Comparison of Physics-Based and Machine Learning-Based Molecular Docking Methods

**Abstract**

Molecular docking holds significant potential to transform drug development. However, the comparative effectiveness of machine learning based molecular docking methods versus traditional physics-based methods remains unclear. This study evaluates the accuracy of various molecular docking programs to determine whether machine learning based docking programs are comparable or superior to physics-based programs. The docking programs were tested using ligands and proteins from the Database of Useful Decoys: Enhanced (DUD-E). Specifically, the physics-based docking method Smina was compared to the machine learning based methods DiffDock and RoseTTAFold All-Atom. Accuracy was assessed using three metrics: the root mean squared deviation (RMSD) of the crystal ligand compared to the redocked crystal ligand, the Spearman rank correlation coefficient of the predicted binding affinities/confidence scores with the actual binding energies and the enrichment factor of the models. The results provide insights into the benefits and trade-offs of physics-based and machine learning-based methods. The study also sheds light on scenarios where certain methods may be more favorable, which will better inform the docking process.

**Introduction**

Molecular docking plays a crucial role in the field of drug discovery. It is the computational process of predicting the position and orientation of a ligand (a small molecule that binds to a protein) when bound to a protein.

In drug discovery, potential chemical compounds are tested with proteins to see if the compounds will bind to the protein and cause the desired biological effect in the human body. Experimentally testing each compound with its target protein is a time consuming and costly process, but molecular docking significantly accelerates this process. Beyond speeding up the testing, molecular docking also aids in compound optimization by allowing medicinal chemists to visualize the binding interactions and make informed modifications–such as additions, deletions or substitutions–as needed on the compound to improve its binding affinity with the protein. They can also use docking to understand how to alter the ligand to optimize properties like toxicity. Furthermore, by screening large libraries of chemical compounds, molecular docking allows diverse compounds to be tested. Due to this, novel compounds that can bind to proteins are found. Because the novel compounds have different properties than the existing active compounds, it can potentially lead to the discovery of drugs with lower toxicity, more desirable ADME and fewer side effects. Ultimately, molecular docking has the power to revolutionize the drug development field, reducing both the cost and time required to develop drugs, and thus positively impacting healthcare.

In molecular docking, the ligands can take on a variety of positions when bound to the protein, and certain positions will form a more favorable interaction between the ligand and the protein. Binding poses that form a more favorable interaction with a protein have a lower binding energy. After the program predicts poses, it is still important to experimentally verify that the top predictions are correct because docking algorithms may not always be accurate.

In the early stages of drug discovery, the interactions between ligands and proteins were primarily tested experimentally. This approach was inherently time consuming as it required testing a large amount of compounds against the target protein1. As technology advanced, computational methods were introduced to streamline the process, enabling the screening of large libraries of compounds to predict which compounds would bind favorably to the protein2. These early computational methods were primarily physics-based. Physics-based docking programs create a simulation of the ligand’s interaction with the protein. Possible binding poses of the ligand are sampled and physics based principles like Van der Waals interactions, hydrogen bonding and electrostatic forces are used to dock the ligands and predict their binding affinity with the protein2. Physics based molecular docking programs directly predict the binding energy between the ligand and protein, and represent this as a docking score. The docking scores are used to rank the predicted poses based on their binding affinity2. Notable physics-based models include Autodock Vina and Smina. These models sample possible binding poses and use scoring functions to predict docking scores4.

More recently, machine learning-based molecular docking programs have emerged as powerful alternatives to traditional physics-based methods. Unlike physics-based approaches which rely on physical properties and principles, machine learning-based methods leverage machine learning models in molecular docking. These models have multiple parameters and use preexisting data to train themselves on how the ligands would interact with the proteins2. Some machine learning methods produce a confidence score for each of their predicted poses and rank the poses based on the confidence score5-6. Prominent machine learning models include Equibind, TankBind, DiffDock, DiffDock-L and RoseTTAFold All-Atom7. A significant benefit of machine learning methods is their ability to continuously improve as new experimental data becomes available, allowing the models to constantly learn and train themselves, adapting and refining their predictions over time. This continuous learning ability contrasts with physics based models which are typically static and cannot evolve over time. Furthermore, the machine learning models can address complexities in the docking process that physics-based models cannot. For example, machine learning-based models can take into account the flexibility of the protein when predicting poses2, a critical factor that can significantly influence binding accuracy and which is challenging for physics based methods.

