DOCTORAT BRETAGNE LOIRE / MATHSTIC

Soutenance de thèse de doctorat

université angers

Par **Jules Leguy**

le 9 décembre 2022

Recherche combinatoire guidée par apprentissage artificiel en chimie moléculaire

---

Membres du jury

Rapporteurs :	Laetitia Jourdan	Professeure des universités, CRIStAL, Université de Lille
	Jean-Claude Crivello	Chargé de recherche (HDR), ICMPE, CNRS
Examinateur :	Gaël Varoquaux	Directeur de recherche, INRIA, INRIA Saclay
Dir. de thèse :	Béatrice Duval	Professeure des universités, LERIA, Université d'Angers
Co-enc. de thèse :	Benoit Da Mota	Maître de conférences, LERIA, Université d'Angers
Co-enc. de thèse :	Thomas Cauchy	Maître de conférences, MOLTECH-Anjou, Université d'Angers

Introduction

Recherche de molécules satisfaisant des propriétés moléculaires cibles

- Historiquement, recherche par amélioration itérative des molécules connues (exemple de l'aspirine).
- Actuellement, recherche de méthodes pour générer automatiquement des molécules [a, b].

Exemple : histoire de l'aspirine

- Écorce de saule déjà consommée par l'Homme de Néandertal pour ses propriétés pharmaceutiques [c].
- Début du 19ÈME siècle : la molécule d'acide salicylique est isolée.
- 1852 : découverte de la molécule d'acide acétylsalicylique par modification de l'acide salicylique. Commercialisation sous le nom d'aspirine.

a. David H. Freedman. "Hunting for New Drugs with AI". In : *Nature* 576 (déc. 2019)
b. Debleena Paul et al. "Artificial intelligence in drug discovery and development". In : *Drug Discovery Today* 26 (jan. 2021)
c. Laura S. Weyrich et al. "Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus". In : *Nature* (avr. 2017)

Chimie et représentations moléculaires

Chimie moléculaire

- Molécules.
- Représentations moléculaires.
- Règles de valence (validité).

Chimie organique

- Sous-ensemble de la chimie moléculaire.
- Atomes de carbone, hydrogène, oxygène, azote, etc.

Figure 1 : Nuage d'atomes.

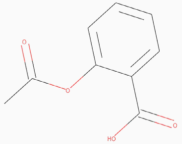


Figure 2 : Graphe moléculaire.

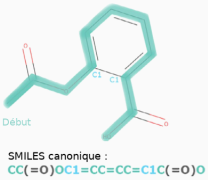
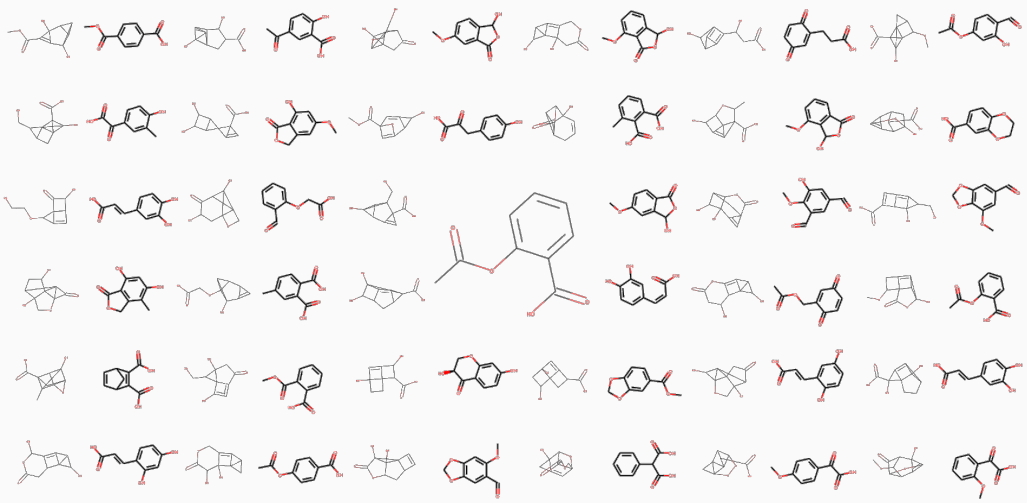


Figure 3 : SMILES (parcours de graphe)

Espace de recherche combinatoire



C 9
H 8
O 4

Figure 4 : Molécule d'acide acétylsalicylique et sélection d'une soixantaine de molécules parmi des dizaines de milliers possédant la même composition atomique (isomères).

Propriétés moléculaires d'intérêt

QED : *Quantitative Estimate of Drug-likeness* [a]

- Propriété « jouet » de la chimie pharmaceutique : ressemblance à des médicaments.
- Valeurs $\in [0, 1]$.
- Coût très faible.

Énergie HOMO : *Highest Occupied Molecular Orbital*

- Propriété électronique : énergie de l'électron de plus haute énergie.
- Valeurs environ $\in$ [-10, -1] eV.
- Dépend de calculs en chimie quantique : coût très important (10^3 s).
- Calculs liés à la géométrie moléculaire.

a. G. Richard Bickerton et al. "Quantifying the chemical beauty of drugs". In : *Nature Chemistry* 4 (fév. 2012)

État de l'art : génération moléculaire

Approches évolutionnaires

- Méthodes récentes : ChemGE [a], GB-GA [b], CReM [c], MolFinder [d].
- Représentation moléculaire :
  - SMILES (ChemGE, MolFinder).
  - Graphe moléculaire (GB-GA, CReM).
- Limites potentielles :
  - Réalisme des solutions.
  - Représentation SMILES : génération de molécules invalides.

---

a. Naruki Yoshikawa et al. "Population-based De Novo Molecule Generation, Using Grammatical Evolution". In : *Chemistry Letters* (2018)

b. Jan H. Jensen. "A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space". In : *Chemical Science* 10.12 (2019)

c. Pavel Polishchuk. "CReM : chemically reasonable mutations framework for structure generation". In : *Journal of Cheminformatics* 12.1 (2020), p. 28

d. Yongbeom Kwon et Juyong Lee. "MolFinder : an evolutionary algorithm for the global optimization of molecular properties and the extensive exploration of chemical space using SMILES". In : *Journal of Cheminformatics* 13 (mars 2021), p. 24

Approches basées sur des modèles d'apprentissage profond

- Deux paradigmes :
  - Apprentissage d'un générateur moléculaire [a, b].
  - Apprentissage d'un espace moléculaire latent continu et optimisation continue [c, d].
- Limites potientielles :
  - Disponibilité des données ?
  - Données d'entraînement : restriction de l'espace de recherche accessible ?

a. Nicola De Cao et Thomas Kipf. "MolGAN : An implicit generative model for small molecular graphs". In : *arXiv :1805.11973 [cs, stat]* (mai 2018)

b. Niklas W. A. Gebauer, M. Gastegger et Kristof T. Schütt. "Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules". In : *NeurIPS*. 2019

c. Rafael Gómez-Bombarelli et al. "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules". In : *ACS Central Science* 4 (fév. 2018), p. 268-276. ISSN : 2374-7943

d. Robin Winter et al. "Efficient multi-objective molecular optimization in a continuous latent space". In : *Chemical Science* 10.34 (2019)

Génération moléculaire : contraintes et enjeux

- Taille de l'espace moléculaire : jusqu'à 10^{60} molécules [a].
- Applications principales :
  - Chimie pharmaceutique.
  - Chimie des matériaux moléculaires organiques (propriétés électroniques coûteuses).
- Dépendance à des jeux de données : introduction d'un biais peu contrôlable ?
- Enjeu d'interprétabilité du processus de génération : étude structure-propriété.

a. Regine S. Bohacek, Colin McMartin et Wayne C. Guida. "The art and practice of structure-based drug design : A molecular modeling perspective". In : *Medicinal Research Reviews* 16 (1996)

Plan des contributions

1. EvoMol : un algorithme évolutionnaire pour l'optimisation de propriétés moléculaires.
   - Générique et interprétable.
   - Prédominance de intensification de l'espace de recherche.

2. Optimisation de la diversité moléculaire.
   - Maximisation de la diversité au sein de la population d'EvoMol.
   - Compromis intensification/exploration.

