# Prédiction et Réactivité avec Exploration Dynamique et Innovation par la Chimie Théorique (PREDICT)

Luc Brun, Laurent Joubert, Vincent Tognetti

Mars 2025

## 1 Conditions de la thèse

**Bourse CNRS environnementée** finançant le salaire (2200€ brut), le matériel et des frais de déplacements en conférences

**Lieu de travail :** Rouen (Institut CARMeN - ex laboratoire COBRA, UMR 6064) et Caen (GREYC, UMR 6072), selon des modalités à discuter avec le doctorant

**Contacts (candidatures et informations) :**

— luc.brun@ensicaen.fr,
— laurent.joubert@univ-rouen.fr,
— vincent.tognetti@univ-rouen.fr

La thèse se déroulera dans le cadre d'une collaboration étroite entre l'Institut CARMeN (ex laboratoire COBRA) [1] (Chimie Théorique) et GREYC [2] (Informatique).

## 2 Description de la problématique

Ce projet multidisciplinaire (chimie théorique et informatique) se concentre sur les interactions non-covalentes appelées "à trous sigma". Elles se produisent lorsqu'un atome, comme un halogène ou un chalcogène, présente une région appauvrie en densité électronique, appelée "trou sigma". Cette région peut interagir avec des espèces riches en électrons, influençant ainsi la stabilisation des structures moléculaires et cristallines. Ces interactions sont cruciales en biochimie, en ingénierie des matériaux et en catalyse.

Prédire la stabilité du complexe ainsi formé entre une espèce déficiente en électrons (dite "électrophile") et une autre riche en électrons (le "nucléophile")

1. https ://https ://www.labex-synorg.fr/laboratoires/cobra-rouen/
2. https ://www.greyc.fr

est donc un enjeu majeur dans l'optimisation de structures supramoléculaires. L'approche généralement employée repose sur l'étude des propriétés intrinsèques de chacun des réactifs. En effet, il a été montré expérimentalement (échelle de Mayr-Patz) que les vitesses de réaction peuvent s'exprimer en fonction de paramètres caractéristiques de chaque entité indépendante.

En général, ceux-ci sont calculés sur les structures de plus basse énergie. Cependant, cela ne prend pas en compte la dynamique de chacun des partenaires, induite par la température, et qui ne peut nullement être négligée. Celle-ci peut être simulée à partir de calculs de chimie quantique, malheureusement très coûteux en termes de ressources computationnelles, limitant son application à des jeux réduits de petites molécules. Un mouvement relativement récent (2017) consiste à utiliser des méthodes de machine learning et plus précisément de Deep learning pour approximer efficacement le calcul de l'énergie afin d'en déduire (rapidement) la dynamique de la molécule.

Cette thèse se propose d'adresser cette problématique et se situe donc à l'intersection entre la chimie théorique et le machine learning. Une approche classique de l'apprentissage profond à la chimie consiste à utiliser un graphe moléculaire où les sommets correspondent aux atomes et les arêtes aux liaisons (simple, double, triple, aromatiques) entre atomes. Les atomes sont généralement décrits par leur type sans utiliser de coordonnées. Ce type d'utilisation de l'apprentissage ne peut intrinsèquement pas prendre en compte l'évolution de la position des atomes et donc la dynamique de la molécule.

L'utilisation de l'apprentissage profond pour le calcul dynamique de propriétés moléculaires doit également adresser un certain nombre de contraintes : Il se doit d'être équivariant [3] aux rotations et invariant [4] aux translations qui définissent les coordonnées des molécules. L'application d'une symétrie sur la molécule doit également produire un résultat équivariant. Enfin, l'ordre des sommets étant arbitraire, le réseau doit être équivariant à ce dernier. Notons également qu'un calcul de dynamique moléculaire s'effectue à l'aide des dérivées de l'énergie moléculaire. Un réseau calculant cette énergie doit donc fournir un résultat deux fois différentiable ce qui interdit l'utilisation de fonctions non $C^1$ comme la fonction RELU [SKS+17].