This data analysis study aims to compare the accuracy of physics-based and machine learning-based models in predicting the binding affinities and poses in molecular docking. To achieve this, ligands were docked to proteins using various models. The models used were Smina, DiffDock and RoseTTA Fold-All-Atom.

Smina is a physics-based molecular docking program that is an enhanced version of Autodock Vina and improves on Autodock Vina’s performance. Smina takes as input a protein, the ligands and a predefined binding box. The binding box is the place on the protein where the ligands will bind. To create this binding box, the autobox function was used, and the autobox was created using the same ligand docked to the protein. Additionally, a buffer of 6 was added to all six sides of the autobox and exhaustiveness was set to 16. Exhaustiveness represents how rigorously Smina will search for correct poses. Using the input, Smina creates possible poses for the ligands bound to the protein, and ranks the poses based on a docking score. The docking score represents the binding affinity of the ligand and the protein, and the lower the docking score, the better the pose4.

Diff Dock is a machine learning based algorithm. More specifically, it is a generative diffusion model. Instead of being given a certain binding pocket for the ligand to be docked on, Diff Dock predicts the binding pocket itself. Given an input of a ligand and a protein, Diff Dock runs reverse diffusion to generate possible poses for the ligand. Diff Dock then ranks the poses based on a confidence score. The confidence score represents the likelihood that a pose will fall below a certain RMSD value, and the higher the confidence score the better the pose7.

RoseTTAFold All-Atom is a machine learning algorithm and it predicts an entire assembly of proteins and non-proteins using a neural network. Its input includes elements of non-polymer atoms, chirality, and chemical bonds. The input represents the information about different components of the assembly. The input is passed through the neural network and used to predict an output assembly. RoseTTAFold All-Atom also gives a confidence score for each predicted assembly5.

These models were chosen because they are proven to be accurate and efficient from past research. For example, Smina improves on Autodock Vina’s performance as it allows for multi-ligand files, multiple file types, autobox creation, more generated docking poses, improved minimization algorithms and user customized scoring functions4. Diff Dock shows promising results and accuracy compared to past machine learning based models like Equibind and Tankbind7. RoseTTAFold All-Atom was chosen because it outperforms baseline pipelines, has high generalizability, and has RMSD values <2.5Å for 46% of covalent modifications7.

A Venn diagram comparing the similarities and differences among Smina, DiffDock and RoseTTAFold All-Atom is shown below to illustrate their respective features and capabilities

