INTRINSICALLY MOTIVATED GOAL EXPLORATION PROCESSES (IMGEP) FOR THE DISCOVERY OF DIVERSE PROTEIN-LIGAND DOCKING PROFILES

Zacharie BUGAUD
INRIA, Flowers Team, France
zacharie.bugaud@inria.fr

ABSTRACT

In the context of drug discovery on an identified protein target, protein-ligand docking is a commonly used in-silico approach to predict molecular interactions between ligands and proteins. However, the focus is often on finding the most stable ligands, sometimes neglecting the potential exploration of other interaction profiles that may initially seem less favorable in terms of energy, but could be interesting from a therapeutic standpoint. In this study, we propose using Intrinsically Motivated Goal Exploration Processes (IMGEP) to systematically explore chemical space in order to identify new binding profiles on a given protein pocket. By iteratively moving through chemical space, we were able to generate new ligands yielding original docking profiles, capturing a broader range of possible interactions and conformations. The results show that the IMGEP approach allows for the identification of novel ligand configurations, providing new perspectives for drug discovery.

1 Introduction

Protein-ligand docking is a key area in molecular modeling, primarily used in the search for new drugs. It aims to predict the preferred orientation of a ligand (small molecule) when it interacts with a protein target (typically an enzyme or receptor). This prediction is based on molecular interactions, such as hydrogen bonds, hydrophobic interactions, and other electrostatic interactions that form between the ligand and the protein.

Docking simulations aim to find the most stable configuration of a ligand within a protein's binding pocket. However, these approaches are generally employed to search for ligands that come closest to a global minimum binding energy. By focusing solely on finding ligands with the highest affinities, these methods may overlook conformations that could still be relevant and non-trivial from a therapeutic standpoint. The objective of this study is to systematically and diversely explore affinity profiles with a given protein pocket, identifying original docking profiles that may be overlooked by traditional approaches. To this end, we propose using an approach called *Intrinsically Motivated Goal Exploration Processes* (IMGEP) [1]. Initially developed in the field of robotics, this method can also be employed for exploration of chemical space in the context of protein-ligand docking.

By applying this method to protein-ligand docking, we propose to introduce controlled chemical mutations into the ligand space and explore a broadened set of possible interaction modes within a given protein pocket. The goal is to generate new docking profiles that may reveal unexpected molecular interactions, thus providing new perspectives for drug discovery.

2 IMGEP

2.1 In General Cases

Intrinsically Motivated Goal Exploration Processes (IMGEP) algorithms are designed to autonomously explore complex environments. They rely on the idea of exploration guided by self-defined goals, rather than pre-defined objectives or

explicit rewards. To understand their functioning, it is important to distinguish between two key concepts: parameter space and behavior space.

The parameter space encompasses all the actions or configurations that the algorithm can try. It represents all possible combinations of parameters the algorithm can modify to generate different outcomes. In contrast, the behavior space represents the observable results of the actions taken by the algorithm. It can be any measurement or characteristic of the outcome that describes what has been produced. For example, if one aims to explore different behaviors of a system, the behavior space could represent measured characteristics such as shape, speed, or trajectory of a process.

The IMGEP algorithm relies on the idea of continuous interaction between these two spaces. At each iteration, the algorithm selects a goal in the behavior space, corresponding to a type of behavior it wishes to explore. It then selects a configuration in the parameter space to attempt to produce this behavior. The algorithm then evaluates the result of this action and updates its knowledge of the behavior space based on the observed results.

This approach allows the algorithm to discover interesting areas of the behavior space on its own, maximizing the diversity of observed behaviors. It promotes curiosity-driven exploration, where the algorithm chooses to explore new or unexpected behaviors rather than simply following a single objective.

2.2 In the Context of Protein-Ligand Docking

Here, we adapt the IMGEP framework for the exploration of chemical space in the context of protein-ligand docking. The goal is to systematically and diversely explore the possible interaction profiles between a ligand and a fixed protein pocket throughout the exploration. The parameter space corresponds to the chemical space, the space of possible ligands that can be generated and docked in the protein pocket. The behavior space represents the observed docking profiles, that is, the types of interactions formed between the ligand and the protein pocket. We aim to explore a diverse set of docking profiles by autonomously navigating the chemical space to identify original ligand-protein configurations within the protein pocket.

3 Methodology

3.1 Specification of the Protein and the Pocket

The protein structures used in this study were selected from the Protein Data Bank (PDB) [2], which is a public resource containing three-dimensional structures of proteins obtained through X-ray crystallography or nuclear magnetic resonance (NMR).

To identify potential binding pockets on the protein surface, the CavitySpace database [3] was used. This database contains identified potential pockets for many protein structures. A bounding box is then defined around the binding pocket, constituting the physical search space for the docking simulations.

