sanofi

Drug Design with diffusion models

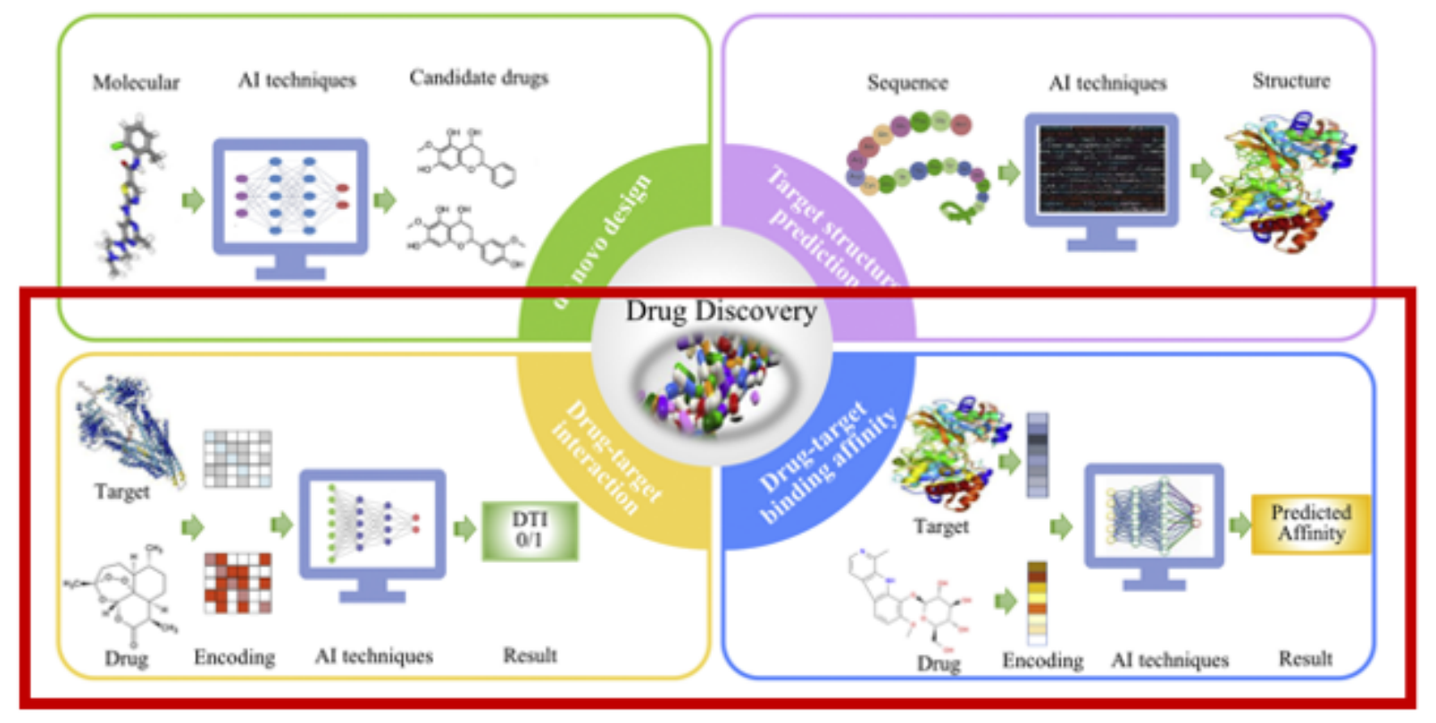
Gaspard Beaudouin, Maxime Muhlethaler, Titouan Pottier

