Sorbonne Université

Rapport de Stage

# Conformation sampling of small molecules for molecular dynamics of large molecules

**Auteur :**
Shuyun Deng

**Référent :**
Grégoire Sergeant-Perthuis
**Co-supervisé par :**
Léo Boitel

24 juin 2025

# Table des matières

# 1 Introduction et objectifs

La dynamique moléculaire classique génère des trajectoires atomiques : à chaque pas de temps $\Delta t$, typiquement de l'ordre de 1 à 2 femtosecondes (fs), on obtient les coordonnées de tous les atomes du système ainsi que les énergies et forces associées. Par exemple, des outils tels que GROMACS, utilisant le champ de force AMBER, permettent de simuler ces trajectoires et d'accéder à l'énergie potentielle, à la topologie et aux coordonnées atomiques à chaque instant [1].

Pourtant, pour explorer plus largement la diversité structurale du système, c'est-à-dire simuler les changements de conformation majeurs, les approches classiques de MD montrent rapidement leurs limites. Les transitions entre états conformationnels séparés par des barrières d'énergie libre élevées sont rarement échantillonnées dans une simulation directe, même longue. Pour contourner ce problème, on utilise une méthode appelée Replica Exchange (RE), qui consiste à faire évoluer en parallèle plusieurs copies du système à différentes températures et à permettre des échanges d'états entre elles [5]. Cela permet de franchir plus efficacement les barrières énergétiques et d'obtenir un meilleur contrôle sur l'échantillonnage des conformations. Cependant, malgré son efficacité, RE présente des limitations majeures pour les grands systèmes : le nombre de répliques nécessaires augmente rapidement avec la taille du système, rendant son application très coûteuse, voire irréaliste pour des systèmes de type ribosome [6].

Face à ces limitations, des approches alternatives se développent. Parmi elles, les modèles de diffusion se révèlent prometteurs pour explorer plus efficacement le paysage conformationnel. L'objectif de ce stage est d'expérimenter l'utilisation d'un modèle de diffusion pour générer des transitions conformationnelles plausibles à partir d'une configuration initiale. Cette méthode sera d'abord évaluée sur l'alanine dipeptide, système modèle présentant des transitions bien caractérisées, avant d'envisager son extension à des systèmes plus complexes. Enfin, une perspective majeure de ce travail est d'étendre cette approche aux systèmes de grande taille, tels que le ribosome. L'idée serait de découper le système global en sous-ensembles plus petits, d'appliquer des modèles de diffusion pour échantillonner efficacement les conformations de chaque sous-ensemble, puis de recoller ces "blocs" en utilisant des algorithmes de passage de message [2, 3, 4]afin de reconstruire l'ensemble du paysage conformationnel à l'échelle du système entier.

# 2 Matériels et méthodes

## 2.1 Préparation des données

Les données utilisées proviennent du dépôt *mdshare*, plus précisément du jeu de données « alanine dipeptide ». L'alanine dipeptide est une petite molécule de 22 atomes, couramment utilisée comme système modèle en dynamique moléculaire. Le dataset inclut 250 000 images (frames) issues de 5 trajectoires indépendantes de 50 000 frames chacune, simulées à 300 K avec le champ de force AMBER. Chaque frame fournit les coordonnées atomiques (`.xtc`) et la structure de référence (`.pdb`). Les angles dièdres $\phi$ (C-N-$C_\alpha$-C) et $\psi$ (N-$C_\alpha$-C-N), qui constituent les coordonnées réactionnelles principales de l'alanine dipeptide et caractérisent sa diversité conformationnelle, sont extraits à l'aide de la bibliothèque `mdtraj`, ce qui permet d'obtenir un jeu de données de dimension $(N, 2)$ (ici $N = 250\,000$) utilisé comme entrée du modèle.