3. Optimisation basée sur un modèle de substitution.
   - Optimisation efficace de propriétés coûteuses.
   - Modèle de substitution : connaissance issue des appels précédents à la fonction objectif.

EvoMol : un algorithme évolutionnaire pour l'optimisation de propriétés moléculaires

Motivation pour la définition d'un algorithme évolutionnaire

Objectifs et caractéristiques des algorithmes évolutionnaires

1. Définition d'une méthode généraliste : cadre générique (espace de recherche, fonction objectif).
2. Contrôle des biais : pas de dépendance à un jeu de données spécifique.
3. Interprétabilité de la recherche : possibilité d'enregistrer et de représenter graphiquement l'historique des mutations.

Méthode : algorithme EvoMol

Algorithme évolutionnaire pour l'optimisation de propriétés moléculaires (EvoMol)

Soit une population pop de taille constante et f la fonction objectif boîte-noire
tant que le critère d'arrêt n'est pas atteint **faire**
  calcul de P une pile des individus de pop triée selon les valeurs de f
  sélection de $L \subseteq \mathrm{pop}$ la liste d'individus à remplacer à l'étape courante
  pour chaque individu ind de la liste L **faire**
    recherche améliorant mut de ind par mutation des solutions de P par ordre de priorité
    remplacement de ind par mut dans pop
  fin pour
fin tant que

- Intuition : les solutions les moins bonnes sont remplacées par la mutation des meilleures.
- Prédominance intensification.
- Pas d'opérateur de recombinaison : arbre d'exploration.

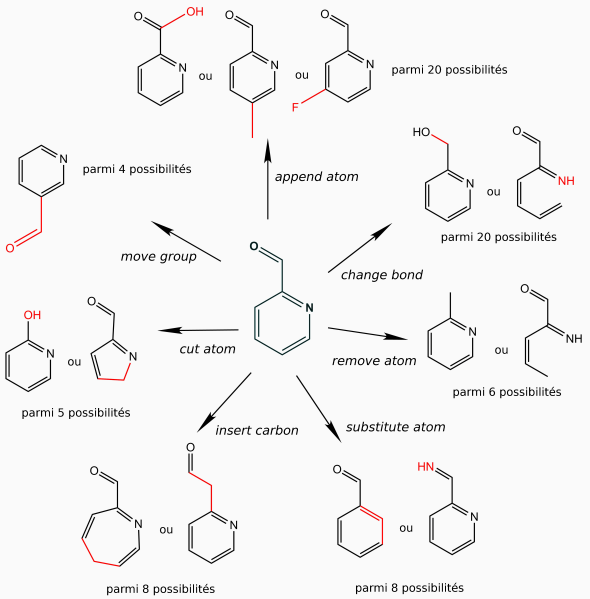
Méthode : opérateur de mutation

Actions sur le graphe moléculaire

- 7 actions.
- Modifications locales intuitives.

Opérateur de mutation

- Application aléatoire des actions sur le graphe moléculaire.
- Filtrage des actions basé sur les règles de valence : garantie de validité des solutions résultantes.



Expériences menées (extrait des résultats)

- Optimisation de propriétés moléculaires peu coûteuses issues de la chimie du médicament : QED, plogP.

  Résultats compétitifs avec l'état de l'art.

- Passage du *benchmark* GuacaMol [a].

  Meilleures performances de l'état de l'art au moment de la publication des résultats [b].

- Optimisation de propriétés électroniques coûteuses (énergies HOMO et LUMO).

  Succès de l'optimisation : redécouverte d'une molécule connue.

a. Nathan Brown et al. "GuacaMol : Benchmarking Models for de Novo Molecular Design". In : *Journal of Chemical Information and Modeling* 59.3 (mars 2019)

b. Jules Leguy et al. "EvoMol : a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation". In : *Journal of Cheminformatics* 12.1 (sept. 2020)

Arbre d'exploration

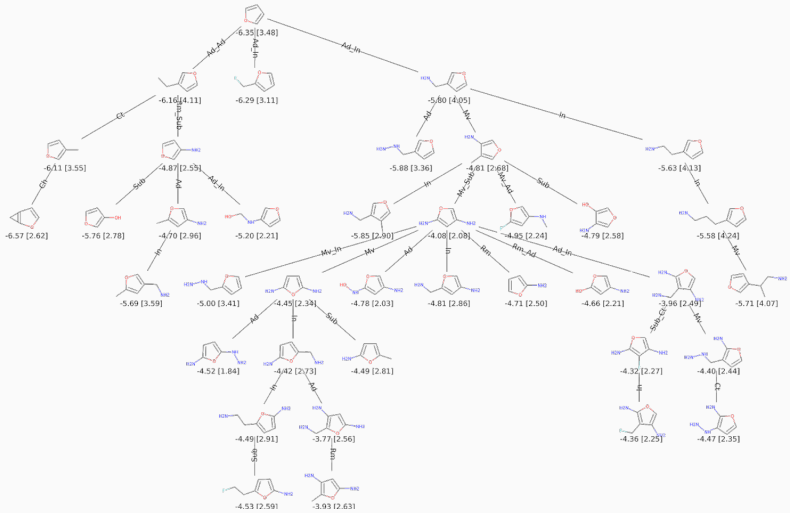
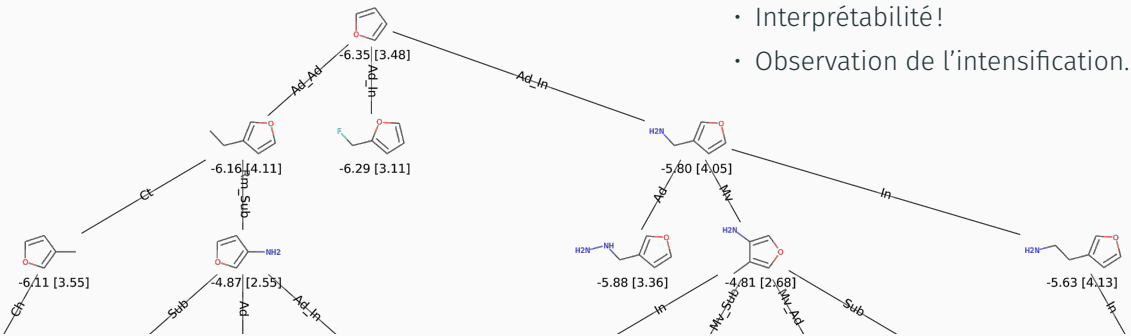


Figure 5 : Arbre d'exploration pour la maximisation de l'énergie HOMO.

Arbre d'exploration (extrait)



- Interprétabilité !
- Observation de l'intensification.

Figure 6 : Arbre d'exploration pour la maximisation de l'énergie HOMO (extrait).

Optimisation de la diversité moléculaire

Optimisation de la diversité moléculaire

Motivation

- Un manque de diversité dans les jeux de données moléculaires peut affecter la capacité de généralisation des modèles d'apprentissage artificiel [a].
- Favoriser la diversité lors de l'optimisation évolutionnaire permet de favoriser l'exploration de l'espace de recherche [b].

→ **Proposition d'une approche pour maximiser la diversité au sein de la population d'EvoMol.**

Pré-requis : Diversité définie à partir d'un descripteur moléculaire.

---

a. Marta Glavatskikh, Jules Leguy et al. "Dataset's chemical diversity limits the generalizability of machine learning predictions". In : *Journal of Cheminformatics* 11.1 (déc. 2019)

b. Y. Tsujimura et M. Gen. "Entropy-based genetic algorithm for solving TSP". In : *1998 Second International Conference. Knowledge-Based Intelligent Electronic Systems. Proceedings KES'98 (Cat. No.98EX111)*. T. 2. Avr. 1998

Descripteur moléculaire : vecteur de *shingles*

Descripteur moléculaire

Transformation d'une représentation moléculaire introduisant des invariances et mettant en évidence des caractéristiques pertinentes pour résoudre un problème donné.

Shingle [a]

Sous-graphe moléculaire de rayon r.