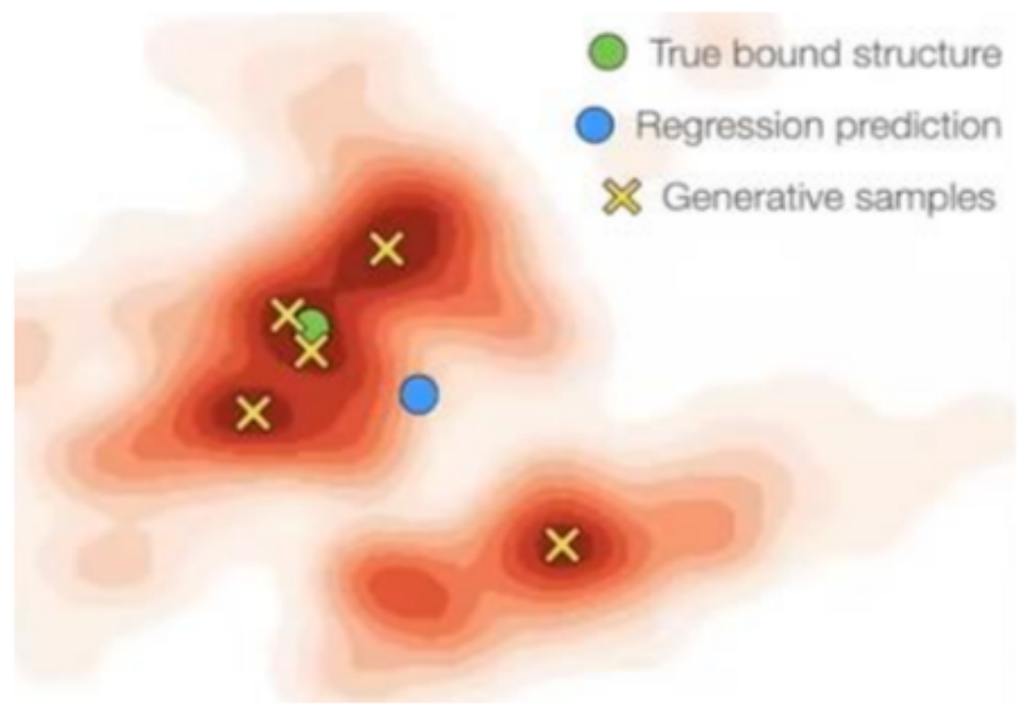
Ecole des Ponts ParisTech x Sanofi

École des Ponts ParisTech

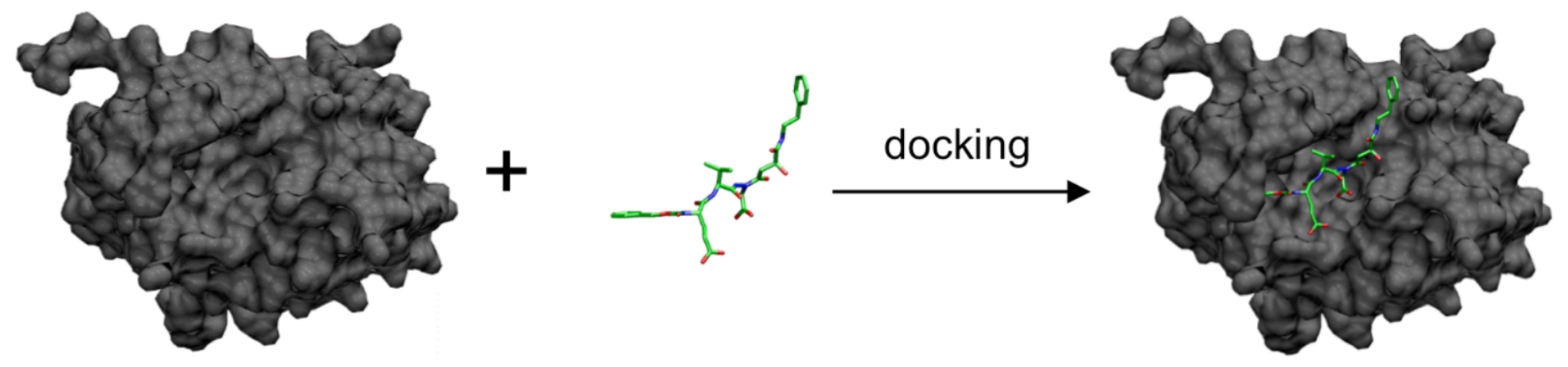
Abstract

- Artificial intelligence, through **generative models**, has emerged as a crucial tool in the realm of **drug discovery**.
- Diffusion models, which utilize **stochastic diffusion equations**, have proven successful in generating new data points to expand the molecular landscape for drug discovery.





- The **DiffDock software** represents a significant innovation in **molecular docking**, a computational technique vital for predicting the binding interactions between small molecules (ligands) and target proteins and the ligand's pose including its position and orientation.
- Molecular docking is essential for determining the potential affinity and stability of the interaction between a ligand and its protein, facilitating the identification of compounds that can effectively bind to **protein sites linked to diseases**. It is crucial for the drug discovery process.