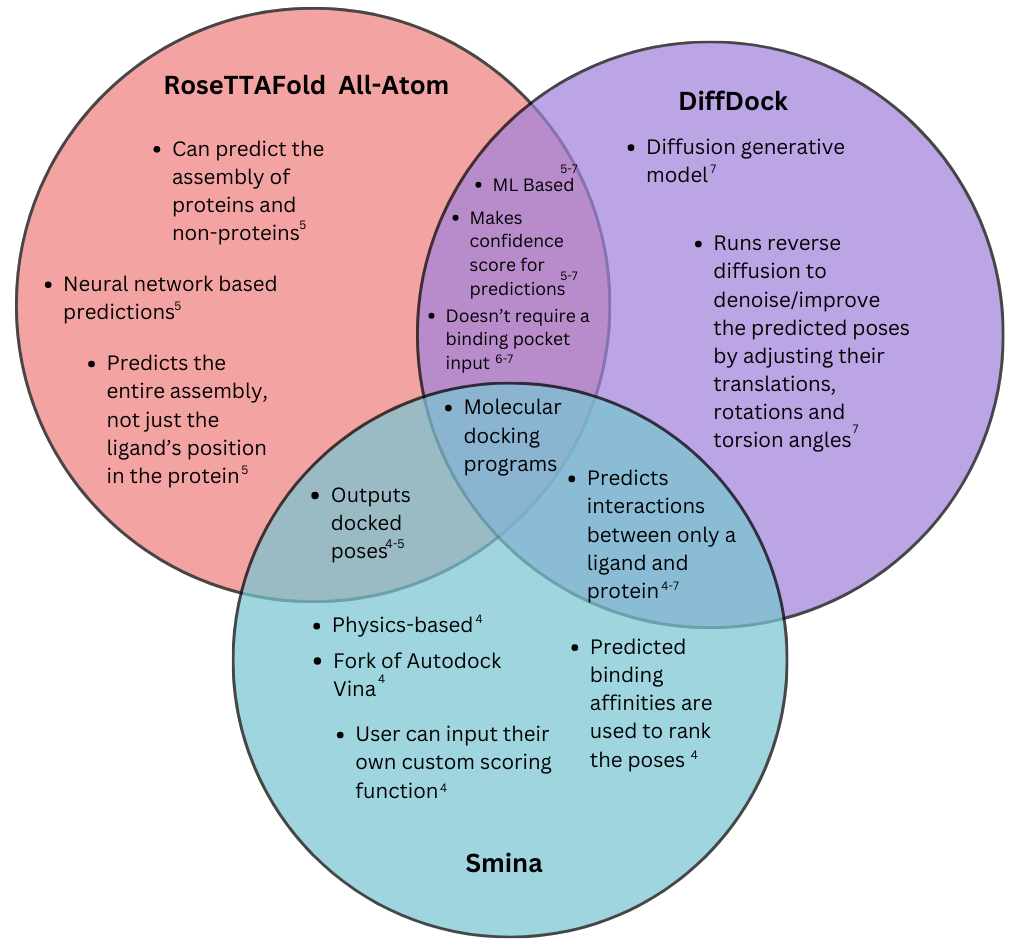


Fig 1. Venn diagram outlining similarities and differences between Smina, DiffDock and RoseTTAFold All-Atom

While both machine learning and physics based approaches have shown promise in molecular docking, it remains unclear which method is more effective under different conditions. Studying these methods and their effectiveness can inform their use and help optimize the drug development process.

To that end, this study aims to evaluate the accuracy and capabilities of these methods, comparing them to determine whether machine learning-based models can match or surpass the performance of physics-based models. By analyzing these methods in depth, this research seeks to clarify their respective strengths and limitations, ultimately providing guidance on their applicability in various docking scenarios.

**Methodology**

These models were run on clean data from the Database of Useful Decoys: Enhanced (DUD-E). From DUD-E, the .pdb files of 25 different proteins were collected. For each protein, 201 ligands were collected, 20 of which were active ligands, 180 of which being decoy ligands and one was the crystal ligand. Active ligands are compounds that bind to the protein, decoys are compounds that have similar properties to active compounds but they themselves do not actually bind to the proteins, and the crystal ligand is a compound that demonstrates the true pose for the bound ligand on the protein8. To prepare the data for Smina, it was necessary to use OpenBabel to add hydrogens to all the ligands and proteins9.

Smina was first used to dock each ligand to their respective protein. The docking scores for all the active ligands were collected, and the top 20 ligands docked to each protein were collected as well. This process was repeated for DiffDock and RoseTTAFold All-Atom.

To evaluate the performance of the three models, Spearman rank correlation coefficient, RMSD values, and enrichment factor were used. The Spearman rank correlation coefficient measures the correlation between certain values10. In this case, it is measuring the correlation between the predicted affinities and true affinities to understand the model’s ability to predict docking scores that correlate with ligands’ actual binding affinities. Additionally, PyMol was used to calculate the RMSD value between the redocked crystal ligand and its true position on the protein. PyMol is a molecular visualization tool. The RMSD value represents the distance between the true pose of the crystal ligand and the predicted pose of the crystal ligand, and therefore is a good measure of the model’s ability to make accurate predictions of possible poses4. The enrichment factor was also calculated for each protein. The enrichment factor is a value that measures how much the model was able to enrich the library with actives8.

After running Smina on each protein, the RMSD was calculated to compare the redocked crystal ligand to the true crystal ligand. This was repeated for each protein. Then the docking scores of the 20 actives were compared with the true binding affinities of the actives, which were also found from the DUD-E database. The Spearman rank correlation coefficient was used to calculate the correlation between the true and predicted binding affinities. The enrichment factor was calculated for each protein. The distribution of the data was visualized in python. This process was repeated with the output of DiffDock and RoseTTAFold All-Atom.

**Data**

201 ligands were docked to each protein, 20 were actives, 180 were decoys and 1 was a crystal ligand. Both the ligands and proteins were downloaded from the Database of Useful Decoys: Enhanced (DUD-E).