3.2 Docking

Protein-ligand docking was performed using AutoDock Vina [4] [5], a widely used molecular docking tool for predicting interactions between ligands and proteins. It allows for the prediction of the most stable conformations of a ligand within a protein pocket, in a relatively short amount of time.

3.3 Definition of the Goal Space

Hydrogen bonds play a crucial role in stabilizing protein-ligand complexes. Thus, for an identified binding pocket, atoms capable of forming hydrogen bonds are listed. These atoms, totaling N, are denoted as A_i, $i \in [1, N]$.

For each ligand pose obtained after the docking process, the hydrogen bonds formed with the atoms A_i are encoded in a vector $\mathbf{D} = [d_h(A_1), d_h(A_2), ..., d_h(A_N)]$,

where $d_h(A_i)$ represents the distance in Ångström between the atom A_i and the hydrogen bond donor or acceptor located on the ligand, or if no hydrogen bond is formed, the maximum accepted distance MAX_D for a hydrogen bond.

We then use the goal space $[0, 1]^N$, with the embedding $E = \frac{d_{MAX}-D}{d_{MAX}-d_{MIN}}$ where d_{MAX} and d_{MIN} are the maximum and minimum accepted distances for a hydrogen bond. We consider $d_{MIN} = 1.5$ Å and $d_{MAX} = 3.5$ Å. This encoding allows us to quantify the quality and extent of the hydrogen bonds formed, constituting a basic yet informative signature of the docking configuration. The representation is then a finite vector of the same size throughout the exploration, independent of the complexity or pose of the ligand.

3.4 Exploration of Chemical Space

The exploration of the potential ligand space was carried out by generating molecular derivatives from initial ligands. To do this, we used the tool *CReM* (Chemically Reasonable Mutations), an open-source framework for generating chemical structures based on fragments. The principle of CReM relies on the idea of "matched molecular pairs," where molecular fragments in identical contexts can be replaced to produce new chemically valid and likely synthesizable structures. Successive random mutations enable iterative movement in chemical space.

3.5 Implementation of IMGEP

IMGEP is initialized here with a predefined ligand, additionally initializing the history of explored ligands and their associated hydrogen binding profiles. The goals are sampled uniformly in $[0, 1]^N$. Once a goal is generated, the closest ligand from the history of explored ligands is searched. A nearest neighbor search method is used to identify the ligand closest to the goal. We then use CReM to generate a new ligand from this closest ligand.

We randomly select the mutated ligand from a set consisting of at most 10 mutated ligands by substitution via CReM and 10 mutated ligands by augmentation via CReM.

The resulting ligand is then docked in the protein pocket, and its hydrogen binding profile is calculated. We update the history of explored ligands and iterate the process by generating a new goal. The algorithm stops after a predetermined number of iterations. The ADTool framework [7] has been used to implement the IMGEP algorithm, providing an interactive visualization of the results and an API to define the algorithm components specific to each application.

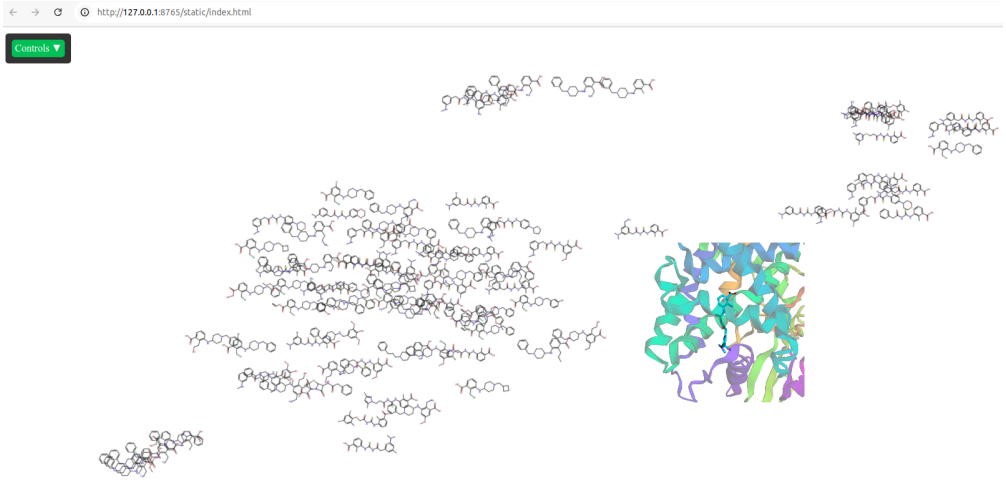

Figure 1: Interactive visualization of the results of chemical space exploration with ADTool.