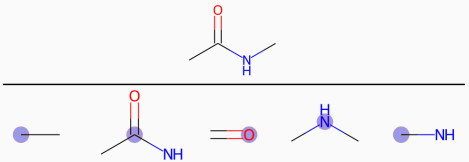


Figure 7 : Molécule de N-methylacetamide et ses 5 *shingles* de rayon 1.

Descripteur : vecteur entier de *shingles* ($r = 1$)

- Vecteur des occurrences des *shingles*.
- Taille : 10^3 caractéristiques.

a. Daniel Probst et Jean-Louis Reymond. "A probabilistic molecular fingerprint for big data settings". In : *Journal of Cheminformatics* 10 (déc. 2018)

Définition d'une mesure de contribution à la diversité

Calcul de la diversité de la population

- Choix de l'entropie de Shannon.
- L'entropie évalue l'ensemble de la population.

Mesure de contribution d'un individu à la diversité

- Nécessaire pour l'utilisation de méthodes d'optimisation de propriétés moléculaires.

Entropie de Shannon [a]

$$H(X) = -\sum_{i=1}^{n} P_i(X) \log P_i(X) + P_i(\overline{X}) \log P_i(\overline{X})$$

$P_i(X)$: proportion d'éléments du jeu de données X contenant la caractéristique d'indice i.

Mesure de contribution à l'entropie de la population

(m remplacé par m' dans X)

$$\Delta_{\text{remplacement}}(m, m', X) = H(X \setminus \{m\} \cup \{m'\}) - H(X)$$

a. C. E. Shannon. "A mathematical theory of communication". In : *The Bell System Technical Journal* 27 (juill. 1948)

Approximation de la contribution à la diversité moléculaire

Limite

- Calcul naïf de $\Delta_{\text{remplacement}}$ coûteux.
  Taille descripteur $\approx 10^3$.
  Taille population X $\approx 10^5$-10^6.

Approximations

1. Caractéristiques absentes ignorées.
2. Traitement par lot :
   - $\Delta_{\text{remplacement}}(m, m', X)$ ne dépend que des caractéristiques de m et m'.
   - Cache des valeurs de $P_i(X) \log P_i(X)$ selon un état fixe de la population.

Entropie de Shannon [b]

$$H(X) = -\sum_{i=1}^{n} P_i(X) \log P_i(X) + P_i(\overline{X}) \log P_i(\overline{X})$$

$P_i(X)$: proportion d'éléments du jeu de données X contenant la caractéristique d'indice i.

Mesure de contribution à l'entropie de la population

(m remplacé par m' dans X)

$$\Delta_{\text{remplacement}}(m, m', X) = H(X \setminus \{m\} \cup \{m'\}) - H(X)$$

- Génération d'un jeu de données avec une diversité élevée.

- Optimisation conjointe de la diversité et de la QED :

$$f_{\text{obj}}(m, m', X) = \text{QED}(m') + \boldsymbol{\omega}\Delta_{\text{remplacement}}(m, m', X)$$

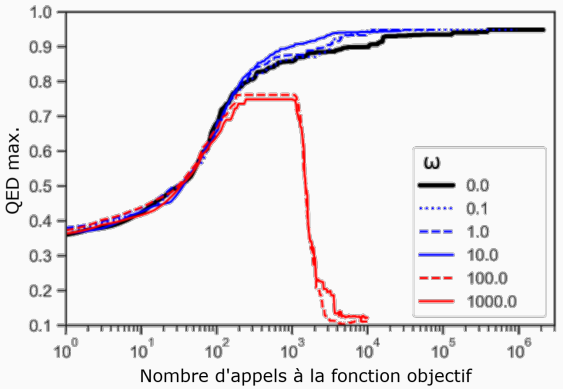


Figure 8 : Meilleure valeur de QED (moyenne parmi 10 exécutions) en fonction du nombre d'appels à la fonction objectif selon la valeur de $\boldsymbol{\omega}$.

Optimisation basée sur un modèle de substitution

Optimisation de propriétés moléculaires coûteuses

Motivation

- Approche évolutionnaire : peut effectuer un grand nombre d'appels à la fonction objectif (pas de connaissance extraite des appels antérieurs).

$\rightarrow$ Nécessité d'une alternative pour l'optimisation de propriétés coûteuses.

Optimisation de propriétés moléculaires coûteuses

Comment réduire le coût de l'optimisation des propriétés moléculaires coûteuses ?

→ Prédiction des propriétés (apprentissage profond, modèles linéaires, GPR, etc.) [a, b, c].

Pourquoi ne pas optimiser les valeurs prédites par un modèle prédictif ?

- Domaine de validité du modèle ? Existe données pertinentes ?
- Problèmes potentiels de convergence si optimisation des valeurs du modèle [d].

→ Cadre de l'optimisation boîte-noire basée sur un modèle de substitution (prédiction).

---

a. Felix A. Faber et al. "Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error". In : *Journal of Chemical Theory and Computation* 13 (nov. 2017)

b. K. T. Schütt et al. "SchNet – A deep learning architecture for molecules and materials". en. In : *The Journal of Chemical Physics* 148.24 (juin 2018), p. 241722

c. Volker L. Deringer et al. "Gaussian Process Regression for Materials and Molecules". In : *Chemical Reviews* 121 (août 2021)

d. Donald R. Jones. "A Taxonomy of Global Optimization Methods Based on Response Surfaces". In : *Journal of Global Optimization* 21.4 (déc. 2001), p. 345-383

Optimisation de propriétés moléculaires coûteuses

Optimisation boîte-noire basée sur un modèle de substitution [a].

- Procédure itérative d'apprentissage et d'optimisation.
- Optimisation d'une fonction de mérite et non du modèle de substitution.
- Approche généralement définie pour des problèmes d'optimisation continue.

a. Ky Khac Vu et al. "Surrogate-based methods for black-box optimization". en. In : *International Transactions in Operational Research* 24.3 (2017)

Optimisation à l'aide d'un modèle de substitution

Algorithme générique pour l'optimisation basée sur un modèle de substitution

entrée : f la fonction objectif boîte-noire, ▷ f fonction coûteuse
s le modèle de substitution de f,
m la fonction de mérite exprimée à partir des valeurs de s

$k \leftarrow 0$
$X_0 \leftarrow$ sélection des données initiales ▷ données couvrant au mieux l'espace de recherche
tant que le critère d'arrêt n'est pas atteint **faire**
$k \leftarrow k + 1$
calcul des valeurs de f pour les données X_{k-1} ▷ coût important
entraînement du modèle s à partir des données $\{(x, f(x)), \forall x \in \bigcup_{i=0}^{i<k} X_k\}$
$X_k \leftarrow$ recherche de solutions prometteuses par résolution de $\max_x m(s(x))$ ▷ coût modéré
fin tant que

Optimisation à l'aide d'un modèle de substitution

Algorithme générique pour l'optimisation basée sur un modèle de substitution

entrée : f la fonction objectif boîte-noire, ▷ fonction coûteuse
s le modèle de substitution de f,
m la fonction de mérite exprimée à partir des valeurs de s

$k \leftarrow 0$
$X_0 \leftarrow$ sélection des données initiales ▷ données couvrant au mieux l'espace de recherche
tant que le critère d'arrêt n'est pas atteint **faire**
$k \leftarrow k + 1$
calcul des valeurs de f pour les données X_{k-1} ▷ coût important
entraînement du modèle s à partir des données $\{(x, f(x)), \forall x \in \bigcup_{i=0}^{i<k} X_k\}$
$X_k \leftarrow$ recherche de solutions prometteuses par résolution de $\mathbf{max}_x\, m(s(x))$ ▷ coût modéré
fin tant que

Optimisation à l'aide d'un modèle de substitution

Algorithme générique pour l'optimisation basée sur un modèle de substitution

entrée : f la fonction objectif boîte-noire, ▷ fonction coûteuse
s le modèle de substitution de f,
m la fonction de mérite exprimée à partir des valeurs de s

