

Développement de méthodes de sélection et d'optimisation de conditions réactionnelles pour des transformations chimiques.

Le projet de thèse s'inscrit dans le cadre d'une collaboration entre le Laboratoire d'Infochimie (UMR 7551), l'équipe Apprentissage et Fouille de Données du Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection (LSIIT, UMR 7005) ainsi que la société *Novalyst Discovery*. L'objectif de ce travail est de développer dans le cadre des réactions d'hydrogénation, des méthodes informatiques permettant d'établir des liens entre la structure chimique des réactifs, les conditions expérimentales (température, pression, et solvant), le catalyseur, la structure des produits et le rendement réactionnel.

Ce projet de thèse s'inscrit dans la stratégie générale de l'entreprise, qui consiste à lier la chémo-informatique, la synthèse parallèle et la catalyse dans le but de construire une "encyclopédie" de la connaissance chimique. Par une approche systématique s'appuyant sur des technologies propriétaires, *Novalyst Discovery* explore de nouveaux systèmes catalytiques pour découvrir, caractériser et répertorier les transformations chimiques. Actuellement, la société dispose d'une grande quantité de données expérimentales sur la réaction d'hydrogénation (composition des bruts réactionnels et rendements des différent produits) obtenus dans différentes conditions expérimentales.

Le travail impliquera l'organisation de données la récolte sur les réactions d'hydrogénation ainsi que le développement d'un système expert permettant d'anticiper les conditions expérimentales optimales pour une réaction donnée.

Sur la **figure 1**, nous avons représenté la base de donnée qui couvre l'espace des trois variables, les catalyseurs, les conditions réactionnelles et les molécules sondes

Les molécules sondes permettent d'étalonner la réactivité des différents groupes fonctionnels qui entrent en jeu dans la réaction. Les catalyseurs sont représentatifs de la diversité réactionnelle possible. Enfin les conditions réactionnelles sont représentatives des conditions les plus communément décrites dans la littérature. Par extrapolation, il est alors possible de modéliser par un point une réaction de catalyse.

L'objectif de ce travail de thèse est de mettre au point un système d'analyse chimie informatique permettant de comparer une transformation requête avec celles présentes dans la base pour en faire ressortir les conditions et le catalyseur à utiliser pour effectuer cette réaction. Cette base de donnée permettra également de faire ressortir des propriétés de sélectivité qui ne sont pas accessible à l'heure actuelle.

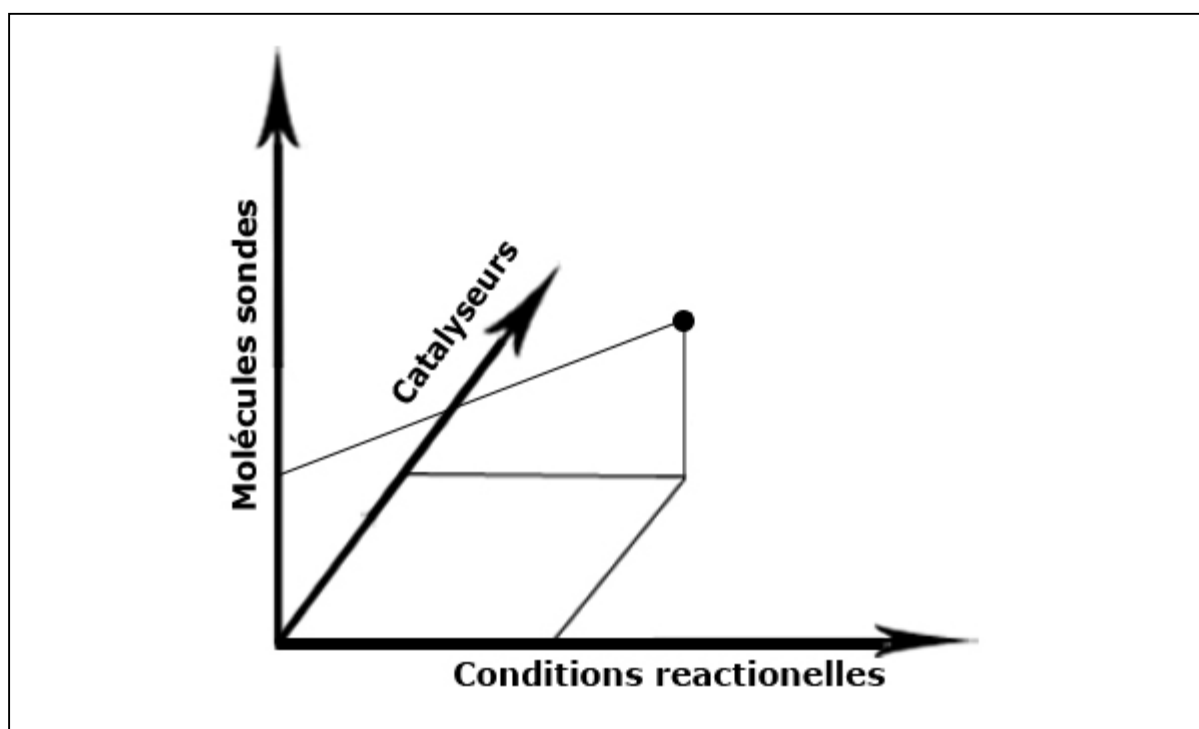


Figure 1

Les données expérimentales sont stockées dans une base de données permettant de les organiser, d'effectuer des recherches et d'exporter au format RDF. A partir de cette base, les jeux de données correspondant à des conditions spécifiques seront sélectionnés puis traités par les méthodes de chimie informatique développées dans le cadre de cette thèse.

Le système expert intégrera deux approches complémentaires : (1) la recherche par similarité et, (2) les simulations par réseaux neuronaux. L'approche 1 impliquera les algorithmes permettant de sélectionner à partir d'une requête, les réactions les plus similaires stockées dans la base de données et donc d'extraire les conditions et le catalyseur les plus efficaces pour la réaction donnée. Les algorithmes de recherche de similarité de deux réactions seront développés et adaptés aux systèmes expérimentalement étudiés. Ils devront permettre de faire des requêtes intelligentes sur la base de données de façon la plus automatique possible. L'approche 2 concernera la modélisation « structure – activité » en utilisant la méthode de réseaux de neurones. L'objectif sera d'établir les relations quantitatives entre la structure des réactifs, les conditions réactionnelles, le type du catalyseur, d'une part, et la composition de produits et le rendement réactionnel d'autre part. Ainsi, ces deux approches complémentaires permettront de faire de meilleurs choix de conditions réactionnelles et de catalyseur pour la réaction étudiée.

Dans toutes les étapes de ce travail, le concept des descripteurs « fragmentaux » ainsi que les algorithmes développés dans le laboratoire d'Infochimie du Professeur Alexandre VARNEK (UMR 7551) seront utilisés. Le Docteur Nicolas LACHICHE, Maître de Conférences au Département d'Informatique de l'IUT Robert Schuman, membre de l'équipe Apprentissage et Fouille de Données du LSIIT (UMR 7005), veillera en parallèle aux aspects informatique du travail. Les compétences de cette équipe dans la mise en œuvre de diverses techniques de fouille de données, en particulier celles capables de prédire les rendements à partir des données telles que les approches neuronales, génétiques ou probabilistes, d'une part, et son expérience dans l'utilisation de données complexes, en particulier les données relationnelles, d'autre part, seront mises à profit.