ESSENCE-Dock: A Consensus-Based Approach to Enhance Virtual Screening Enrichment in Drug Discovery

Jochem Nelen[1], Miguel Carmena-Bargueño[1], Carlos Martínez-Cortés[1], Alejandro Rodríguez-Martínez[1], José Manuel Villalgordo-Soto[2] and Horacio Pérez-Sánchez[1,*]

[1]Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), HiTech Innovation Hub, UCAM Universidad Católica de Murcia, 30107 Murcia, Spain
[2]Eurofins-Villapharma, Parque Tecnológico de Fuente Álamo, E-30820 Murcia, Spain
[*] Email: hperez@ucam.edu

Abstract

Drug development is a complex, costly, and time-consuming endeavor. While high-throughput screening (HTS) plays a critical role in the discovery stage, it is one of many factors contributing to these challenges. In certain contexts, virtual screening can complement HTS, potentially offering a more streamlined approach in the initial stages of drug discovery. Molecular docking is an example of a popular virtual screening technique that is often used for this purpose, however, its effectiveness can vary greatly. This has led to the use of consensus docking approaches, which combine results from different docking methods to improve the identification of active compounds and reduce the occurrence of false positives. However, many of these methods do not fully leverage the latest advancements in molecular docking.
In response, we present ESSENCE-Dock (Effective Structural Screening ENrichment ConsEnsus Dock), a new consensus docking workflow aimed at decreasing false positives and increasing the discovery of active compounds. By utilizing a combination of novel docking algorithms, we improve the selection process for potential active compounds. ESSENCE-Dock has been made to be user-friendly, requiring only a few simple commands to perform a complete screening, while also being designed for use in high-performance computing (HPC) environments.

Introduction

The development of novel drugs is a very long and expensive process: depending on the context, the costs of bringing new drugs to market can span from several hundred million to over 4 billion U.S. dollars (1). Many factors contribute to these high costs, including costly clinical trials and toxicity studies among others. Another part of the high costs can be attributed to hit discovery through expensive, time-consuming high-throughput screening (HTS) (2).

Screening small libraries in this manner is done routinely, but chemical space is vast, and navigating it using HTS exclusively can be very costly, both in time and money.

Virtual screening techniques are capable of computationally identifying potential hit compounds, thereby lowering the number of compounds to test experimentally. This can be done much faster and cheaper than HTS, and thus applied to many more compounds, allowing the exploration of a larger chemical space (3). In short, virtual screening techniques are valuable tools for lowering costs and speeding up novel medicine development. Many different types of virtual screening methods exist, but one of the most used techniques is molecular docking (4). For this structure-based screening method, a 3D structure of both the target protein and ligand needs to be available. The goal of molecular docking is to find the optimal ligand pose, which typically has an associated docking score. Subsequently, this docking score can be used to compare different ligands and rank them to obtain a list of the most promising candidates.

The performance of docking algorithms can vary significantly across different targets (5). Examples of popular docking methods include AutoDock Vina (6), and its subsequent forks Smina (7) and Gnina (8), but also commercial software such as LeadFinder (9) and Glide (10) are routinely used in both industry and academia. These methods can be effective on their own, but consensus docking approaches have been shown to improve reliability (11). More specifically, by implementing a consensus approach one can reduce false positives and improve the overall accuracy of results. However, many of these existing consensus methods rely on older docking techniques with lower accuracy. Furthermore, they often overlook the importance of the predicted pose similarity, despite its demonstrated effectiveness (12).

To provide a brief overview of existing consensus approaches, we begin with the early yet effective work by A. Kukol (13). In this work, different docking rankings from different algorithms were combined in order to determine which combination of docking methods is the most effective. Specifically, the Directory of Useful Decoys (DUD) (14) was used for this purpose. It was reported that AutoDock 4 (15) combined with AutoDock Vina yielded the best results overall. Houston and Walkinshaw (16) built on these findings by incorporating a simple form of binding pose similarity for additional consensus, which proved to be quite effective at lowering false positives. Their method started by performing molecular docking using both AutoDock Vina and AutoDock 4. Subsequently, they quantified the binding pose similarity by calculating the root mean square deviation (RMSD) between the binding poses of each ligand. In case the RMSD was larger than 2Å, the molecule was rejected and filtered out from further calculations due to the disagreement between the two docking methods. If the RMSD was smaller or equal to 2Å, the compounds were kept for further analysis and ranking. Although simple, this filter step based on RMSD pose-similarity managed to identify many of the decoy compounds, thus demonstrating its value in the whole protocol.