$k \leftarrow 0$
$X_0 \leftarrow$ sélection des données initiales ▷ données couvrant au mieux l'espace de recherche
tant que le critère d'arrêt n'est pas atteint **faire**
$k \leftarrow k + 1$
calcul des valeurs de f pour les données X_{k-1} ▷ coût important
entraînement du modèle s à partir des données $\{(x, f(x)), \forall x \in \bigcup_{i=0}^{i<k} X_k\}$
$X_k \leftarrow$ recherche de solutions prometteuses par résolution de $\mathbf{max}_x\, m(s(x))$ ▷ coût modéré
fin tant que

Optimisation à l'aide d'un modèle de substitution

Algorithme générique pour l'optimisation basée sur un modèle de substitution

entrée : f la fonction objectif boîte-noire, ▷ fonction coûteuse
s le modèle de substitution de f,
m la fonction de mérite exprimée à partir des valeurs de s

$k \leftarrow 0$
$X_0 \leftarrow$ sélection des données initiales ▷ données couvrant au mieux l'espace de recherche
tant que le critère d'arrêt n'est pas atteint **faire**
$k \leftarrow k + 1$
calcul des valeurs de f pour les données X_{k-1} ▷ coût important
entraînement du modèle s à partir des données $\{(x, f(x)), \forall x \in \bigcup_{i=0}^{i<k} X_k\}$
$X_k \leftarrow$ recherche de solutions prometteuses par résolution de $\max_x m(s(x))$ ▷ coût modéré
fin tant que

Optimisation à l'aide d'un modèle de substitution

Algorithme générique pour l'optimisation basée sur un modèle de substitution

entrée : f la fonction objectif boîte-noire, ▷ fonction coûteuse
s le modèle de substitution de f,
m la fonction de mérite exprimée à partir des valeurs de s

$k \leftarrow 0$
$X_0 \leftarrow$ sélection des données initiales ▷ données couvrant au mieux l'espace de recherche
tant que le critère d'arrêt n'est pas atteint **faire**
$k \leftarrow k + 1$
calcul des valeurs de f pour les données X_{k-1} ▷ coût important
entraînement du modèle s à partir des données $\{(x, f(x)), \forall x \in \bigcup_{i=0}^{i<k} X_k\}$
$X_k \leftarrow$ recherche de solutions prometteuses par résolution de $\max_x m(s(x))$ ▷ coût modéré
fin tant que

Sélection des données initiales

- Approches de référence pour des espace de recherche continus.
- Pas de solution évidente pour l'espace des graphes moléculaires.
- Heuristique simple.
  - Connaissance minimale (méthane).
  - Sous-ensemble aléatoire d'un jeu de données.

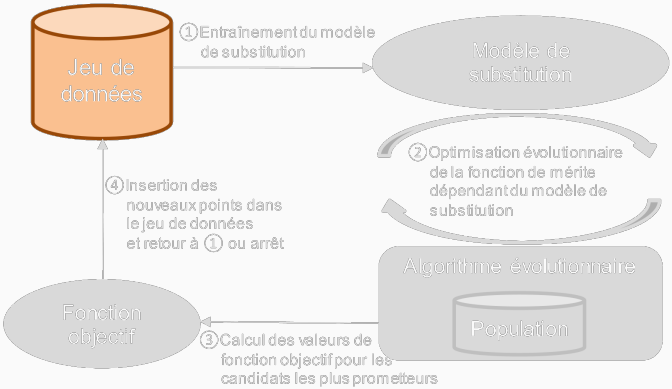


Figure 9 : Représentation schématique du fonctionnement de notre approche d'optimisation basée sur un modèle de substitution.

Modèle de substitution

- Modèle de régression par processus Gaussien (GPR). [a]
- Prédiction d'une distribution gaussienne.
- Dépend d'une fonction noyau.

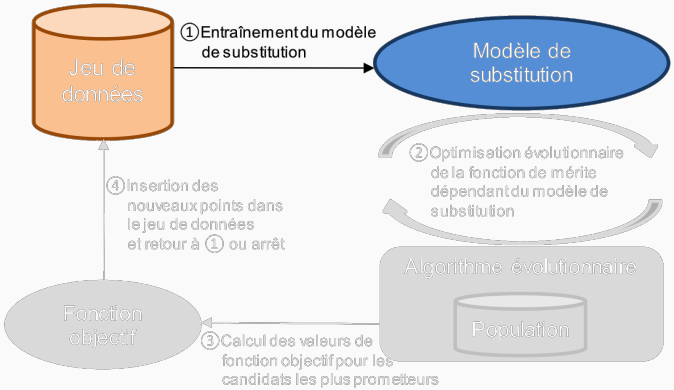


Figure 10 : Représentation schématique du fonctionnement de notre approche d'optimisation basée sur un modèle de substitution.

a. Carl Edward Rasmussen et Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006

Fonction de mérite

- Choix d'une fonction probabiliste.
  - Probability of Improvement (POI).
  - Expected Improvement (EI) : quantification de l'amélioration.

Fonctions de mérite probabilistes

$$\mathrm{POI}(x) = \mathbb{P}(f(x) \geq f(x^+) + \xi)$$
$$\mathrm{EI}(x) = \mathbb{E}[\max(Y - f(x^+) - \xi, 0)]$$

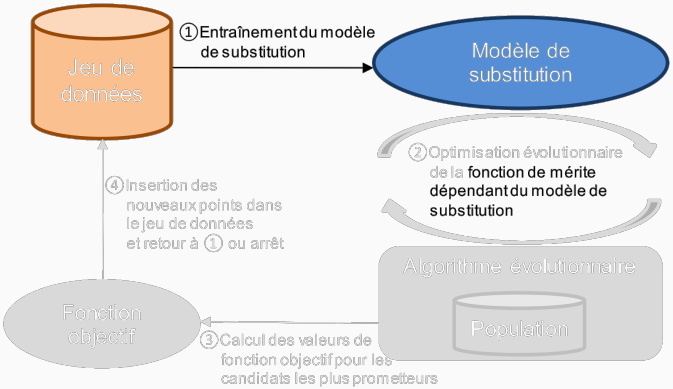


Figure 11 : Représentation schématique du fonctionnement de notre approche d'optimisation basée sur un modèle de substitution.

Optimisation de la fonction de mérite

- Espace de recherche discret.
- Utilisation de notre algorithme EvoMol.

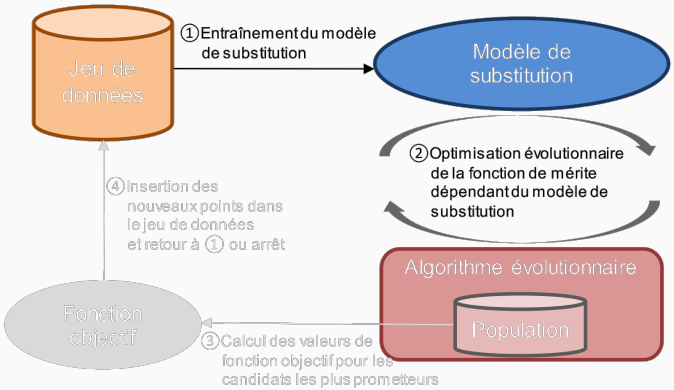


Figure 12 : Représentation schématique du fonctionnement de notre approche d'optimisation basée sur un modèle de substitution.

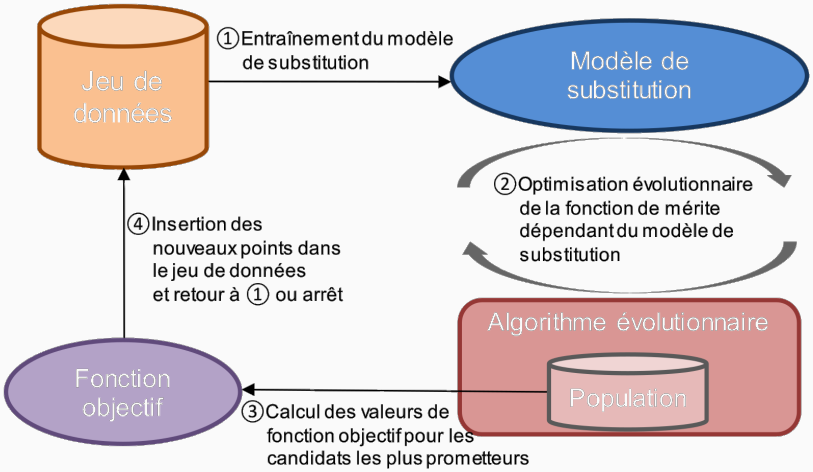


Figure 13 : Représentation schématique du fonctionnement de notre approche d'optimisation basée sur un modèle de substitution.