4 Experiments

4.1 Experimental Protocol

The objective is to evaluate whether the IMGEP approach allows for better coverage of the goal space compared to random navigation in the parameter space. The baseline is constructed by the following procedure:

- Initialization of the history of explored ligands with the initial ligand (the same as for IMGEP)
- Random selection of a ligand from the history
- Random mutation of the ligand and addition to the history. The mutation operator is the same as for IMGEP.
- Calculation of the affinity profile and repetition of the process

To measure the coverage of the goal space ($[0, 1]^N$) obtained after each exploration method, we subdivide the space into M^N equal bins, by discretizing each dimension into M equal parts. . We then measure after each exploration method the number of bins containing at least one hydrogen binding profile. We also measure the average distance between the obtained hydrogen binding profiles for each exploration method.

The SMILES of the initial ligand is arbitrarily CC1(C(N2C(S1)C(C2=O)NC(=O)CC3=CC=CC=C3)C(=O)O)C (Penicillin G), and the hrefPDF of the protein structure is P22680_3V8D_A. The pocket used is in CavitySpace with ID 2.

4.2 Results

The use of IMGEP has allowed for the generation of a significantly greater diversity of hydrogen binding profiles compared to random exploration.

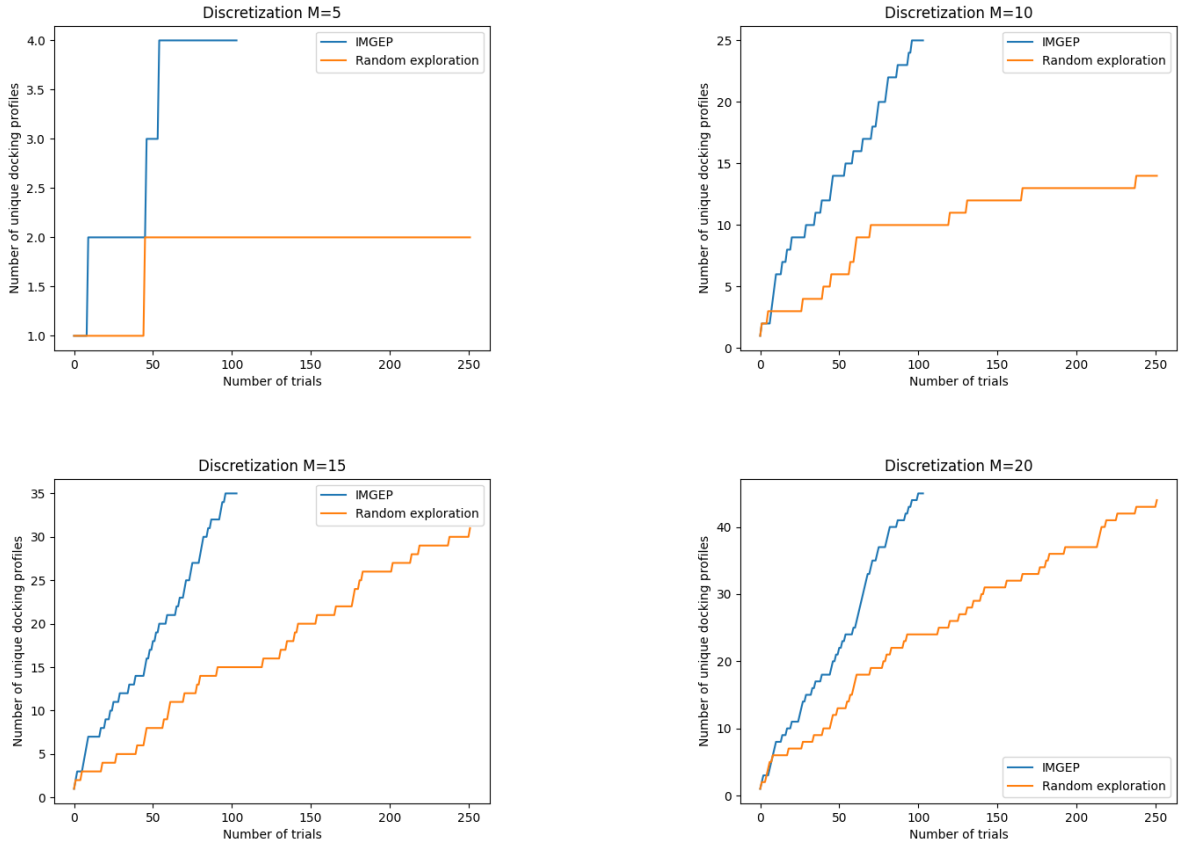

Figure 2: Number of unique bins containing at least one hydrogen binding profile for each exploration method, depending on the discretization for bins.