## 2.2 Modélisation par Diffusion (DDPM)

Le modèle de diffusion est entraîné à partir des couples $\phi/\psi$. L'ensemble du pipeline peut se décomposer en trois étapes principales (Figure 1) :

1. **Forward diffusion** : Un bruit gaussien est progressivement ajouté aux configurations initiales $\tau_0$, générant une suite de vecteurs bruités $\tau_t$ pour différents pas de temps $t$. Cette étape simule la dégradation contrôlée des données selon l'équation différentielle stochastique (SDE) suivante :

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{w} \tag{1}$$

où $\mathbf{w}$ est un processus de Wiener standard, c'est-à-dire un mouvement brownien d'espérance nulle et de variance croissante linéairement avec le temps ($\mathbb{E}[w(t)] = 0$, $\mathrm{Var}[w(t)] = t$). Ce processus est couramment

utilisé pour modéliser le mouvement aléatoire de particules fines (mouvement brownien) dans un milieu fluide. $f(\cdot,t):\mathbb{R}^d\to\mathbb{R}^d$ est le coefficient de dérive et $g(\cdot):\mathbb{R}\to\mathbb{R}$ est le coefficient de diffusion. Dans notre cas spécifique avec les angles dièdres, le processus forward transforme progressivement les données structurées en bruit gaussien.

2. **Score matching** : Un réseau de neurones profond est entraîné à prédire le bruit ajouté à chaque étape de diffusion, en se basant sur les paires $(\tau_t, t)$. L'objectif d'entraînement utilise une généralisation continue des objectifs de débruitage [7] :

$$\theta^* = \arg\min_\theta \mathbb{E}_t\left[\lambda(t)\mathbb{E}_{\mathbf{x}(0)}\mathbb{E}_{\mathbf{x}(t)|\mathbf{x}(0)}\left[\left\|s_\theta(\mathbf{x}(t),t)-\nabla_{\mathbf{x}(t)}\log p_{0t}(\mathbf{x}(t)|\mathbf{x}(0))\right\|_2^2\right]\right] \tag{2}$$

où $\lambda:[0,T]\to\mathbb{R}_+$ est une fonction de pondération positive, $t$ est échantillonné uniformément sur $[0,T]$, $\mathbf{x}(0)\sim p_0(\mathbf{x})$ provient de la distribution des données, $\mathbf{x}(t)\sim p_{0t}(\mathbf{x}(t)|\mathbf{x}(0))$ suit le noyau de transition, et $s_\theta(\mathbf{x}(t),t)$ est la prédiction du réseau de neurones pour le score. Le terme $p_{0t}(x(t)|x(0))$ représente le noyau de transition issu de l'équation de diffusion forward (équation 1), c'est-à-dire la probabilité d'observer $x(t)$ à partir d'une condition initiale $x(0)$ au temps $t$. Cette approche, dite score-based generative modeling with SDEs [7], permet d'estimer la direction optimale à suivre pour restaurer les données originales à partir d'un état bruité.

3. **Reverse diffusion** : Une fois le réseau entraîné, on utilise le processus inverse selon la SDE reverse-time [7] :

$$d\mathbf{x} = \left[f(\mathbf{x},t) - g(t)^2\nabla_{\mathbf{x}}\log p_t(\mathbf{x})\right]dt + g(t)d\bar{\mathbf{w}} \tag{3}$$

où $\bar{\mathbf{w}}$ est un processus de Wiener standard lorsque le temps s'écoule vers l'arrière de $T$ à 0. Cette SDE inverse correspond à la dynamique optimale pour générer des échantillons réalistes à partir du bruit, telle que démontrée rigoureusement dans [7]. L'ajout du terme de score $\nabla_{\mathbf{x}}\log p_t(\mathbf{x})$ assure la convergence vers la distribution réelle des données. À partir d'un bruit aléatoire, le modèle applique séquentiellement les corrections prédites par $s_\theta$ pour générer de nouveaux échantillons réalistes d'angles dièdres. Chaque étape consiste à débruiter la configuration précédente selon la direction calculée.