Evaluation : étude approfondie pour l'optimisation d'une propriété peu coûteuse

- Maximisation des valeurs de QED.
- Descripteur moléculaire : *shingles* ($r = 1$).
- Méthodologie
  1. Étude des performances du modèle de substitution pour la prédiction des valeurs de QED.
  2. Étude de l'influence d'un ensemble de paramètres.

Paramètre	Valeur
Jeu de données initial	Méthane, QM9 [a], ChEMBL [b]
Fonction de mérite	POI, EI, id (identité)
Fonction noyau	k_{RBF}, $k_{\mathrm{DotProduct}}$

Table 1 : Grille de paramètres étudiés pour l'optimisation des valeurs de QED.

a. Raghunathan Ramakrishnan et al. "Quantum chemistry structures and properties of 134 kilo molecules". In : *Scientific Data* 1 (2014)
b. Anna Gaulton et al. "The ChEMBL database in 2017". In : *Nucleic Acids Research* 45.D1 (2017)

Extrait des résultats : maximisation des valeurs de QED

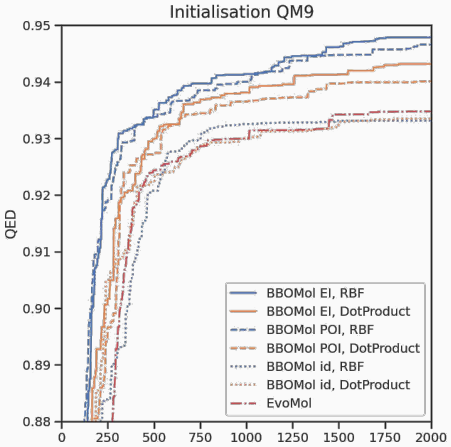


Figure 14 : Meilleure solution (moyenne parmi 10 exécutions) en fonction du nombre d'appels à la fonction objectif.

Evaluation : optimisation d'une propriété moléculaire coûteuse

Descripteur *many-body tensor representation* (MBTR)

- Description de la géométrie moléculaire.
- Somme de fonctions gaussiennes encodant
  - k=1 les numéros atomiques.
  - k=2 les distances.
  - k=3 les angles.

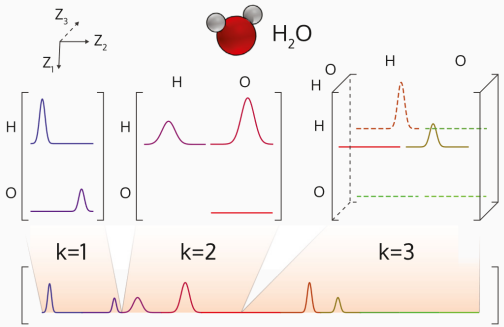


Figure 15 : Représentation du calcul de MBTR pour la molécule d'eau. Figure reproduite de (Himanen et al. 2019)[a]. CC BY

a. Lauri Himanen et al. "DScribe : Library of descriptors for machine learning in materials science". In : *Computer Physics Communications* 247 (fév. 2020)

Évaluation : optimisation d'une propriété moléculaire coûteuse

Maximisation des valeurs d'énergie HOMO

- Jeu de paramètres unique :
  - Fonction de mérite EI.
  - Noyau k_{RBF}.
  - Descripteur MBTR.

Résultats

- Optimisation plus efficace en nombre d'appels à la fonction objectif et en temps de calcul par rapport à EvoMol.

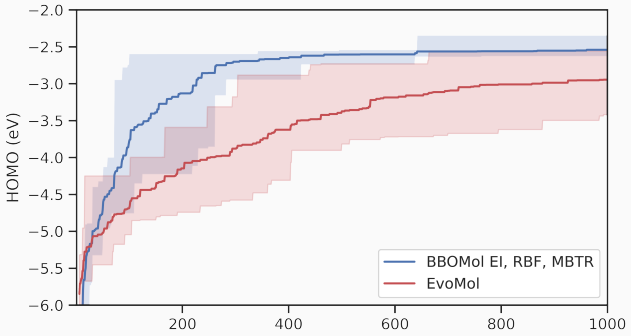


Figure 16 : Meilleure solution (moyenne parmi 10 exécutions) en fonction du nombre d'appels à la fonction objectif. Mise en évidence des valeurs minimales et maximales.

Conclusion et perspectives

Contributions principales et publications

- Algorithme évolutionnaire pour l'optimisation moléculaire [a].
- Approche pour l'optimisation efficace de la diversité moléculaire au sein de la population d'un algorithme évolutionnaire [b].
- Méthode d'optimisation basée sur un modèle de substitution pour l'optimisation efficace de propriétés moléculaires coûteuses [c, d].

---

a. Jules Leguy et al. "EvoMol : a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation". In : *Journal of Cheminformatics* 12.1 (sept. 2020)

b. Jules Leguy et al. "Scalable estimator of the diversity for de novo molecular generation resulting in a more robust QM dataset (OD9) and a more efficient molecular optimization". In : *Journal of Cheminformatics* 13 (oct. 2021), p. 76

c. Jules Leguy et al. "Surrogate-Based Black-Box Optimization Method for Costly Molecular Properties". In : *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. Nov. 2021, p. 780-785

d. Jules Leguy et al. "Surrogate-Based Black-Box Optimization Method for Costly Molecular Properties (Poster)". In : *RSC-CICAG Artificial Intelligence in Chemistry. Diss, UK (Virtual)* (2021)

Contributions supplémentaires et publications

- Filtrage de l'espace de recherche basé sur une liste blanche de caractéristiques pour favoriser le réalisme des solutions : introduction de connaissance [a].
- Application d'EvoMol : génération d'explications contrefactuelles pour l'explication de modèles de classification moléculaire binaire [b].

a. Thomas Cauchy, Jules Leguy et Benoit Da Mota. "Definition and exploration of realistic chemical spaces using the connectivity and cyclic features of ChEMBL and ZINC.". en. In : (déc. 2022). doi : *10.26434/chemrxiv-2022-2b41l*

b. Jules Leguy et al. "Génération d'explications contre-factuelles pour la chimie moléculaire". In : *Atelier EXPLAIN'AI hébergé à EGC 2022*. Jan. 2022

Publications supplémentaires

État de l'art

- Jules Leguy et al. "Chapter 2 - Goal-directed generation of new molecules by AI methods". In : *Computational and Data-Driven Chemistry Using Artificial Intelligence*. Sous la dir. de Takashiro Akitsu. Elsevier, jan. 2022, p. 39-67

Prédiction de la géométrie moléculaire (travaux de master)

- Jules Leguy et al. "Des réseaux de neurones pour prédire des distances interatomiques extraites d'une base de données ouverte de calculs en chimie quantique". In : *Extraction et Gestion des connaissances, EGC 2019, Metz, France, January 21-25, 2019*. 2019, p. 9-20
- Jules Leguy et al. "Predicting Interatomic Distances of Molecular Quantum Chemistry Calculations". In : *Advances in Knowledge Discovery and Management : Volume 9*. Sous la dir. de Rakia Jaziri et al. Studies in Computational Intelligence. Springer International Publishing, 2022, p. 159-174

Impact de la diversité moléculaire sur la qualité des prédictions (travaux de master)

- Marta Glavatskikh, Jules Leguy et al. "Dataset's chemical diversity limits the generalizability of machine learning predictions". In : *Journal of Cheminformatics* 11.1 (déc. 2019)

Perspectives

Perspectives principales

- Apprentissage par renforcement : stratégie dynamique pour la sélection des opérateurs de mutation.
- Stratégie pour la sélection du jeu de données initial pour l'optimisation basée sur un modèle de substitution.

Merci pour votre attention.