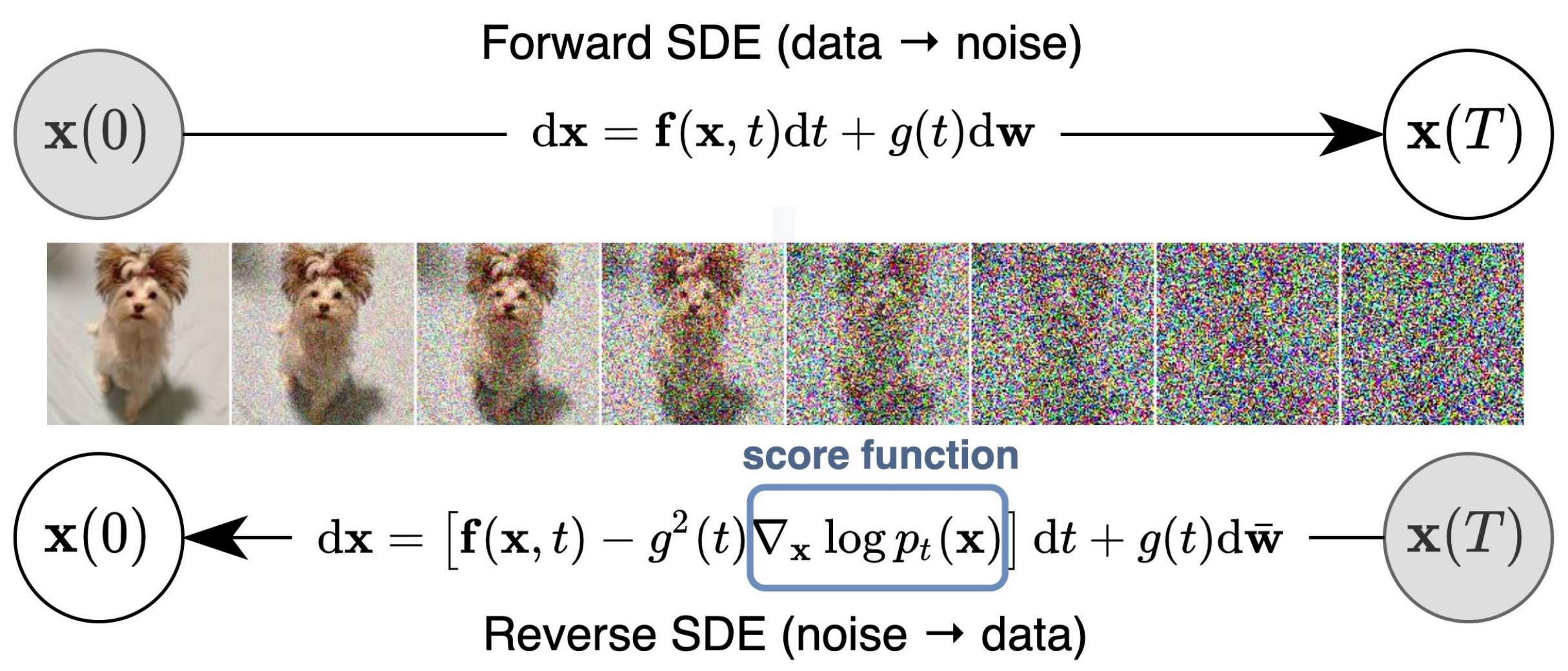


Focus on diffusion models with score-based models

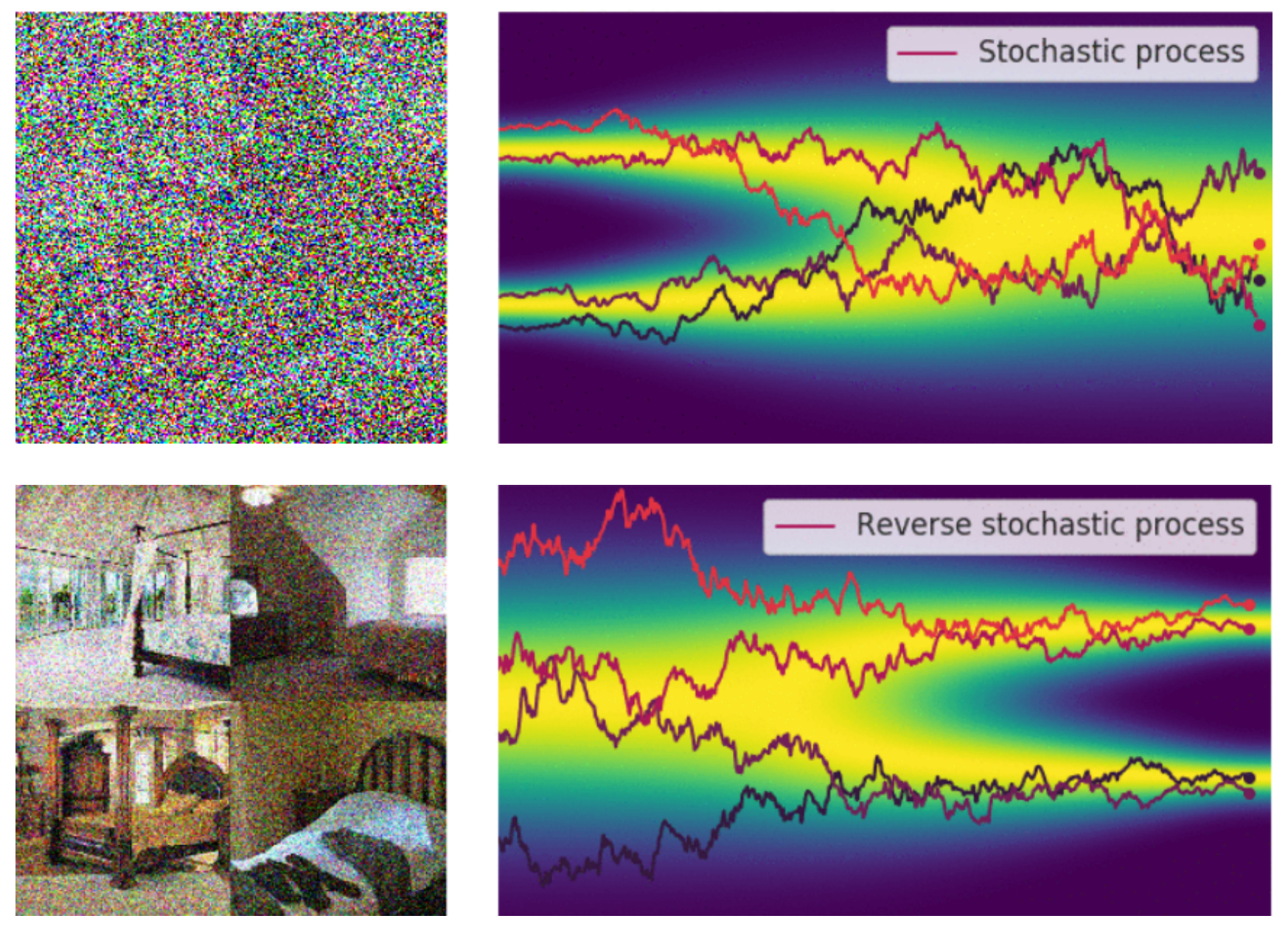
- **Diffusion models** work by gradually transforming an initial data distribution (noise) into a target data distribution, through an iterative process. This process is guided by models that **learn to reverse these diffusions**, allowing the **generation of samples from noise** by following the **inverse steps of diffusion.**
- **Score models** learn score functions (gradients of logarithmic probability densities) over many data distributions perturbed by noise through **Stochastic Differential Equation (SDE)**, and then generate samples via the reverse SDE.
- SDEs and their reverse counterparts (reverse SDEs) commonly take the following forms :

$$dX_t = f(X_t, t)\,dt + g(t)\,\mathbf{B}_t$$

$$dX_t = \left[f(X_t, t) - g^2(t)\nabla_x \log p(X_t)\right]\,dt + g(t)\,\mathbf{B}_t$$


Forward SDE (data → noise)
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$
$\mathbf{x}(T)$
score function
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}]\,\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$
$\mathbf{x}(T)$
Reverse SDE (noise → data)

Process of Diffusion and Reverse Diffusion in Score-based Generative Models



Stochastic Trajectory of Diffusion Models :
from Random Noise to Structured Representation with Score Functions

We restrict ourselves to a simple two-dimensional setting and want to generate a spiral distribution from a gaussian distribution

- Forward Noising Process with the following SDE :