Cet ensemble de contraintes a induit des réseaux convolutifs très différents de ceux habituellement utilisés en chémoinformatique. L'opération de convolution consiste à considérer à chaque itération un graphe dynamique reliant chaque atome à tous ces voisins situés à une distance inférieure à un seuil (entre 10 et 30 Å). Ces interactions entre atomes devant être invariantes aux rotations, de nombreux auteurs ont dans un premier temps utilisé uniquement la distance (réelle) entre atomes. On obtient ainsi un schéma général (voir par exemple [SKS+17] ou [UM19]) où après un plongement des atomes (codant le type et les coordonnées de chaque atome) un ensemble de modules d'interactions est appliqué avec des connections résiduelles. Les matrices de poids permettent de pondérer les mises à jour en fonction de ces distances. Notons que les modules d'interac-

3. Une fonction $f$ est équivariante à une transformation $P$ si $f(P(x)) = P(f(x))$
4. Une fonction $f$ est invariante à une transformation $P$ si $f(P(x)) = f(x)$

tion sont généralement basés sur des convolutions mais des architectures à base de transformeurs ont également été proposées [PSG+24]. L'énergie de chaque atome est ensuite estimée avant de sommer les contributions de tous les atomes pour avoir une estimation de l'énergie globale de la molécule.

Ce schéma est réducteur pour plusieurs raisons :

1. La simple utilisation d'une distance (donc d'une valeur réelle) ne permet pas de prendre en compte la position relative des atomes. Ce point à été partiellement corrigé par DimNet [GGG22, KGMG20] puis par GemNet [GBG24].
2. La restriction des interactions à un certain seuil est une simplification. Ce point a été noté assez tôt, puisque par exemple, PhysNET [UM19] propose une correction de l'estimation de l'énergie à partir d'interactions non locales entre les atomes. Cette correction est toutefois appliquée à posteriori. Une approche plus intéressante est proposée par SpookyNet [UCG+21] qui introduit des interactions non locales dans le module d'interaction.
3. L'initialisation du plongement des atomes avec leurs types et leur positions, ne prend pas en compte les propriétés électroniques des atomes (les électrons). Cette limitation est importante dans notre cas pour la prédiction des propriétés électroniques de la molécule telles que les trous sigma.

# 3 Plan de travail

Le travail que nous allons entreprendre débutera par un benchmarking des modèles existants en privilégiant les approches utilisant les fonctions angulaires au moins au premier ordre (voir DimNet [GGG22, KGMG20]) car elles paraissent fondamentales pour une prédiction efficace des descripteurs de réactivité. En outre, les modèles incorporeront nécessairement, dans les données d'entrée, les propriétés électroniques des atomes, à la suite des travaux de l'équipe de chimie théorique qui ont identifié de manière précise les descripteurs quantiques pertinents pour l'étude de ce type d'interactions et qui ont été implémentés dans le logiciel de référence ADF par cette même équipe. Ces modifications s'accompagneront de l'évaluation de la complexité du modèle obtenu afin de garantir une des principales qualités de l'approche par apprentissage machine, à savoir un temps de calcul compatible avec le traitement de grandes bases.

Enfin, afin d'atteindre les applications pratiques, une collaboration avec l'équipe du Dr. Robin Weiss (Chaire Professeur Junior, intégrant la nouvelle UMR CARMeN en janvier 2025 à Caen), spécialiste de la synthèse de la caractérisation physicochimique des trous sigma[6], est prévue et devrait permettre la mise au point de nouveaux systèmes d'intérêt sociétal. Cette synergie s'appuiera également sur le fait que les deux équipes de Rouen et de Caen sont toutes deux membres du GDR sigma-hole.

# Références


[GBG24] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet : Universal directional graph neural networks for molecules, 2024.

[GGG22] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs, 2022.

[KGMG20] Johannes Klicpera, Shankari Giri, Johannes T. Margraf, and Stephan Gunnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *ArXiv*, abs/2011.14115, 2020.

[PSG+24] Raul P. Pelaez, Guillem Simeon, Raimondas Galvelis, Antonio Mirarchi, Peter Eastman, Stefan Doerr, Philipp Thölke, Thomas E. Markland, and Gianni De Fabritiis. Torchmd-net 2.0 : Fast neural network potentials for molecular simulations. *Journal of Chemical Theory and Computation*, 20(10) :4076–4087, 2024. PMID : 38743033.

[SKS+17] Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet : A continuous-filter convolutional neural network for modeling quantum interactions, 2017.

[UCG+21] Oliver T. Unke, Stefan Chmiela, Michael Gastegger, Kristof T. Schütt, Huziel E. Sauceda, and Klaus-Robert Müller. Spookynet : Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature Communications*, 12(1) :7273, Dec 2021.

[UM19] Oliver T. Unke and Markus Meuwly. Physnet : A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of Chemical Theory and Computation*, 15(6) :3678–3693, 2019. PMID : 31042390.