Jeux de données : caractéristiques et tailles des molécules

	Molécules			Taille molécules		
	Total	{C, N, O, F}	Charges	Min.	Med.	Max.
QM9	133 885	133 885	1 845	1	9	9
ChEMBL	1 817 795	922 494	128 658	1	27	139

Table 2 : Description des jeux de données QM9 et ChEMBL.

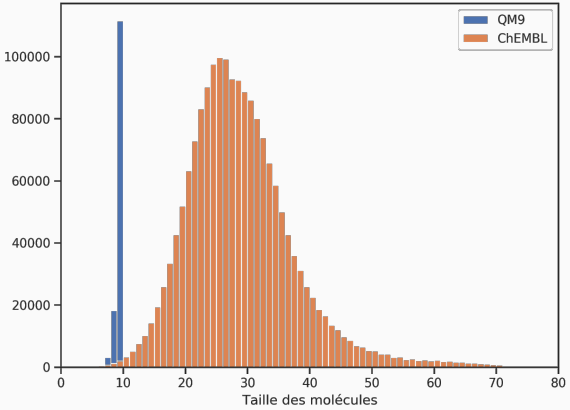


Figure 17 : Distribution des tailles de molécules en fonction du jeu de données.

Jeux de données : distribution des valeurs de propriétés

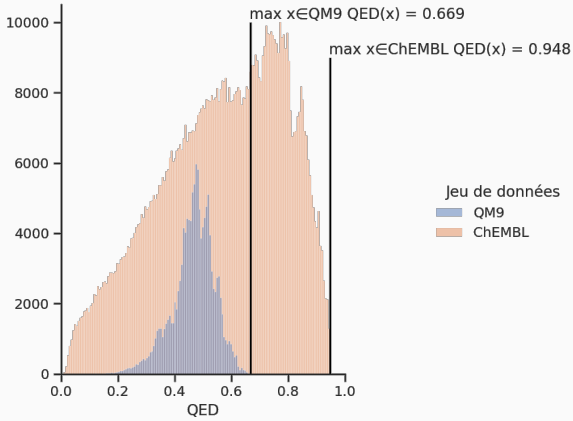


Figure 18 : Distribution des valeurs de QED au sein de QM9 et ChEMBL.

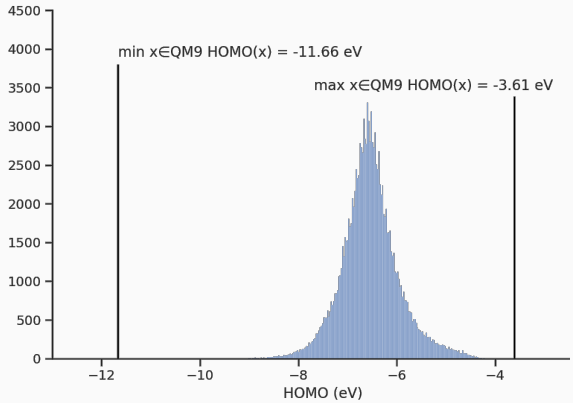


Figure 19 : Distribution des valeurs d'énergie HOMO (eV) au sein de QM9.

Arbres d'exploration (1/2)

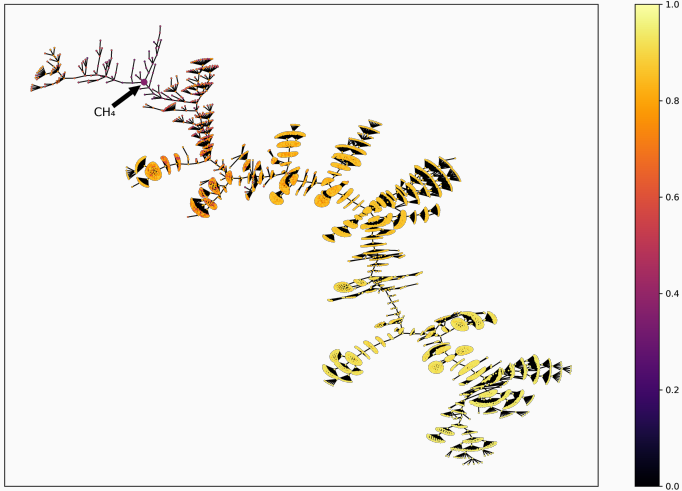


Figure 20 : Arbre d'exploration pour la maximisation de la valeur de QED. Une arête représente la mutation liant deux solutions. Toutes les solutions qui ont fait partie de la population sont représentées. Le point de départ (méthane) est indiqué à l'aide d'une flèche. Les solutions sont colorées selon leur valeur de QED.

Arbres d'exploration (2/2)

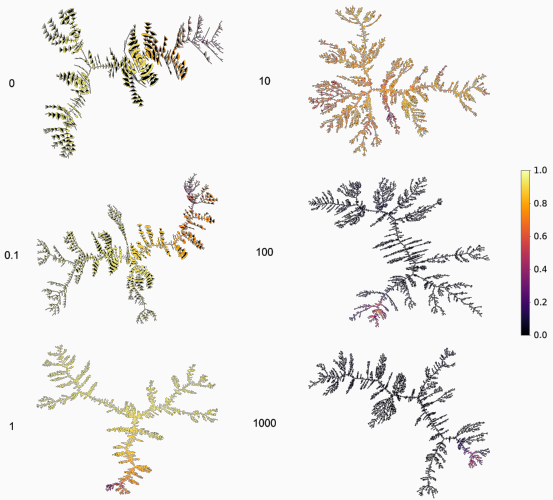


Figure 21 : Arbres d'exploration lors de l'optimisation conjointe de la QED et de la diversité des shingles de niveau 1 pour différentes valeurs de ω.

Filtrage de l'espace de recherche

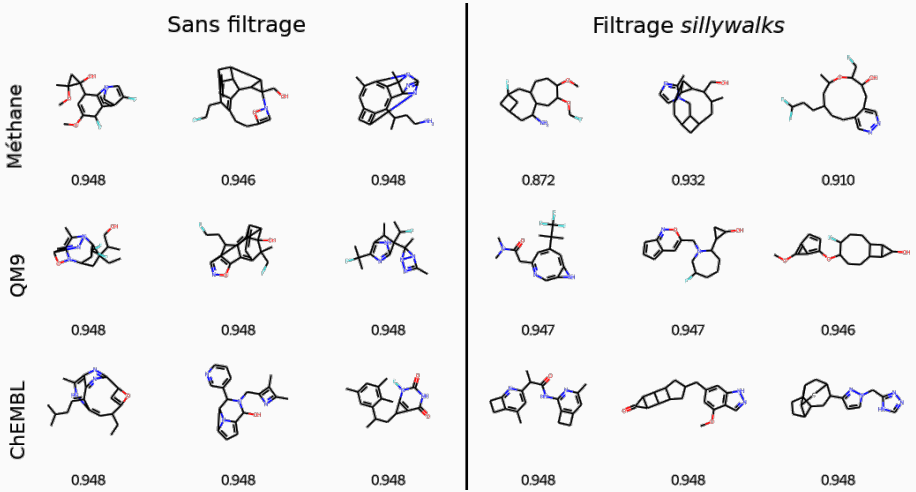


Figure 22 : Exemples de solutions obtenues par BBOMol avec et sans application de la contrainte sillwyalks en fonction de la stratégie d'initialisation. Les résultats affichés correspondent à l'utilisation de la fonction de mérite EI et de la fonction noyau k_{RBF}.

Filtrage de l'espace de recherche (filtrage des caractéristiques cycliques)

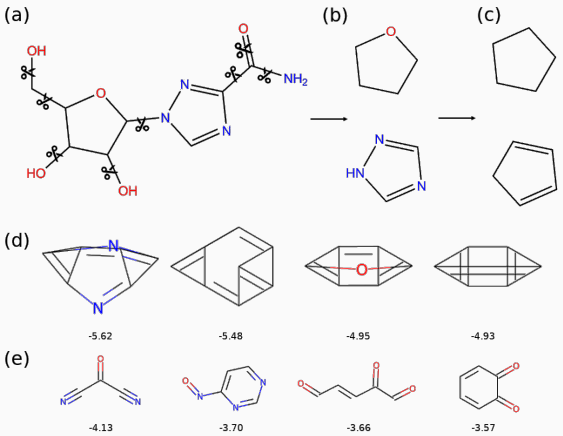


Figure 23 : Représentation du processus de génération des caractéristiques GCF (a, b, c) et étude de l'application de la contrainte GCF (d, e). La partie (d) correspond aux 4 meilleures solutions obtenues pour la minimisation de l'énergie HOMO avec la contrainte sillywalks, et la partie (e) correspond aux 4 meilleures solutions obtenues pour le même problème avec la contrainte sillywalks et la contrainte GCF.