After these developments, consensus docking has continuously been further explored by many researchers. A brief overview of this has been provided in a review of molecular docking written by Torres et al. (11). Some of the most recent consensus docking publications include DockECR (17) and MILCDock (18). Palacio-Rodríguez et al. described an exponential consensus ranking

(ECR) method that uses an exponential distribution for each individual rank of every docking method (19). Their method is implemented in such a way that compounds that score very well in a single method can still be ranked highly, even though they score worse in the other docking methods. The DockECR publication is an in-depth exploration of the ECR method and was collaboratively written with one of the original authors of the exponential consensus ranking paper. They developed an open-source program that has implemented the ECR method and supports several docking methods, such as LeDock (20), rDock (21), Smina (7), and AutoDock Vina (6). Additionally, they added a binding pose similarity metric, which was shown to improve the results in most cases. Their protocol also supports the use of parallelization, enabling simultaneous docking runs across multiple CPUs. Finally, one of the latest works regarding consensus docking is MILCDock. Here, the authors describe a consensus docking method that is enhanced by machine learning. To gather training data, the authors performed molecular docking on datasets such as DUD-E (22) using various open-source docking algorithms including AutoDock Vina, AutoDock 4, and rDock. Subsequently, they used the resulting docking data and binding pose similarities to train machine learning models that can be used as a scoring function. Their results show substantial improvements when compared to the individual docking runs.

Inspired by this prior research, we developed ESSENCE-Dock: a docking workflow integrating data generated by different docking methods, effectively reducing false positives and enhancing the enrichment of active compounds. This is achieved by prioritizing consensus from different docking techniques, both in docking scores and binding poses, effectively leveraging the strengths of each docking method while mitigating a lot of the individual limitations. High consensus scores from our protocol correlate with a higher likelihood of compound activity, as demonstrated by our reported high enrichment factors across various DUD-E targets compared to the base methods (Table 1 in the Results section). These observed enrichment factors highlight its promising efficacy and applicability in the hit identification process. Our workflow is flexible, user-friendly, and able to be performed using High-Performance Computing (HPC) clusters. The entire screening process can be streamlined to just four commands: one for each of the individual docking methods — DiffDock (23), Gnina (8), and LeadFinder (9) — and a fourth to initiate the actual consensus calculations which performs the final rescoring. For this, ESSENCE-Dock uses the information from the individual docking runs such as docking scores, binding pose similarity, and ligand flexibility in order to compute a comprehensive consensus score which is used to finally rerank all of the screened compounds. Additionally, all of the top results can be saved to an interactive PyMol Session (PSE) file (24) for visual inspection and user convenience. Furthermore, predicted protein-ligand interactions for these top results are calculated using the Protein-Ligand Interaction Profiler (PLIP) (25), and added to the PSE file as well. These protein-ligand interaction results are also saved in tabular and JSON formats for easy integration with potential further post-docking analysis.

In summary, we introduce ESSENCE-Dock (Effective Structural Screening ENrichment ConsEnsus Dock), a novel consensus docking workflow that addresses some of the identified gaps in current consensus docking methods. ESSENCE-Dock leverages state-of-the-art docking techniques and takes into account binding pose similarity and ligand flexibility to

calculate a final consensus score. Our workflow is easy to use, only requiring a few commands for the whole protocol. It is also able to run on HPC clusters using Singularity and has the potential to streamline drug discovery efforts by focusing resources on acquiring and testing compounds with high consensus scores.

Materials and Methods

As established in the introduction, consensus analysis of docking techniques can substantially improve the outcome of a virtual screening campaign (4). With this in mind, we set out to develop our own consensus docking procedure, with a focus on lowering false positives in order to obtain a high enrichment in the top fraction of high-scoring compounds.