Proteins Used

|  |  |  |
| --- | --- | --- |
| Protein | PDB ID | Description |
| aa2ar | 3eml | Adenosine A2a receptor |
| abl1 | 2hzi | Tyrosine-protein kinase ABL |
| adrb1 | 2vt4 | Beta-1 adrenergic receptor |
| akt1 | 3cqw | Serine/threonine-protein kinase AKT |
| bace1 | 3l5d | Beta-secretase 1 |
| csf1r | 3krj | Macrophage colony stimulating factor receptor |
| dyr | 3nxo | Dihydrofolate reductase |
| esr1 | 1sj0 | Estrogen receptor alpha |
| fabp4 | 2nnq | Fatty acid binding protein adipocyte |
| fak1 | 3bz3 | Focal adhesion kinase 1 |
| fnta | 3e37 | Protein farnesyltransferase/geranylgeranyltransferase type I alpha subunit |
| fpps | 1zw5 | Farnesyl diphosphate synthase |
| gcr | 3bqd | Glucocorticoid receptor |
| glcm | 2v3f | Beta-glucocerebrosidase |
| hivpr | 1xl2 | Human immunodeficiency virus type 1 protease |
| ital | 2ica | Leukocyte adhesion glycoprotein LFA-1 alpha |
| kit | 3g0e | Stem cell growth factor receptor |
| met | 3lq8 | Hepatocyte growth factor receptor |
| parp1 | 3l3m | Poly [ADP-ribose] polymerase-1 |
| ppara | 2p54 | Peroxisome proliferator-activated receptor alpha |
| prgr | 3kba | Progesterone receptor |
| reni | 3g6z | Renin |
| try1 | 2ayw | Trypsin I |
| urok | 1sqt | Urokinase-type plasminogen activator |
| xiap | 3hl5 | Inhibitor of apoptosis protein 3 |

Fig 2. Table detailing the proteins used in the study, their pdb id and their description from DUD-E

**Results**

This study compared the performance of the physics-based method Smina with the machine learning-based approaches DiffDock and RoseTTAFold All-Atom using three metrics: enrichment factor, Spearman rank correlation coefficient, and RMSD.

Note: I’m still working on getting the results for RoseTTAFold All-Atom, and this page will be updated after that.

Smina**:**

* Enrichment Factor: 60% of docked proteins had an enrichment factor between 0 and 2, indicating Smina mostly did not enrich the library with new actives. A low enrichment factor suggests nonoptimal performance of the docking model.
* Spearman rank correlation coefficient: 64% of binding affinity sets had a Spearman rank correlation coefficient between -0.4 and 0.4. This suggests a weaker correlation between predicted docking scores and the actual binding energies since these numbers are not very close to 1. This suggests Smina may not be the best choice for accurate binding affinity predictions between a docked protein and a ligand.
* RMSD: 61% of docked crystal\_ligands had an RMSD below 2 Å. This shows that Smina is capable of accurately sampling poses that align with how the ligand would truly bind to the protein. Note: Only 18 of the 25 crystal ligands were used to calculate RMSDs due to errors.

Recommendation**:** Smina may be more suited for situations where accurate pose generation is crucial, such as structural studies or when the focus is on visualizing how a ligand fits into a protein's active site. However, due to its weaker enrichment and binding affinity prediction capabilities, it is less suitable when prioritizing active compounds for screening or predicting precise binding affinities.