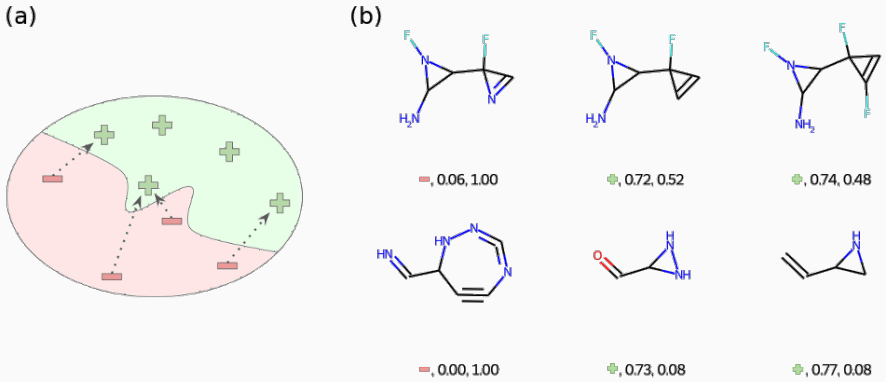


Figure 24 : (a) : Représentation schématique de la frontière de décision d'un modèle de classification binaire séparant un ensemble d'instances négatives et positives. (b) : Exemples d'application de notre approche pour générer des explications d'un modèle prédictif de la stabilité moléculaire. Chaque molécule est étiquetée selon un triplet (classe prédite, valeur prédite par le modèle, similarité avec la molécule de départ).

Approximation du calcul de contribution à la diversité

Calcul de l'entropie

$$H(X) = -\sum_{i=1}^{n} P_i(X) \log P_i(X)$$

Calcul de l'entropie (reformulation)

$$H(X) = \sum_{i=1}^{N} H(D_i, X)$$

$$H(D_i, X) = -\frac{C_i(X)}{|X|} \log \frac{C_i(X)}{|X|}$$

Contribution des caractéristiques

$$\delta_{\text{suppr}}(D_i, x, X) = H(D_i, X \setminus \{x\}) - H(D_i, X)$$

$$\Delta'_{\text{suppr}}(x, X) = \sum_{D_i \in X} \delta_{\text{suppr}}(D_i, x, X)$$

$$\delta_{\text{ajout}}(D_i, x, X) = H(D_i, X \cup \{x\}) - H(D_i, X)$$

$$\Delta'_{\text{ajout}}(x, X) = \sum_{D_i \in X} \delta_{\text{ajout}}(D_i, x, X)$$

$$\Delta'_{\text{remplacement}}(x_s, x_a, X) = \Delta'_{\text{suppr}}(x_s \setminus x_a, X) + \Delta'_{\text{ajout}}(x_a \setminus x_s, X)$$

Gain efficacité : Valeurs de δ_{suppr} et δ_{ajout} en cache pour tous les descripteurs pour un état donné de la population.

Apprentissage des valeurs de QED (1/3)

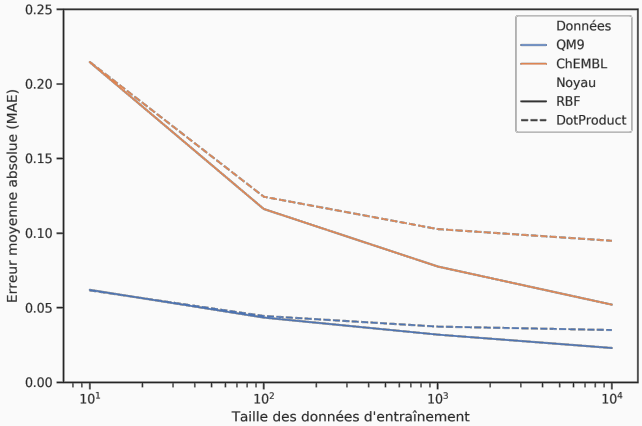


Figure 25 : Erreur moyenne absolue pour la prédiction de la valeur de QED à l'aide du modèle GPR en fonction du jeu de données (QM9 ou ChEMBL), de la fonction noyau, et de la taille du jeu de données d'entraînement.

Apprentissage des valeurs de QED (2/3)

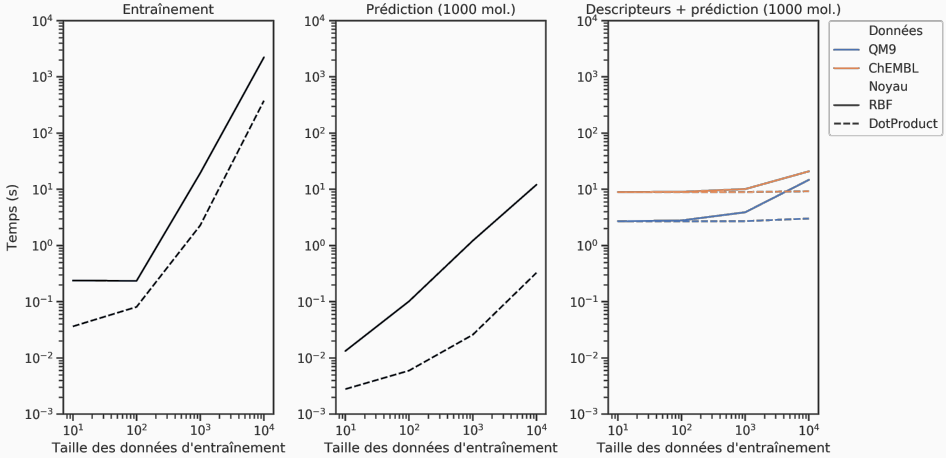


Figure 26 : À gauche : temps pour l'entraînement du modèle de prédiction de la QED. Au centre : temps pour la prédiction de la valeur de QED de 1000 molécules. À droite : temps pour le calcul des descripteurs et la prédiction de la valeur de QED pour 1000 molécules.

Apprentissage des valeurs de QED (3/3)

Fonction noyau	MAE inter-modèles			
	Moyenne		Écart-type	
	QM9	ChEMBL	QM9	ChEMBL
k_{RBF}	0.031	0.078	0.000	0.001
$k_{\mathrm{DotProduct}}$	0.037	0.103	0.000	0.001

Table 3 : Moyenne et écart-type des valeurs de MAE pour la prédiction des valeurs de QED, mesurées sur les différents plis en fonction du jeu de données et de la fonction noyau.

Fonction noyau	Erreur intra-modèle			
	Moyenne		Écart-type	
	QM9	ChEMBL	QM9	ChEMBL
k_{RBF}	0.001	-0.005	0.041	0.100
$k_{\mathrm{DotProduct}}$	0.001	-0.006	0.047	0.129

Table 4 : Moyenne et écart-type des erreurs enregistrées pour le modèle entraîné sur le premier pli pour la prédiction des valeurs de QED.

Apprentissage des valeurs d'énergie HOMO (1/3)

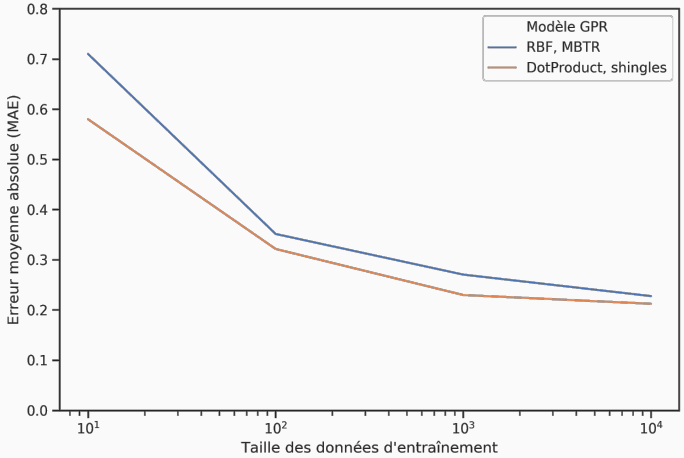


Figure 27 : Erreur moyenne absolue pour la prédiction de la valeur d'énergie HOMO en eV en fonction du modèle et de la taille du jeu de données d'entraînement, sur le jeu de données QM9.

