✔ Meaningful
- ✔ Vector representation
- ✔ Sentiment analysis
- ✔ Document classification

…

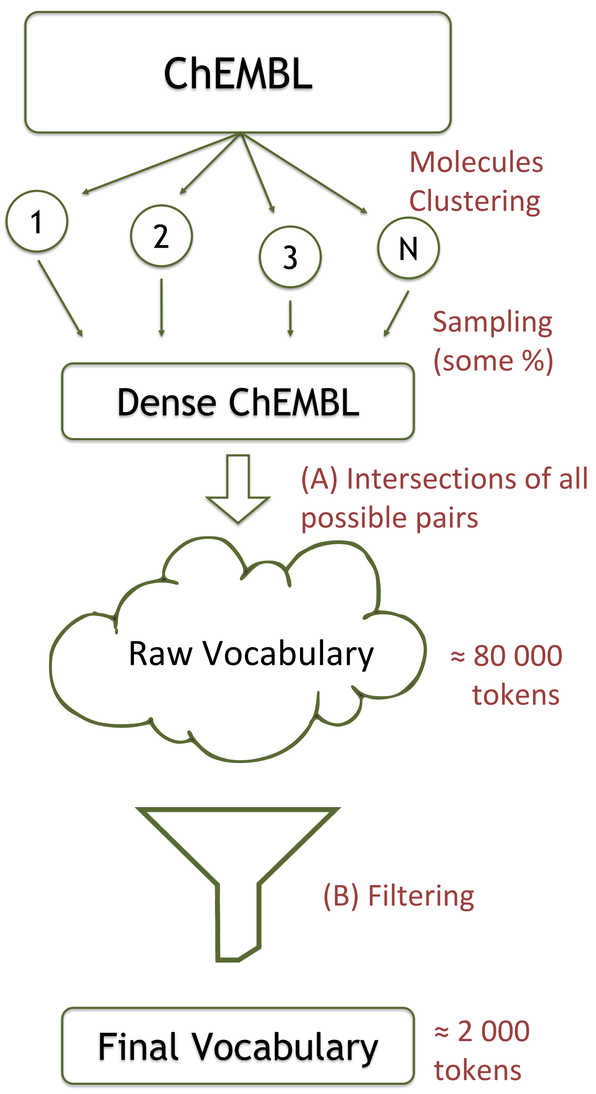
- ? Meaningful
- ? Vector representation
- ? Chemical properties prediction
- ? Molecules generation

**Question**:
Similarly to natural language, can we create a vocabulary of sub-smiles that contain some semantics and leverage it?

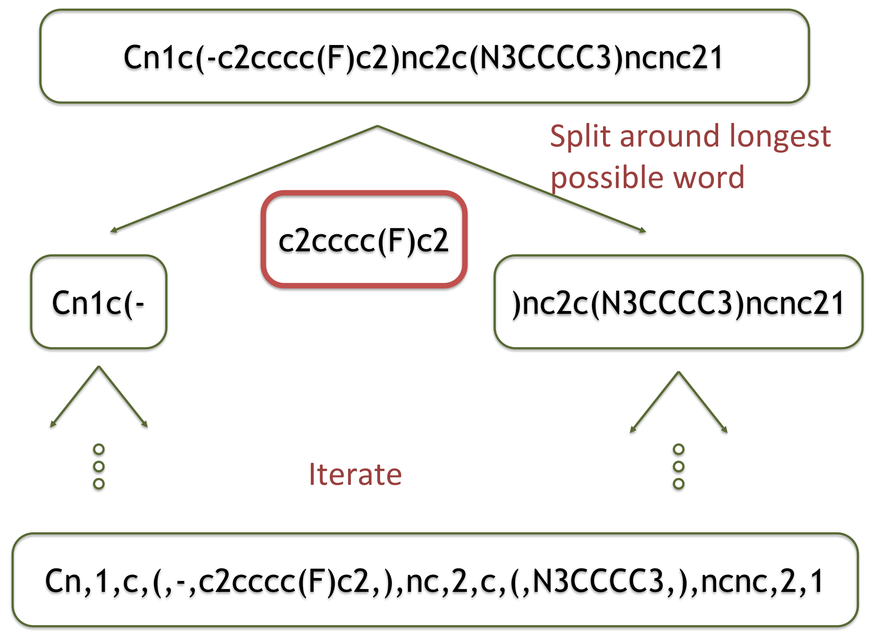
## Build the vocabulary

Our vocabulary must meet **requirements**:

- (A) Construction purely **data-driven**
  —> intersection of Smiles strings
  e.g : *S(=O)(=O)* ∩ *C(=O)* yields *(=O)*
- (B) As **generic** as possible with **valid Smiles** only




## Decompose a Smiles




## Vocabulary topology

We decomposed ChEMBL with the vocabulary and studied how the words were used.