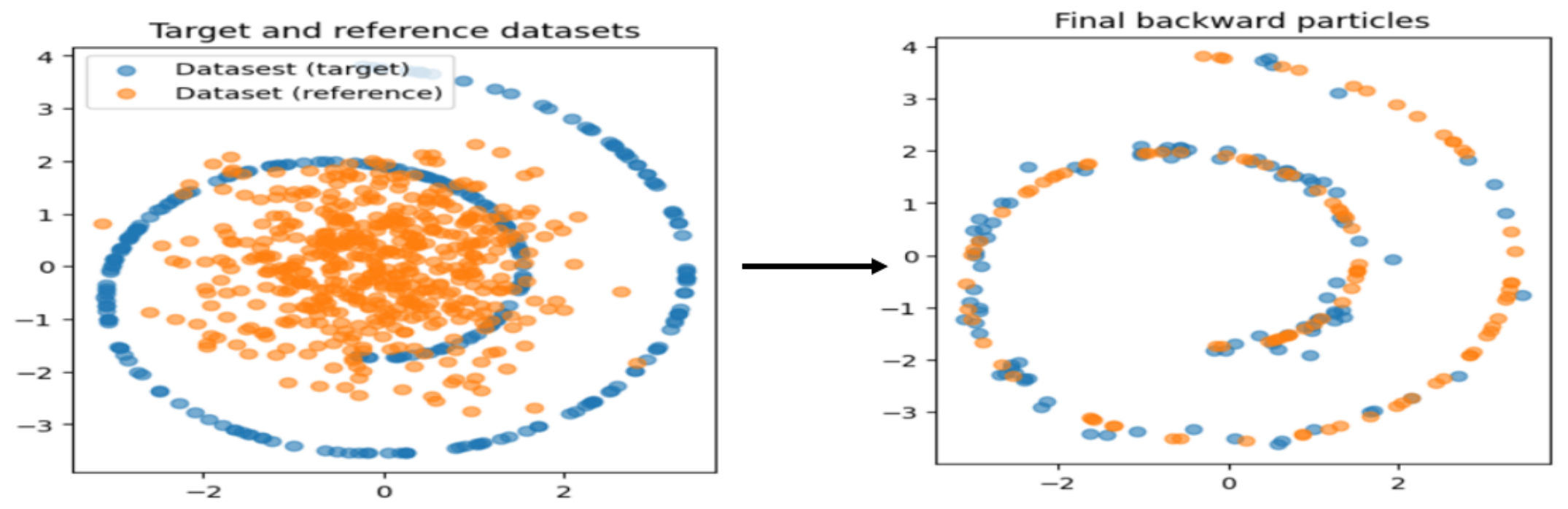
$$\mathrm{d}\mathbf{X}_t = -\frac{1}{2}\mathbf{X}_t\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \quad \mathbf{X}_0 \sim \pi,$$

.

- Backward Denoising Process using the reverse SDE :

$$\mathrm{d}\mathbf{X}_t = \left(\frac{1}{2}\mathbf{X}_t + \nabla \log p_{T-t}(\mathbf{X}_t)\right)\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \quad \mathbf{X}_0 \sim \mathcal{N},$$

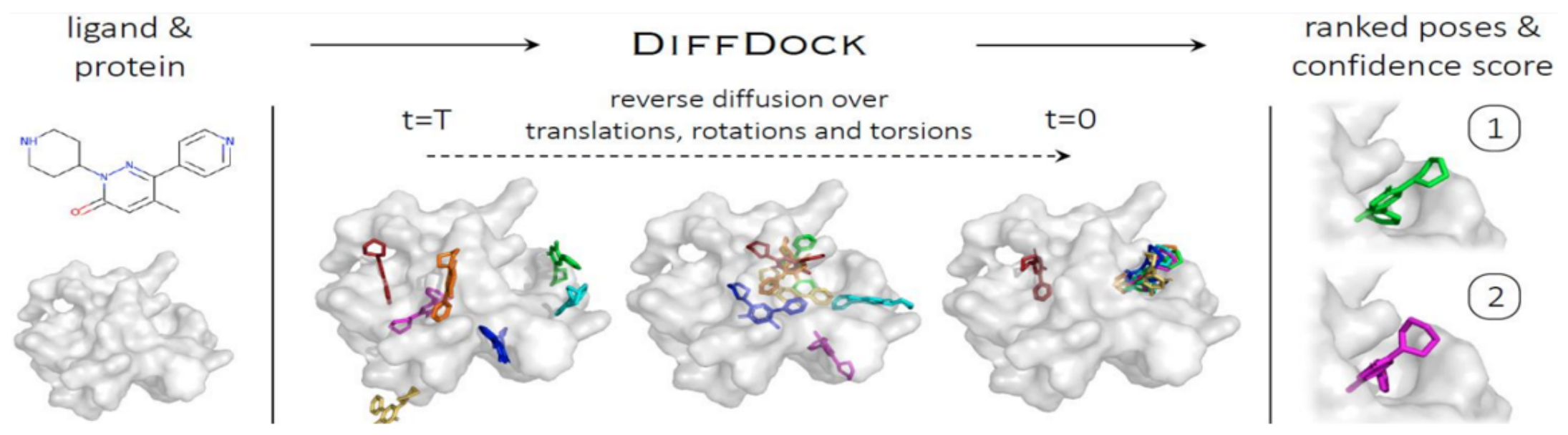
.