Apprentissage des valeurs d'énergie HOMO (2/3)

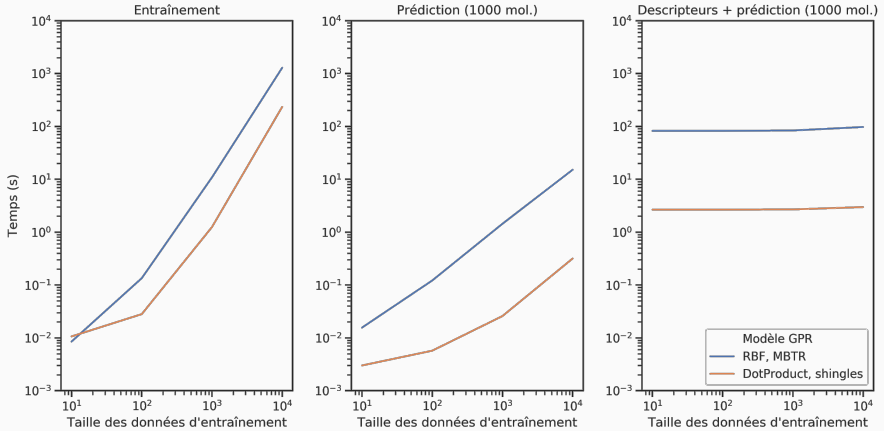


Figure 28 : À gauche : temps pour l'entraînement des modèles de prédiction de l'énergie HOMO. Au centre : temps pour la prédiction de la valeur d'énergie HOMO de 1000 molécules. À droite : temps pour le calcul des descripteurs et la prédiction de la valeur d'énergie HOMO pour 1000 molécules.

Apprentissage des valeurs d'énergie HOMO (3/3)

Modèle	MAE inter-modèles	
	Moyenne	Écart-type
GPR k_{RBF}, MBTR	0.27	0.00
GPR $k_{\mathrm{DotProduct}}$, shingles	0.23	0.00

Table 5 : Moyenne et écart-type des valeurs de MAE mesurées sur les différents plis en fonction du modèle pour la prédiction des valeurs d'énergie HOMO en eV. L'erreur moyenne du modèle SchNet est également reportée.

	Erreur intra-modèle (eV)	
Fonction noyau	Moyenne	Écart-type
k_{RBF}	0.01	0.37
$k_{\mathrm{DotProduct}}$	0.01	0.30

Table 6 : Moyenne et écart-type en eV des erreurs enregistrées pour le modèle entraîné sur le premier pli pour la prédiction des valeurs de HOMO (paramètre $\sigma_n^2 = 0.1$).

Maximisation des valeurs de QED (BBOMOl)

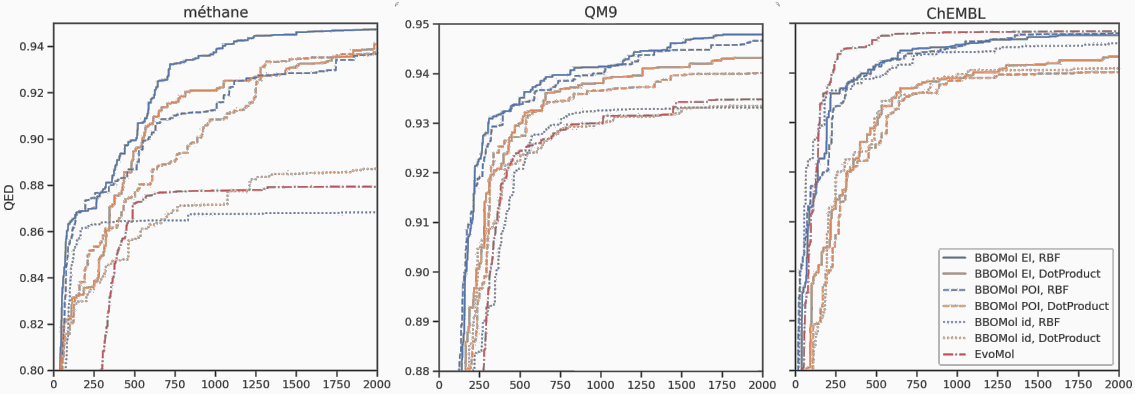


Figure 29 : Moyenne des meilleurs scores obtenus en fonction du nombre d'appels à la fonction objectif, pour différents paramétrages de BBOMol et pour EvoMol.

Maximisation des valeurs de QED (BBOMOl) : paramètre d'exploration

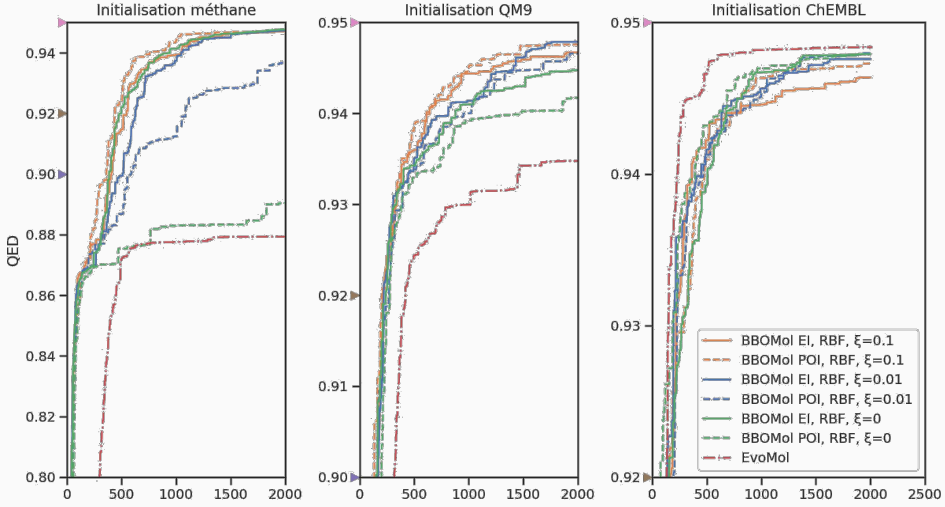


Figure 30 : Étude du paramètre d'exploration ξ. Moyenne des meilleurs scores obtenus en fonction du nombre d'appels à la fonction objectif, pour différents paramétrages de BBOMol et pour EvoMol.

Maximisation des valeurs d'énergie HOMO (BBOMol)

Méthode	ERT (cible -3.0 eV)		Succès (/10)
	Appels à f	Temps de calcul (h)	
BBOMol EI, k_{RBF}, MBTR	177	11.3	10
EvoMol	1186	74.2	6

Table 7 : Mesure d'espérance du coût de l'exécution (ERT) pour l'obtention d'une solution possédant une énergie HOMO supérieure à -3.0 eV en nombre d'appels à la fonction objectif *f* ou en temps de calcul. La colonne succès représente le nombre de fois que la cible a été atteinte parmi 10 exécutions.

Calcul de la mesure d'ERT (expected runtime)

$$\mathrm{ERT}(X, c) = \frac{\sum_{x \in X} \mathrm{coût\_min}(x, c)}{\sum_{x \in X} \mathrm{cible\_atteinte}(x, c)}$$

- X : ensemble d'exécutions devant être évalué.
- c : valeur cible de fonction objectif.
- $\mathrm{coût\_min}$: coût dépensé par l'exécution x pour obtenir la première solution avec une valeur de fonction $\geq$ à c.
- $\mathrm{cible\_atteinte}$ renvoie 1 si la cible c a été atteinte par l'exécution x et 0 sinon.