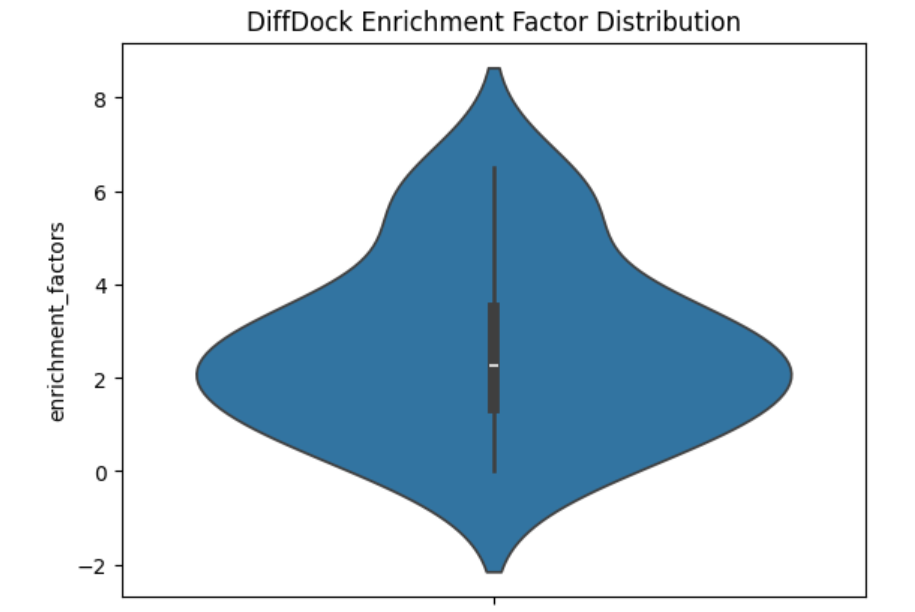
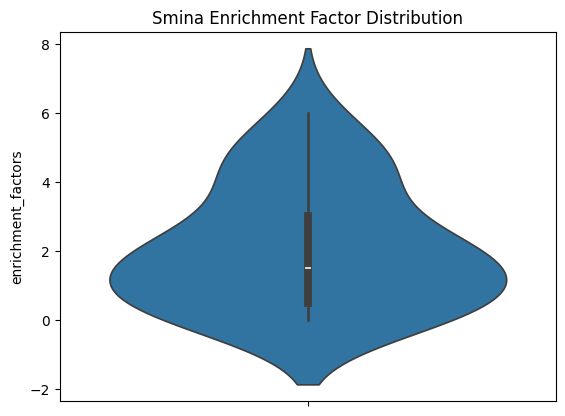


Fig 3. Violin plots showing the enrichment factor calculated for each protein. One graph shows the enrichment factor results of Smina and the other graph shows the enrichment factor results of DiffDock

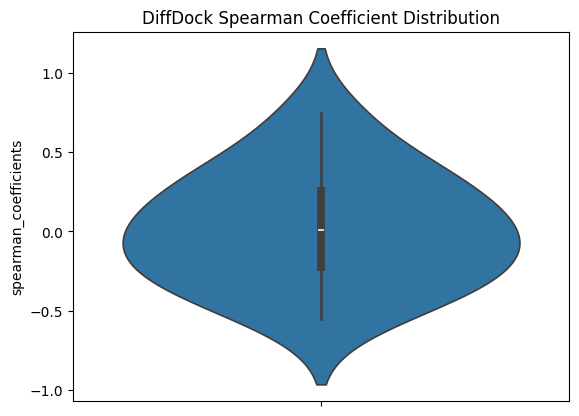
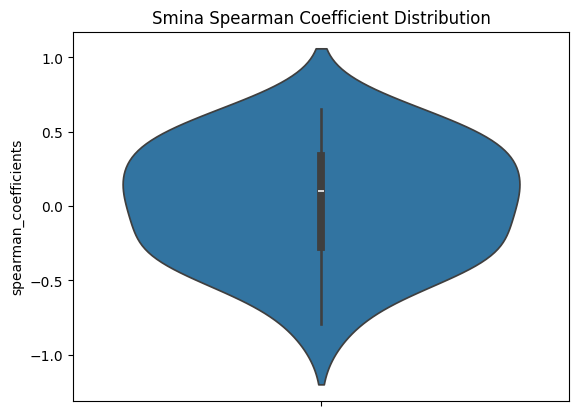
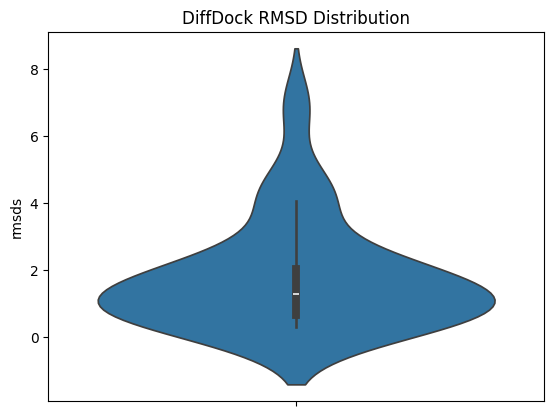
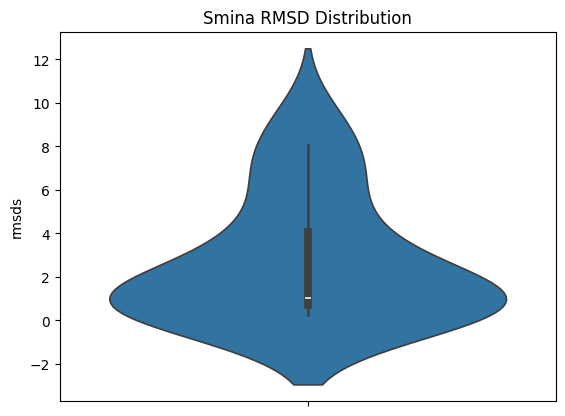