Reverse Diffusion Process

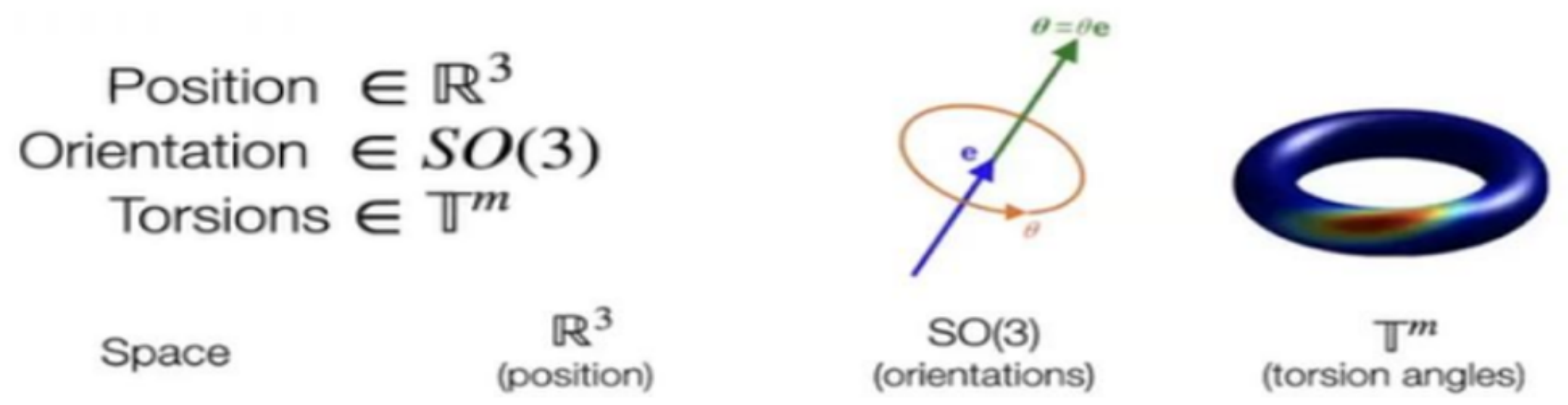
Diffdock and results

- DiffDock uses diffusion models, to simulate the binding process between ligands and proteins. By temporally reversing stochastic diffusion equations, DiffDock generates **predictions about the optimal position and orientation** of ligand molecules relative to the target protein.


ligand & protein
DIFFDOCK
ranked poses & confidence score
t=T
reverse diffusion over translations, rotations and torsions
t=0
1
2

We use the following SDE to generate independently position, orientation and torsion of the ligand on the protein :

$$d\mathbf{X}_t = \sqrt{\frac{d\sigma^2(t)}{dt}}d\mathbf{B}_t \quad \text{where} \quad \sigma^2 = \sigma^2_{\text{tr}}, \sigma^2_{\text{rot}}, \text{ or } \sigma^2_{\text{tor}} \quad \text{for } T(3), SO(3), \text{ and } SO(2)^m$$