Dans le contexte spécifique des modèles de diffusion torsionnelle, la fonction de perte de débruitage est définie comme :

$$J_{\mathrm{DSM}}(\theta) = \mathbb{E}_t\left[\lambda(t)\mathbb{E}_{\tau_0\sim p_0,\tau_t\sim p_{t|0}(\cdot|\tau_0)}\left[\left\|s(\tau_t,t)-\nabla_{\tau_t}\log p_{t|0}(\tau_t|\tau_0)\right\|^2\right]\right] \tag{4}$$

où le niveau de bruit $t$ est échantillonné uniformément, $\lambda(t) = 1/\mathbb{E}_{\tau\sim p_{t|0}(\cdot|0)}\left[\left\|\nabla_{\tau_t}\log p_{t|0}(\tau|0)\right\|_2^2\right]$, $s(\tau_t,t)$ est la prédiction du réseau de neurones pour le score et $p_{t|0}(\tau_t|\tau_0)$ est le noyau de perturbation.

Cette approche permet donc de modéliser et de générer la diversité conformationnelle sans recourir explicitement à des trajectoires MD classiques.

## 2.3 Reconstruction et simulations supplémentaires

Les couples $\phi/\psi$ générés par le modèle de diffusion sont convertis en structures 3D complètes à l'aide de la bibliothèque **Chem** (RDKit). Le processus de reconstruction se décompose en plusieurs étapes techniques :

### 2.3.1 Génération des conformations 3D et re-pondération thermodynamique

Pour chaque jeu de données, 100 conformations 3D sont générées à partir des couples d'angles dièdres $(\phi,\psi)$ simulés. La molécule d'alanine dipeptide est reconstruite à l'aide de la bibliothèque `RDKit`, les liaisons rotatoires sont identifiées automatiquement, et chaque angle généré est appliqué via la fonction `AllChem.SetDihedralDeg()`. Chaque conformation est ensuite optimisée géométriquement afin d'éliminer les collisions stériques.

Les conformations obtenues sont évaluées énergétiquement à l'aide du champ de force MMFF94. Les poids de Boltzmann sont alors attribués selon :

$$w_i = \frac{\exp(-E_i/k_BT)}{\sum_{j=1}^N \exp(-E_j/k_BT)} \tag{5}$$

où $E_i$ est l'énergie MMFF94 de la conformation $i$ et $k_BT = 0{,}593$ kcal/mol à 298K.

Un nouvel ensemble de 100 conformations est ensuite rééchantillonné en respectant cette distribution de poids, garantissant ainsi un échantillonnage thermodynamique pertinent, où les conformations de plus basse énergie sont statistiquement favorisées.