As a first step, we attempted to select different types of docking techniques. We opted for methods that make use of very different sampling methods for pose generation and use different types of scoring functions. This makes it more significant when consensus, both in high docking scores and binding poses, is found because the docking algorithms came to the same conclusion by employing completely different methods. We ended up selecting LeadFinder (9), Gnina (8) and DiffDock (23). These docking methods, along with their specific sampling and scoring algorithms are discussed in more detail below. A schematic overview of the general workflow is provided in Fig. 1.

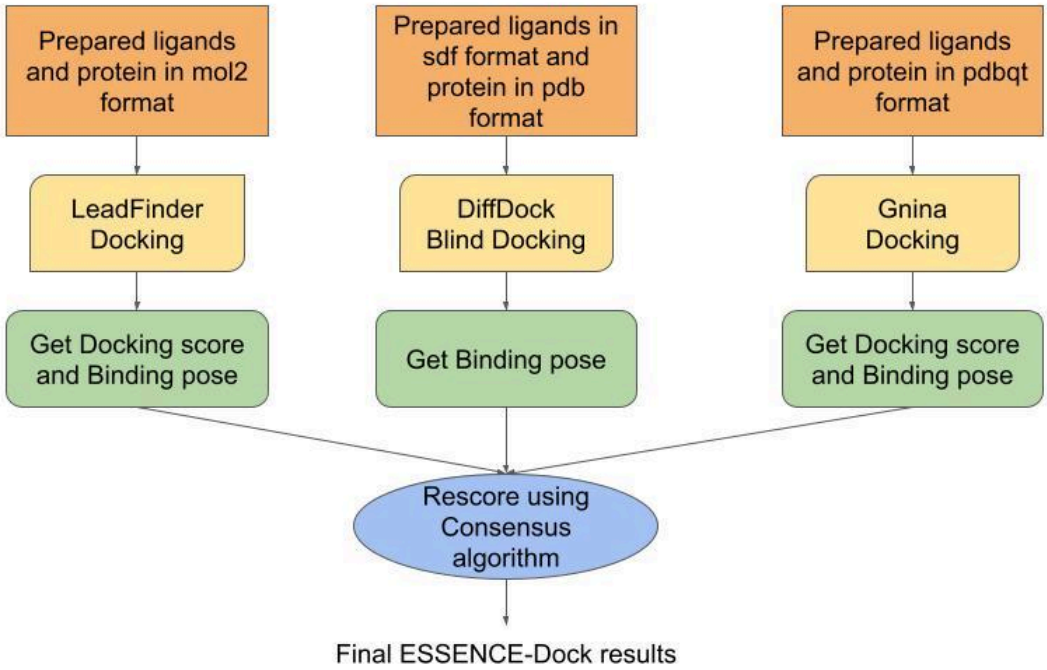

Fig. 1: Overview of the ESSENCE-Dock Workflow. Each ligand is prepared for and docked using LeadFinder, DiffDock, and Gnina. The resulting docking scores and binding poses are subsequently combined into a final ESSENCE-Dock score, which is used for ranking the compounds.

Ligand and Protein Preparation

Proper preparation of protein and ligands is a crucial step when performing molecular docking, and helps to ensure optimal results. Each of the docking techniques used in our work - LeadFinder (LF), DiffDock (DD), and Gnina (GN) - requires distinct formats for both ligands and proteins. LeadFinder requires the use of mol2 format for both ligands and proteins, DiffDock performs best with ligands in SDF format and protein structures in PDB format, while Gnina provides optimal results with structures in pdbqt format for both proteins and ligands. This section will detail the procedures used to prepare the ligands and proteins for all the docking calculations, starting with ligand preparation.

Ligands

The selected DUD-E target sets were downloaded from the DUD-E database, with the files 'decoys_final.ism' and 'actives_final.ism' serving as the starting point for conversion (see Fig. 2).