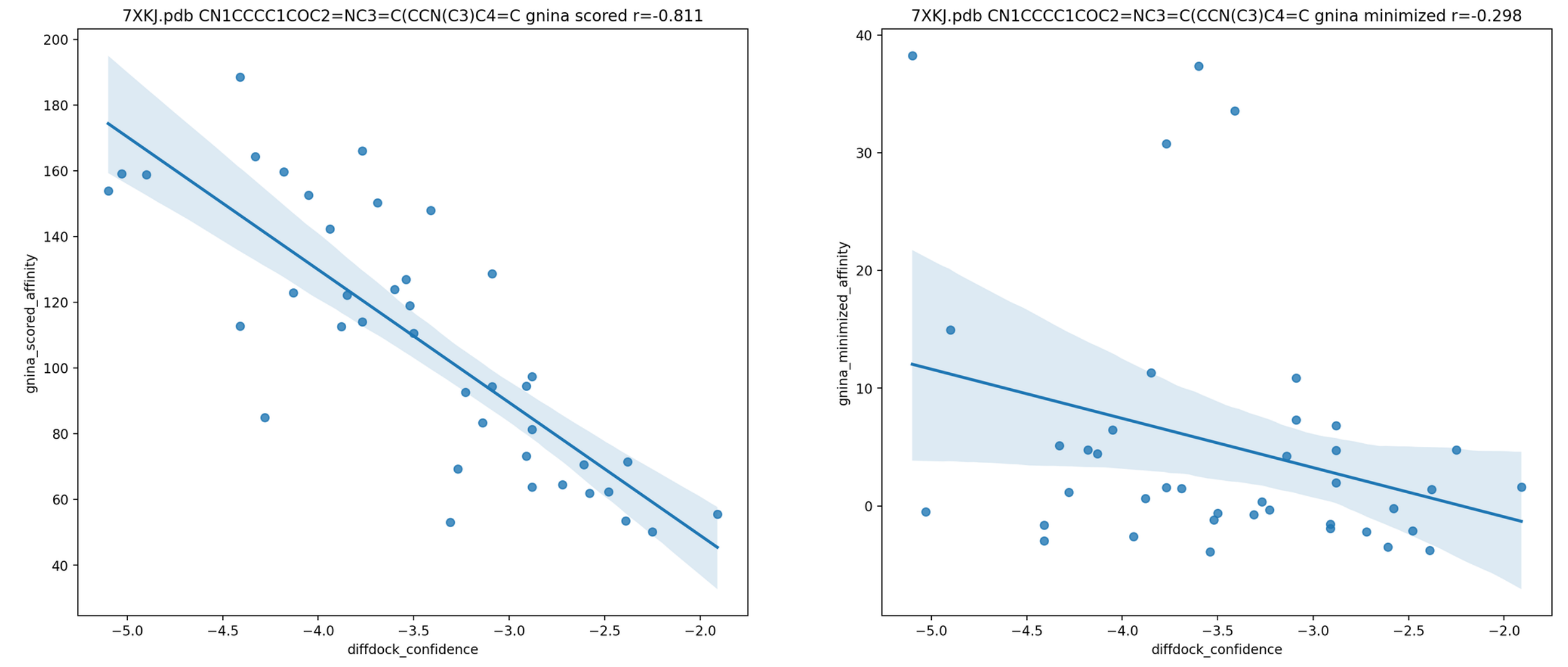

Position $\in \mathbb{R}^3$
Orientation $\in SO(3)$
Torsions $\in \mathbb{T}^m$
$\theta = \theta e$
e
θ
Space
$\mathbb{R}^3$ (position)
SO(3) (orientations)
$\mathbb{T}^m$ (torsion angles)

For our DiffDock test, we will focus on KRAS and Adagrasib molecules

- **KRAS** is a gene encoding the K-Ras protein, crucial in regulating cell growth and differentiation. Mutations in KRAS are prevalent in cancer.
- **Adagrasib** is an small molecule inhibitor targeting the mutated form of K-Ras. Despite challenges due to the absence of a clear binding pocket, Adagrasib has shown promising anti-tumor activity in clinical studies.

Diffdock confidence score

- DiffDock is a two-step process : we train the diffusion model and a **confidence score** model based on the root mean square distance (RMSD) between the structures.