Fig 4. Violin plots showing the Spearman rank correlation coefficient calculated for each protein. One graph shows the Spearman rank correlation coefficient results of Smina and the other graph shows the Spearman rank correlation coefficient results of DiffDock





DiffDock:

* Enrichment Factor**:** 2 is the most common enrichment factor calculated, but it is not the majority, as 20% of the proteins have an enrichment factor of 2. However, 53.3% of the proteins had enrichment factors greater than 2. DiffDock showed higher enrichment, with most proteins achieving an enrichment factor of greater than 2. This suggests that DiffDock can better enrich the library with actives in top ranked poses.
* Spearman rank correlation coefficient: The shape of this graph is roughly symmetrical, and is widest around the center. 46% of the binding affinity sets had Spearman rank correlation coefficients between -0.2 to 0.2. This indicates that DiffDock, like Smina, also is not optimal for producing a confidence score that correlates with the true binding affinity.
* RMSD: 78.2% of the docked crystal ligands had an RMSD < 2 angstroms. Compared to Smina, more docked crystal ligands had an RMSD below 2 Å, and fewer proteins had RMSD values above 2 Å, suggesting that DiffDock is more accurate than Smina in sampling accurate binding poses.

Recommendation: DiffDock should be prioritized in scenarios where identifying active compounds is the primary goal, such as early-stage virtual screening campaigns aimed at selecting candidate molecules for further testing. However, like Smina, its predictions of binding affinities are not highly reliable, so it may not be the best choice for tasks that require precise binding energy estimates.

**Discussion**

Looking at the graphs, the shapes of the distributions of both Smina and DiffDock look similar. They both have lower enrichment factors and lower accuracy of docking scores/confidence scores to true binding affinities. They both had higher accuracy of sampled poses, as shown with the lower RMSD. However, DiffDock had more low RMSD values than Smina and was more heavily centered around a higher enrichment factor. Smina also had more Spearman rank correlation coefficients closer to 1 than DiffDock.

This shows that the accuracy of physics-based methods and ML-based methods are relatively similar with DiffDock being slightly better at enriching the top poses with actives and samling poses, while Smina is slightly more accurate with predicting accurate docking scores.

These findings provide valuable insights into the current capabilities of molecular docking programs. It tells us that both machine learning and physics-based methods struggle and perform well similarly. These findings tell us that physics-based and machine learning-based methods have made good progress in sampling accurate poses. However, the ability to predict the binding affinity of these poses and thereby rank the poses accurately falls short. Additionally, the top scored poses are not usually actives, resulting in a lower enrichment factor. This tells us how similar machine learning-based and physics-based molecular docking methods are and how they can be improved.

For researchers and professionals in drug development, this comparison highlights the strengths and limitations of both physics-based and machine learning-based docking approaches, offering guidance on when each method might be most effective. This could optimize the drug development process and decrease the time and cost necessary for drug development.

Some limitations of this study is that a wide range of molecular docking programs were not studied, so there may not be a full picture of machine learning and physics based methods. This is a potential avenue for future research to go in more depth into the accuracy of different molecular docking methods.

A key recommendation for future research is to expand this comparative analysis by studying a wider range of docking programs. This would offer a more comprehensive understanding of the relative accuracy of physics-based versus machine learning-based methods. Additionally, testing these models under different conditions, such as varying protein structures or ligand complexities, could help identify specific scenarios in which each model performs optimally. This would be especially useful in tailoring molecular docking approaches to the unique challenges of different drug discovery stages.

Molecular docking has the potential to significantly accelerate and optimize drug development, offering cost-effective means of identifying promising compounds. By continuing to investigate and refine these innovative methods, we can drive meaningful advancements in healthcare, with benefits that extend throughout society.