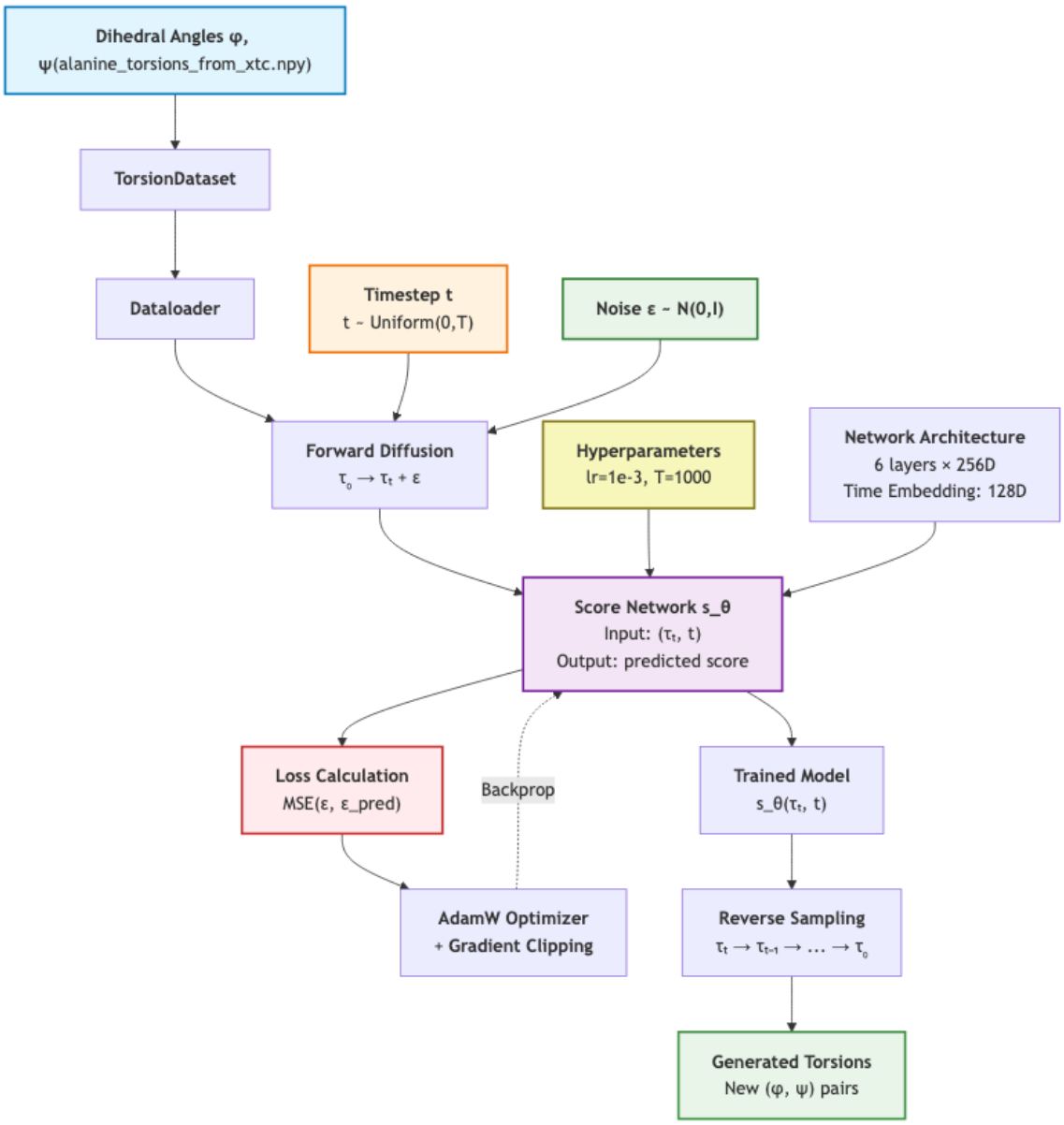


FIGURE 1 – Architecture du pipeline DDPM utilisé pour la modélisation des angles dièdres moléculaires. Le schéma présente les différentes étapes du flux de données : **(1) TorsionDataset** — le jeu de données des angles dièdres est chargé par le *dataloader* ; **(2) Forward diffusion** — un bruit gaussien $\epsilon \sim \mathcal{N}(0,1)$ est ajouté aux angles à chaque pas de temps $t$ tiré uniformément sur $[0,T]$ ; **(3) Score network** $s_\theta$ — un réseau de neurones profond reçoit en entrée les angles bruités et le pas de temps, puis prédit le bruit ajouté ; **(4) Calcul de la perte (MSE)** — la différence entre le bruit ajouté et prédit est minimisée par rétropropagation du gradient (backpropagation), à l'aide de l'optimiseur AdamW ; **(5) Reverse sampling** — une fois le modèle entraîné, le réseau permet de générer de nouveaux échantillons d'angles $(\phi, \psi)$ réalistes par le processus de diffusion inverse. Les hyperparamètres utilisés (par exemple $T = 1000$, structure du réseau) sont également indiqués dans le schéma.

### 2.3.2 Validation par simulations MD

À partir des structures re-pondérées, 100 conformations sont sélectionnées par échantillonnage pondéré, puis chacune est soumise à une courte simulation de dynamique moléculaire en parallèle [8] à l'aide d'OpenMM et du champ de force Amber. Chaque simulation comprend la solvatation explicite de la molécule dans une boîte d'eau TIP3P, une minimisation d'énergie initiale (1000 pas), une phase d'équilibration en conditions NPT (100 ps à 298 K et 1 atm), suivie d'une production de dynamique moléculaire de 1 ns avec un pas de temps de 2 fs.

Ce protocole permet de valider la stabilité structurale des conformations générées et de comparer leur évolution énergétique aux trajectoires issues des simulations classiques.