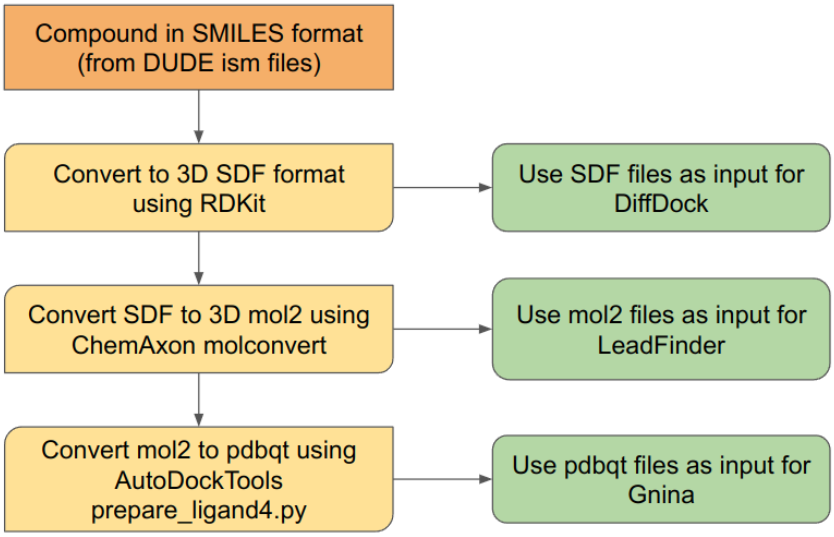

Fig.2: Ligand Preparation Workflow.
Input compounds in SMILES format were processed and converted to 3D SDF, mol2, and pdbqt files for docking with DiffDock, LeadFinder, and Gnina, respectively.

Using Python and RDKit (26), the smiles were processed, hydrogens were added, and a 3D conformation was generated for each compound using RDKit's ETKDG method. The molecules were saved as separate SDF files. Each file was named according to the ID provided in the original .ism files and was tagged as 'active' or 'inactive' for more efficient post-processing after docking.

These SDF files were used as ligands for the DiffDock method but were also converted to mol2 files using ChemAxon's molconvert tool (27). This generated 3D coordinates of a low-energy conformer using the MMFF94 forcefield (28), making the molecules ready to be docked directly using LeadFinder.

Subsequently, these mol2 files were used as input for the conversion to pdbqt format using the "prepare_ligand4.py" script from AutoDockTools (ADT) (15), which were then used for docking with Gnina.

Protein Structures

For the protein preparation of the DUD-E targets, the PDB code of the provided receptor.pdb file was used to obtain the final receptor files. A schematic overview of the protocol is shown in Fig. 3.