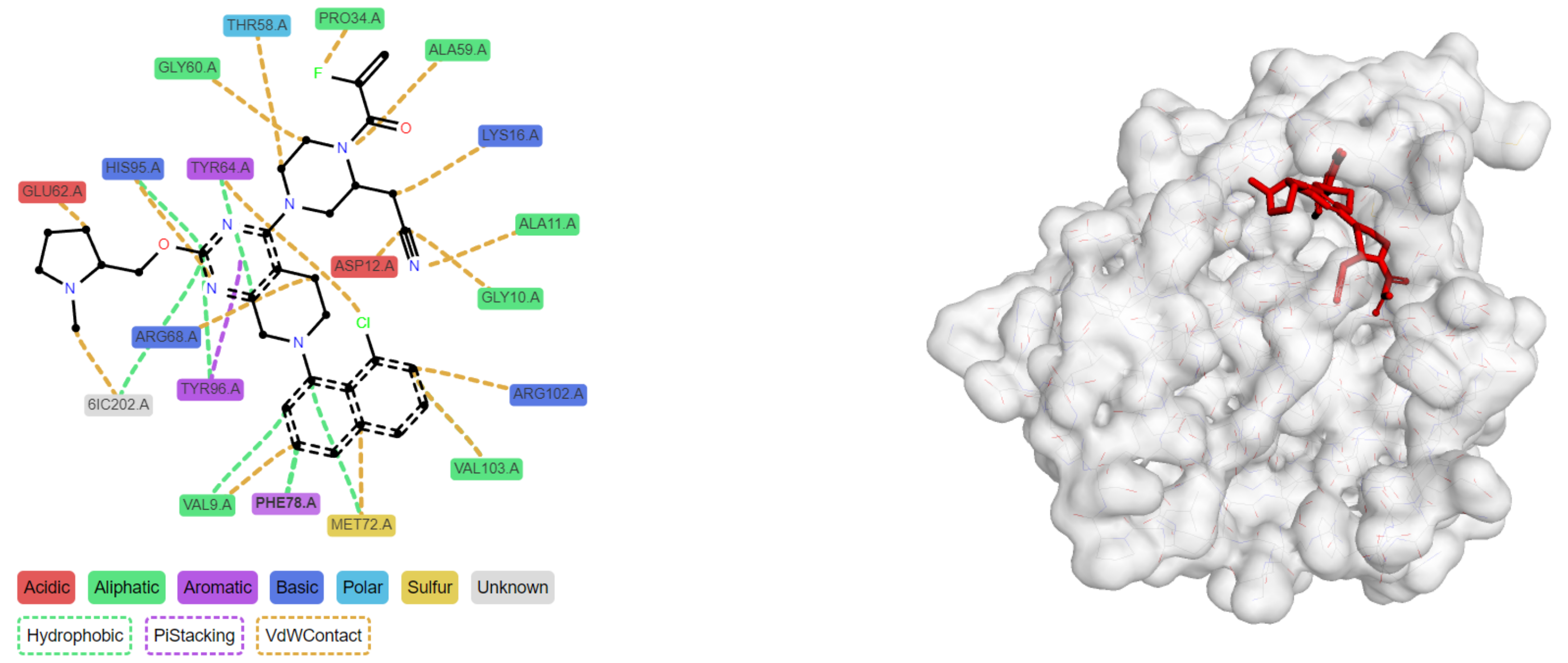


Comparison of DiffDock Confidence with Gnina Predicted Affinity

- We compare our confidence scores with others tools like GNINA affinity scores.
- Then we **optimize the ligand's pose** by locally **minimizing the potential energy of the protein-ligand binding** using the GNINA software.
- GNINA employs evolutionary algorithms to explore various ligand conformations and generate a population of ligand poses, evaluate them based on criteria such as energy and interactions. Then, it optimizes the ligand-receptor system energy using numerical optimization methods like gradient descent to find the most stable and favorable pose.

Our DiffDock results

- The different types of **chemical interactions** could be visualized.
- Such information is crucial for predicting the ligand's **affinity** and **specificity**, as well as for guiding the design of chemical modifications aimed at **enhancing interaction with the target protein**.



Best confidence score result achieved with DiffDock using KRAS and Mirati

References

[1] G. Corso.
Diffdock: Diffusion steps, twists, and turns for molecular docking.
`https://github.com/gcorso/DiffDock`.

[2] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola.
Diffdock: Diffusion steps, twists, and turns for molecular docking.
arXiv, 2017.

[3] Yang Song.
Score-based generative modeling through stochastic differential equations.
`https://yang-song.net/blog/2021/score/`, May 2021.
Published on the blog of Yang Song.

[4] Lilian Weng.
What are diffusion models?
`https://lilianweng.github.io/posts/2021-07-11-diffusion-models/`, July 2021.

Supervisors : Paraskevi Gkeka, Tony Lelievre, Régis Santet, Gabriel Stolz