# 3 Résultats

## 3.1 Diffusion model training loss et génération du 3D structure

La courbe de la loss MSE (Mean Squared Error) au cours de l'entraînement du modèle de diffusion reste stable après une phase initiale de descente rapide, ce qui indique une bonne convergence du modèle pour la tâche de débruitage des angles dièdres $\phi/\psi$ (Figure 2). Par la suite, j'ai utilisé les angles dièdres prédits par le modèle entraîné pour reconstruire la structure 3D complète de la molécule. Après génération, la structure est solvatée dans l'eau TIP3P, ce qui prépare le système pour les simulations de dynamique moléculaire ultérieures. Figure 3 présente une conformation typique de l'alanine dipeptide obtenue par cette méthode. On constate que la structure obtenue et son environnement solvaté sont cohérents avec ceux générés directement à partir d'un fichier `.pdb` suivi de la même solvatation.

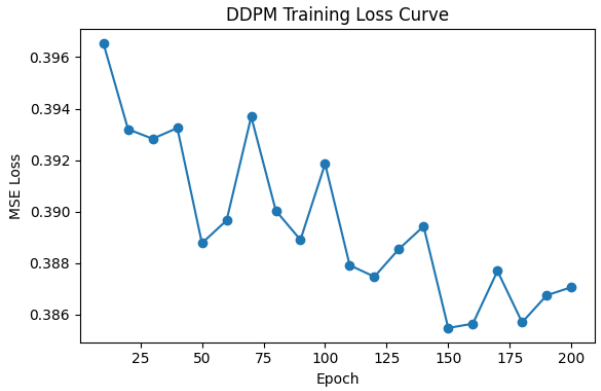


FIGURE 2 – Courbe d'apprentissage du modèle DDPM : la loss MSE (Mean Squared Error, en ordonnée) représente l'erreur quadratique moyenne entre le bruit ajouté et le bruit prédit par le réseau de neurones lors du débruitage des angles dièdres $(\phi, \psi)$. Chaque point correspond à une évaluation sur le jeu d'entraînement à la fin de chaque époque (abscisse, Epoch). Une diminution rapide suivie d'une stabilisation de la loss indique une bonne convergence du modèle.

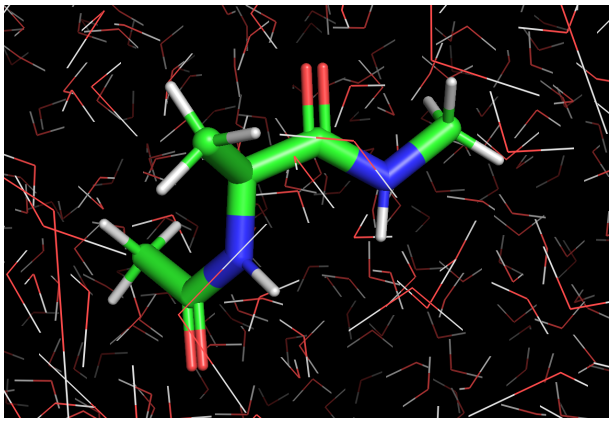

FIGURE 3 – Structure 3D typique de l'alanine dipeptide générée par le modèle. Les couleurs représentent : carbone (vert), hydrogène (blanc), oxygène (rouge), azote (bleu). L'environnement solvaté correspond aux molécules d'eau TIP3P en arrière-plan.

## 3.2 Comparaison de multiples trajectoires générées

Dans la deuxième partie, j'ai comparé les trajectoires de 100 échantillons générés à partir de trois groupes d'`.xtc` différents. Pour chaque ensemble, j'ai analysé le potentiel d'énergie (Figure 4) et le RMSD (Figure 5) tout au long des simulations. Les courbes obtenues pour ces deux indicateurs se superposent parfaitement : les trajectoires issues de trois entraînements indépendants présentent une évolution énergétique et un comportement structural remarquablement similaires. Ces résultats démontrent la stabilité de l'entraînement du modèle de diffusion, ainsi que la reproductibilité des trajectoires générées par l'échantillonnage du modèle, validant ainsi la robustesse de l'architecture que j'ai mise en place dans le cadre de ce travail.

## 3.3 Analyses des TICA

TICA (Time-lagged Independent Component Analysis) est une méthode de réduction de dimension spécialement adaptée à l'analyse des trajectoires de dynamique moléculaire. Elle identifie les coordonnées collectives (TICs) qui capturent les processus dynamiques les plus lents du système, c'est-à-dire ceux associés aux transitions entre états conformationnels séparés par de hautes barrières d'énergie libre.