We can also look at the distribution of distances between the obtained hydrogen binding profiles and the generated goals.

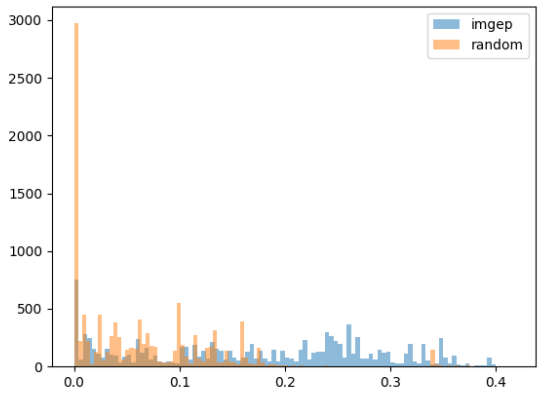

Figure 3: Distribution of Euclidean distances between the hydrogen binding profiles obtained for each exploration method.

	Average distance	Standard deviation	Maximum distance
random	0.062135	0.067486	0.375832
imgep	0.172014	0.110918	0.416773

5 Discussion

5.1 Comparison of Exploration Methods

The results show that the IMGEP approach allows for a more diverse exploration of the hydrogen binding profile space, identifying a greater number of unique bins containing at least one hydrogen binding profile, regardless of the size of the subdivision used. Approximately twice as many unique bins are identified with IMGEP compared to random exploration with the same number of iterations. The finer the discretization, the more the difference narrows between the two methods, as each slight variation in the hydrogen binding profile is more likely to fall into a unique bin. The coarser the discretization, the more significant the differences between the different discoveries.

The distribution of distances between the obtained hydrogen binding profiles is also more extensive for the IMGEP approach, suggesting a greater diversity of the profiles obtained.

By reducing the number of simulations needed to explore chemical space, IMGEP allows for the more rapid identification of promising ligand configurations, thus providing a time and resource advantage where docking simulations are costly.

5.2 Limitations and Future Perspectives

Despite its advantages, the proposed approach also presents certain limitations. For example, the current methodology mainly focuses on hydrogen bonds, although other important interactions may exist in protein-ligand complexes. Future extensions of this work will include incorporating other types of interactions, as well as exploring flexible protein pockets to better capture molecular dynamics. It is also possible to initialize IMGEP with a diverse set of molecular scaffolds to explore a wider and more varied chemical space. Finally, the limitations of the CReM approach [6] could be circumvented by using more sophisticated ligand generation methods, such as generative adversarial networks (GANs) or recurrent neural networks (RNNs). It is also important to note that a single ligand may have multiple conformations within the same protein pocket, and only one is considered in this study. An extension of this research could potentially consider multiple conformations for the same ligand. The calculation of the embedding in the goal space could be improved by also integrating hydrophobic interactions, $\pi - \pi$ interactions, and electrostatic interactions. The sampling of goals might be enhanced by targeting, for example, low-density areas of the goal space. It would also be interesting to verify in more detail the validity of the obtained docking poses, for instance by using validation methods such as PoseBusters [8].

6 Conclusion

This study demonstrates that the application of Intrinsically Motivated Goal Exploration Processes (IMGEP) in the context of protein-ligand docking allows for a more efficient exploration of chemical space, for identifying original docking binding profiles, reducing the number of necessary docking simulations and providing a more diverse set of ligand configurations. The IMGEP approach thus offers a new perspective for drug discovery, identifying novel ligand configurations that may reveal unexpected and potentially therapeutic molecular interactions.

References

[1] Forestier, S., Portelas, R., Mollard, Y., Oudeyer, P.-Y. Intrinsically Motivated Goal Exploration Processes with Automatic Curriculum Learning. arXiv:1708.02190 [cs.AI] (2022). https://arxiv.org/abs/1708.02190

[2] Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-242. doi:10.1093/nar/28.1.235

[3] Wang, S.; Lin, H.; Huang, Z. et al. CavitySpace: A database of potential ligand binding sites in the human proteome. Biomolecules 2022, 12, 967

[4] J. Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. Journal of Chemical Information and Modeling.

[5] O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, Journal of Computational Chemistry 31 (2010) 455–461

[6] Polishchuk, P. CReM: chemically reasonable mutations framework for structure generation. J Cheminform 12, 28 (2020). https://doi.org/10.1186/s13321-020-00431-w

[7] INRIA Flowers Team. 2024. ADTool: Assisted and Automated Discovery for Complex Systems. https://github.com/flowersteam/adtool

[8] Buttenschoen, M., Morris, G. M., Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalize to novel sequences. arXiv:2308.05777 [q-bio.QM] (2023). https://arxiv.org/abs/2308.05777