**Acknowledgements**

Thank you to my mentor Ayush Pandit.

**References**

[1] J.P. Hughes, S. Rees, S.B. Kalindjian, K.L. Philpott. Principles of Early Drug Discovery. *PubMed Central.* **162**, 1239-1249 (2011).

[2] ligand.[bing.com/ck/a?!&&p=11b82e632f927040JmltdHM9MTcyNDgwMzIwMCZpZ3VpZD0wZTJhNTBmZi0yMmI5LTY5NzgtMmVlZS00MWZlMjMwYTY4OTYmaW5zaWQ9NTM5NA&ptn=3&ver=2&hsh=3&fclid=0e2a50ff-22b9-6978-2eee-41fe230a6896&psq=what+is+aligand&u=a1aHR0cHM6Ly93d3cuYnJpdGFubmljYS5jb20vc2NpZW5jZS9saWdhbmQ&ntb=1](https://www.bing.com/search?pglt=41&q=what+is+aligand&cvid=3a45f2195b11401b8576abd04f20cd13&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIICAEQ6QcY_FXSAQgxNzUwajBqMagCALACAA&FORM=ANNAB1&DAF0=1&PC=U531). (2024)

[3] J.M. Paggi, A. Pandit, R.O. Dror. The Art and Science of Molecular Docking. *Annual Review of Biochemistry*. 7.1-7.16 (2024).

[4] D.R. Koes, M.P. Baumgartner, C.J. Camacho. Lessons Learn in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*. **53**, 1893-1904 (2013).

[5] R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G.R. Lee, F.S. Morey-Burrows, I. Anishchenko, I.R. Humphreys, R. McHugh, D. Vafeados, X. Li, G.A. Sutherland, A. Hitchcock, C.N. Hunter, A. Kang, E. Brackenbrough, A.K. Bera, M. Baek, F. DiMaio, D. Baker. Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. *Science*. **384**, eadl2528 (2024).

[6] G. Corso, A. Deng, B. Fry, N. Polizzi, R. Barzilay, T. Jaakkola. Deep Confident Steps to New Pockets: Strategies for Docking Generalization. *Arxiv*. (2024).

[7] G. Corso, H. Stark, B. Jing, R. Barzilay, T. Jaakkola. DiffDock: Diffusion Steps, Twists and Turns for Molecular Docking. *Arxiv*. (2023).

[8] [A. Gimeno](https://pubmed.ncbi.nlm.nih.gov/?term=Gimeno%20A%5BAuthor%5D), [M.J. Ojeda-Montes](https://pubmed.ncbi.nlm.nih.gov/?term=Ojeda-Montes%20MJ%5BAuthor%5D), [S. Tomás-Hernández](https://pubmed.ncbi.nlm.nih.gov/?term=Tom%C3%A1s-Hern%C3%A1ndez%20S%5BAuthor%5D), [A. Cereto-Massagué](https://pubmed.ncbi.nlm.nih.gov/?term=Cereto-Massagu%C3%A9%20A%5BAuthor%5D), [R. Beltrán-Debón](https://pubmed.ncbi.nlm.nih.gov/?term=Beltr%C3%A1n-Deb%C3%B3n%20R%5BAuthor%5D), [M. Mulero](https://pubmed.ncbi.nlm.nih.gov/?term=Mulero%20M%5BAuthor%5D), [G. Pujadas](https://pubmed.ncbi.nlm.nih.gov/?term=Pujadas%20G%5BAuthor%5D), [S. Garcia-Vallvé](https://pubmed.ncbi.nlm.nih.gov/?term=Garcia-Vallv%C3%A9%20S%5BAuthor%5D). The Light and Dark Sides of Virtual Screening: What is There to Know?. *PubMed Central*. **20**, 1375 (2019).

[9] A. Aouidate. A Beginner’s Guide to Molecular Docking with Smina. <https://loopsnstrands.com/a-beginners-guide-to-molecular-docking-with-smina-e1e4360950c2>. (2023).

[10] A. Gupta. Spearman’s Rank Correlation: The Definitive Guide to Understand. <https://www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correlation#:~:text=Spearman's%20rank%20correlation%20measures%20the,represented%20using%20a%20monotonic%20function>. (2024).