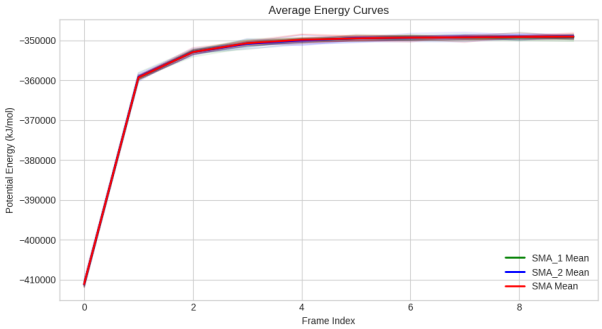


FIGURE 4 – Courbe du potentiel d'énergie moyen : L'axe des abscisses ("Frame index") correspond au numéro de l'image (frame) dans la trajectoire moléculaire analysée.

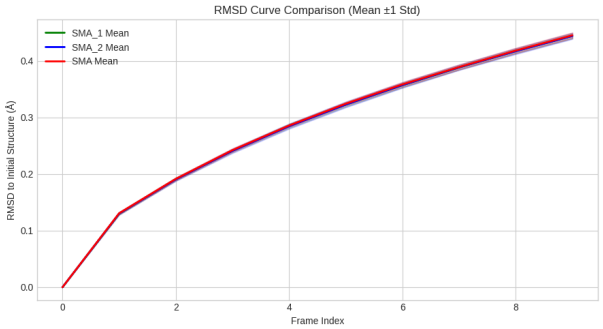


FIGURE 5 – Comparaison des courbes RMSD pour les trois ensembles : L'axe des abscisses ("Frame index") correspond au numéro de l'image (frame) dans la trajectoire moléculaire analysée.

Concrètement, TICA consiste à calculer la matrice de covariance retardée (auto-covariance temporelle) à partir des trajectoires, puis à diagonaliser cette matrice pour extraire les composantes principales indépendantes (TIC0, TIC1, etc.), classées selon leur temps de relaxation décroissant.

Mathématiquement, TICA consiste à résoudre le problème aux valeurs propres généralisé suivant :

$$\mathbf{C}_\tau \mathbf{w} = \lambda \mathbf{C}_0 \mathbf{w}$$

où $\mathbf{C}_0 = \langle \mathbf{x}_t \mathbf{x}_t^\top \rangle$ est la matrice de covariance instantanée, $\mathbf{C}_\tau = \langle \mathbf{x}_t \mathbf{x}_{t+\tau}^\top \rangle$ est la matrice de covariance retardée avec le délai temporel $\tau$, et $\mathbf{w}$ sont les vecteurs propres correspondant aux composantes indépendantes lentes (TICs).

La projection des trajectoires ou des échantillons générés sur ces axes permet d'évaluer la diversité et l'exploration de l'espace conformationnel, en distinguant les différents modes d'échantillonnage explorés par le modèle. La projection du free energy le long de TIC0 (Figure 6) suggère que la trajectoire SMA-0-MD couvre la plage la plus large de coordonnées, ce qui pourrait traduire une exploration efficace des états conformationnels séparés par de hautes barrières d'énergie libre. Les trajectoires SMA-1-MD et SMA-2-MD, bien que plus localisées, semblent également traverser différentes régions stables, ce qui tend à indiquer que le modèle serait capable de franchir les barrières énergétiques ; La projection conjointe (Figure 7) montre que chaque échantillon semble couvrir un sous-espace conformationnel distinct, ce qui pourrait refléter la diversité des états générés par le modèle de diffusion. Pris ensemble, ces résultats laissent penser que le modèle est capable de générer une variété de conformations et d'explorer efficacement l'espace conformationnel du système.