Figure 1: Vocabulary words length distribution

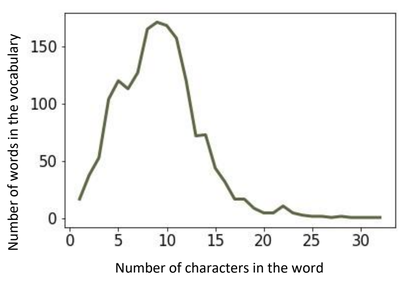



Figure 2: Tokens number distribution of ChEMBL Smiles

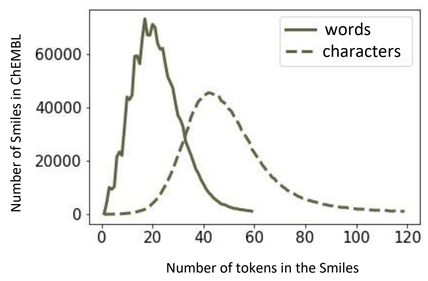
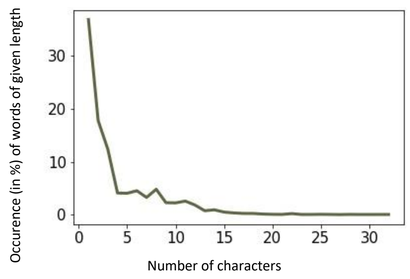
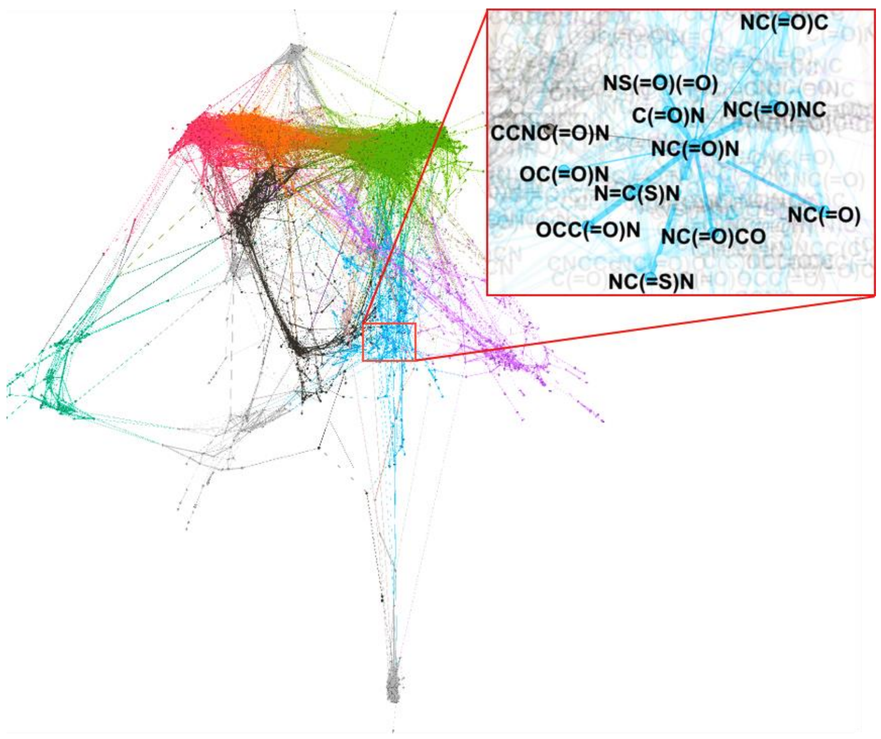



Figure 3: Words occurence distribution by length




## Word embedding

Words In ChEMBL were given a vector representation thanks to the Word2Vec algorithm. The following graph links vectors that have a distance below a given threshold. It highlights groups of similar words and shows that their vector representation caught a form of meaning.




## Generate a dataset

This first task consists in training a generative model. We used recurrent neural networks[2] and compared our approach to the character by character state of the art.

| | Words | Characters (state of the art) |
|---|---|---|
| **Valid Smiles** | 93.0 % | 90.8 % |
| **Redundance** | 4.3 % | **4.1 %** |
| **Internal diversity** | 0.82 | 0.82 |
| **Carbon skeleton** | 42.0 % | 39.7 % |
| **Generation rate** | **620 mol/s** | 300 mol/s |

## Prediction

We used the embedding of our words to form a new molecular fingerprint used in prediction tasks - regression and classification.

| | Words | Morgan |
|---|---|---|
| **Aid (AUC)** | 0.65 | **0.71** |
| **BBBP (AUC)** | 0.88 | **0.92** |
| **Clintox (AUC)** | 0.80 | 0.77 |
| **Ames (AUC)** | 0.81 | **0.88** |
| **Tox21 (AUC)** | 0.78 | **0.82** |
| **ESOL ($r^2$)** | 0.70 | **0.71** |

## Conclusions

- First successful attempt to create a **meaningful** sub-smiles vocabulary and represent it with Word2Vec
- Our approach **outperformed** the state of the art in the unconstrained generation task
- In prediction tasks, our home-made embedding shows promising performances with specific datasets

## Perspectives

- Adapt the vocabulary to problems with particular structures
- Add chemical information to the words, for example pharmacophoric fingerprint to enhance the embedding
- Leverage the similarity between words to extend the applicability domain of predictors

References :
[1]: Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 746-751).
[2]: Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, *9*(1), 48.

**Contact:**
**61 rue Blomet 75015 Paris**
**contact@iktos.com**
**+33 7 68 80 50 76**

IKTOS