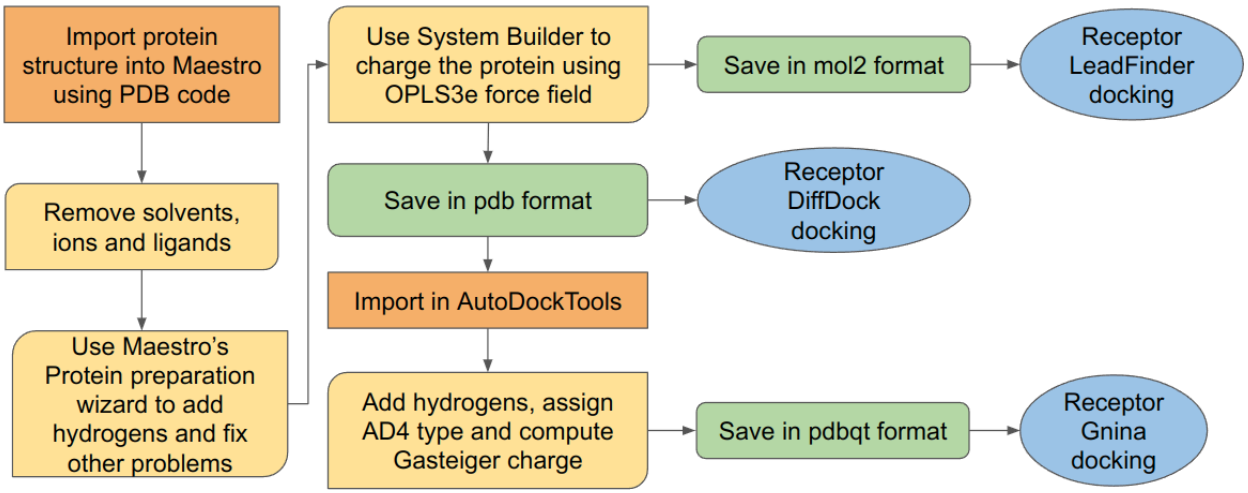

Fig. 3: Protein Structure Preparation Workflow.
Protein structures are imported using their PDB code in Maestro. After processing, the structures are saved as both PDB and mol2 files, which are used for DiffDock and LeadFinder, respectively. Additionally, the PDB file is converted to pdbqt format using AutoDockTools for Gnina docking.

The first step after importing the PDB structure in Schrödinger's Maestro (version 2020-4) (29) was to remove all solvents, ions, and ligands. Subsequently, the Protein Preparation Wizard in Maestro was used to properly (re)assign bond orders, add missing hydrogens, create disulfide bonds where necessary, and fill in missing side chains. Next, Maestro's system builder was used to charge the protein using the OPLS3e forcefield (30). Concretely, this was accomplished by minimizing the box volume, defining the solvent model as none, the box shape as orthorhombic, and 10 Å buffer along every dimension.
After these steps, the processed protein was saved in both mol2 and pdb format. The mol2 file was directly used as the receptor file for the LeadFinder docking calculations, while the pdb file was used as input for DiffDock. The pdb file underwent further processing to create a pdbqt file using AutoDockTools (ADT) from MGLTools (version 1.5.7) (15). Concretely, hydrogens were added with the settings "All Hydrogens" and "noBondOrder", followed by assigning AD4 type to all atoms. As a last step, the Gasteiger charge was computed, and the resulting file was saved as a pdbqt file for use in the Gnina docking process.

Docking Techniques

In the ESSENCE-Dock consensus approach, we deliberately utilize docking methods with different sampling and scoring functions, enhancing the significance of consensus findings in both docking scores and binding poses. Here, we will briefly discuss the individual docking techniques used for the current implementation of our ESSENCE-Dock workflow, detailing their methodology for pose sampling and scoring functions, as well as the docking parameters we used to process the input ligands.

LeadFinder

LeadFinder (LF), developed by BioMolTech and distributed by Cresset, uses a speed- and accuracy-enhanced genetic algorithm for ligand conformation generation (9). This algorithm creates a diversified pool of conformations, which are subsequently ranked using the LF scoring function. The output includes the best compounds with their docking score, predicted binding energy, and corresponding docked pose in mol2 format. LF's scoring function is semi-empirical and explicitly accounts for various types of molecular interactions (9). Overall, LF is a performant docking algorithm that has been shown to perform well across various protein targets (31).

We ran the LF calculations using Metascreener v1.1 (32), an open-source and publicly available tool that allows a user to perform molecular docking on high-performance computing (HPC) clusters. This meant that LF could run in parallel, reducing the overall run time substantially and making it feasible to screen larger libraries. The LF docking was executed using LeadFinder version 2104, along with standard variables and a grid size where X, Y, and Z were all set to 30. As user input, only the docking coordinates, as well as the protein structure and the directory containing the input ligands, were included as input variables.

Gnina

Another docking method we used in our consensus docking approach is Gnina (8). Gnina is a fork of Smina (7), which in turn is a fork of Autodock Vina (Vina) (6). In contrast with LeadFinder, which utilizes a genetic algorithm to explore a ligand’s conformational space, Gnina employs Monte Carlo sampling to generate a diverse ensemble of ligand conformations. It also implemented an ensemble of convolutional neural networks (CNN) as a scoring function, which outperforms the scoring functions of its predecessors Vina and Smina in most cases (8). Although Vina’s scoring function is still used during pose generation and refinement for computational efficiency, Gnina's CNN pose score ensemble model rescores and sorts the final poses. These settings, found to be the most efficient by the authors, were also used for our Gnina docking calculations. As a metric for the final predicted binding energy, we used the Vina score of the most optimal pose as determined by the CNN pose score for further processing.