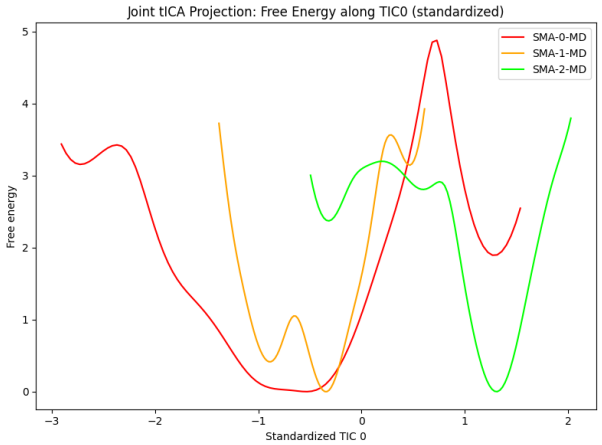


FIGURE 6 – Profil d'énergie libre le long de la première composante lente (TIC0)

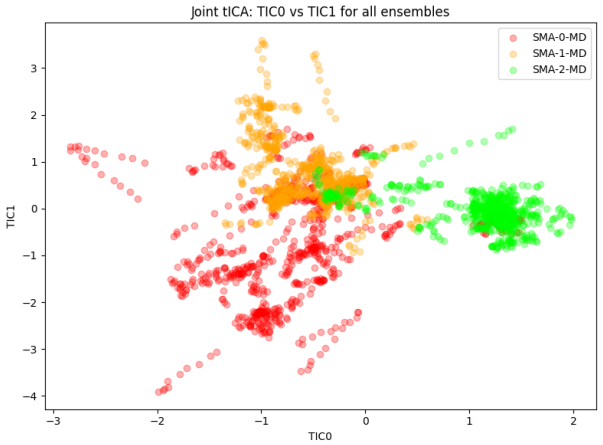


FIGURE 7 – Projection conjointe des trois ensembles (TIC0, TIC1)

# 4 Discussion

Afin de valider la pertinence de mes analyses, j'ai directement extrait les angles dièdres à partir des fichiers `.xtc` d'origine pour réaliser une projection tICA "contrôle" (groupe de référence). En théorie, les courbes de l'énergie libre (Figure 8) obtenues pour chaque simulation originale devraient parfaitement se superposer, chaque échantillon couvrant un même sous-espace conformationnel, avec une distribution plus large et homogène dans l'espace lent (Figure 9).

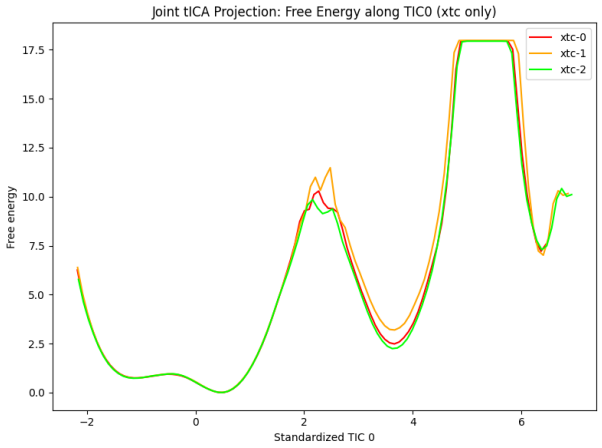


FIGURE 8 – Profil d'énergie libre obtenu par projection tICA directe des données originales. L'axe des abscisses correspond à la première composante indépendante (TIC0), l'axe des ordonnées à l'énergie libre (en unités arbitraires).

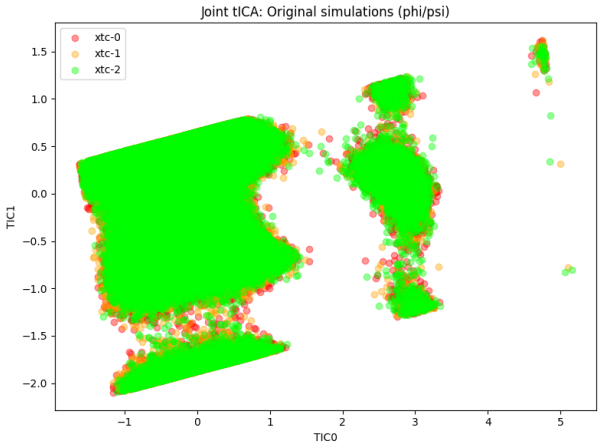


FIGURE 9 – Projection conjointe (TIC0 vs TIC1) des trois trajectoires originales. Chaque point représente une conformation projetée dans l'espace des deux premières composantes indépendantes issues de l'analyse tICA.

Cependant, lorsqu'on compare les résultats obtenus avec ceux des structures générées par le modèle de diffusion, on observe une diversité légèrement inférieure et des sous-espaces moins recouverts, aussi bien sur les courbes d'énergie libre que sur la projection conjointe. Ce constat demeure même après vérification de la validité structurale (PyMOL, potentiel d'énergie, RMSD) des conformations générées.

Cette différence ne semble pas résulter de la qualité des structures générées ni d'un défaut du modèle, mais s'explique principalement par la limitation du nombre de trajectoires produites et du nombre de frames échantillonnées : le modèle n'a pas encore totalement couvert tout l'espace conformationnel exploré par les simulations originales. Ce résultat met en évidence la nécessité d'augmenter le nombre de frames (ou la longueur) de chaque trajectoire générée, ou d'optimiser la stratégie d'échantillonnage, afin d'améliorer la représentativité des nouvelles trajectoires. Ainsi, même si le pipeline que j'ai mis en place permet d'automatiser le processus de génération et d'analyse de nouvelles conformations, il reste encore des marges d'amélioration pour assurer une couverture exhaustive de l'espace des structures accessibles au système étudié.

# 5 Conclusion

Ce protocole, testé ici sur un système modèle, pourrait à terme être appliqué à des biomolécules beaucoup plus complexes. Une perspective particulièrement prometteuse serait de découper un système de grande taille, tel que le ribosome, en sous-unités plus petites, de générer des ensembles conformationnels locaux pour chaque fragment à l'aide de modèles de diffusion, puis de les recombiner via des méthodes globales telles que les chaînes de Markov de Monte-Carlo (MCMC). Une telle approche modulaire permettrait d'explorer de manière réaliste et scalable le paysage conformationnel de systèmes biologiques géants, ouvrant ainsi la voie à une modélisation structurale de nouvelle génération.

# Références


[1] Stefan Doerr, Maciej Majewski, Adrià Pérez, Andreas Krämer, Cecilia Clementi, Frank Noé, Toni Giorgino, and Gianni De Fabritiis. TorchMD : A Deep Learning Framework for Molecular Simulations. *Journal of Chemical Theory and Computation*, 17(4) :2355–2363, 2021. https://doi.org/10.1021/acs.jctc.0c01343

[2] Joseph Anderson, et al. Cascades of information with message-passing. arXiv preprint arXiv :2012.06333, 2020. https://arxiv.org/abs/2012.06333

[3] Grégoire Sergeant-Perthuis, et al. Abstract—Message passing and structure assembly in ribosomal modeling. https://hal.sorbonne-universite.fr/hal-04527780/file/abstract_ACT_submission.pdf

[4] Mojtaba Shafiei, et al. Probabilistic Message Passing for Molecular Systems. arXiv preprint arXiv :2201.11876, 2022. https://arxiv.org/pdf/2201.11876

[5] David J. Earl and Michael W. Deem. On the Infinite Swapping Limit for Parallel Tempering. arXiv preprint arXiv :1110.4984, 2011. https://arxiv.org/abs/1110.4984

[6] Lauren Wickstrom, Alan Okur, and Carlos Simmerling. Multidimensional Replica Exchange Molecular Dynamics Yields a Converged Ensemble of an RNA Tetranucleotide. *Journal of Chemical Theory and Computation*, 5(5) :1347–1352, 2009. https://pmc.ncbi.nlm.nih.gov/articles/PMC3893832/

[7] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based Generative Modeling Through Stochastic Differential Equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. arXiv preprint arXiv :2011.13456. https://arxiv.org/abs/2011.13456

[8] Sara Romeo-Atance, Daniel Viguera-Diez, Pablo Vázquez-Rodríguez, Maciej Majewski, Cecilia Clementi, and Gianni De Fabritiis. Generation of conformational ensembles of small molecules via surrogate model-assisted molecular dynamics. ChemRxiv preprint, 2023. https://doi.org/10.26434/chemrxiv-2023-jz6fc