For our calculations, the Gnina docking algorithm was implemented in Metascreener v1.1, again providing support to run the calculations massively in parallel on HPC clusters. Because of the CNN scoring function, Gnina is able to run a lot faster on a GPU compared to a CPU. Despite this, the parallel execution of calculations on an HPC cluster made processing the selected DUD-E datasets still feasible. Gnina Version 1.0 was used with standard settings, a box size of

30 in all dimensions, along with the user-provided protein, directory of ligands and docking coordinates.

DiffDock

The final docking technique we integrated into our consensus docking workflow is DiffDock (23). DiffDock is a novel, state-of-the-art blind docking method based on geometric deep learning. DiffDock refines randomized ligand conformations at various protein locations via a trained reverse diffusion process, simultaneously determining the best binding pocket and conformation. This eliminates the need for defining a pre-designated search box as required by LeadFinder and Gnina. Because this is a blind docking method, the whole protein and all of its potential binding pockets are explored.

Unlike the other methods, DiffDock does not provide a scoring function. It does offer a confidence score indicating prediction certainty. Although this is useful, it still does not allow a user to estimate how potent a ligand binds, making it difficult to rank different ligands accurately among each other. Still, DiffDock is capable of predicting a ligand's binding pose with state-of-the-art accuracy, so it is still very useful to compare it to the binding poses generated by LeadFinder and Gnina.

To facilitate the use of DiffDock in virtual screening, we developed a fork of DiffDock named DiffDockHPC. Like Metascreener, it can split up the ligand dataset and launch it as separate jobs, meaning the calculations can run in parallel. This significantly speeds up calculations, and is especially useful when processing more extensive libraries. DiffDockHPC is fully open-source and freely available at https://github.com/Jnelen/DiffDockHPC. Like Gnina, DiffDock can be greatly accelerated by GPUs, but running the calculations massively in parallel on CPUs still makes it feasible to be a part of the consensus docking workflow.

Processing of Docking Results

In this section, we describe the processing of docking results, the details of our scoring formula, and the final output files that our ESSENCE-Dock workflow generates.

Scoring Formula

In order to effectively rank compounds, we developed a scoring formula that considers not only individual docking scores, but also incorporates the binding pose similarity and the ligand's flexibility. The complete ESSENCE-Dock rescoring formula is presented in Eq. 1:

$$ESSENCEDockScore = \sqrt{\frac{2}{RMSDaverage}} \cdot MBE \cdot \sqrt{ln(RB + 1) + 1} \quad (1)$$

Here, RMSDaverage denotes the average root mean square deviation (as shown in Eq. 2), and MBE represents the Mean Binding Energy (computed as per Eq. 3). Both $BindingEnergy_{GN}$ and $BindingEnergy_{LF}$ are denoted in kcal/mol. Gnina's final docking score is based on Vina's scoring function, applied to the most optimal pose as determined by Gnina's CNN pose score. RB corresponds to the number of rotatable bonds. The RMSD and the number of rotatable bonds

are calculated using OpenBabel's obrms and obrotamer functionalities respectively (33). During the RMSD calculations, obrms takes potential internal molecular symmetry into account, providing more robust and accurate results.

$$RMSDaverage = \frac{RMSD_{DD\_GN} + RMSD_{DD\_LF} + RMSD_{GN\_LF}}{3} \quad (2)$$

$$MBE = \frac{BindingEnergy_{GN} + BindingEnergy_{LF}}{2} \quad (3)$$

The core principle underpinning the ESSENCE-Dock formula is the concept of consensus, which includes both predicted binding energies and binding poses. Our starting point was a consensus of the docking scores: the Mean Binding Energy, which showed the best performance when comparing the individual docking methods, along with their mean and median scores (Fig. 4 in the Results section). With this as the starting point, we attempted to improve this baseline by including a form of binding pose consensus. In the literature, an RMSD difference smaller or equal to 2 is often chosen as a threshold for similar poses (11). Hence, we incorporated the concept of binding pose consensus by dividing 2 by the average RMSD. If the average RMSD is less than 2, the overall score improves; otherwise, the final score will become worse. To prevent a disproportionate impact of really low or high consensus values, the square root is taken to balance out the impact of extremes. Finally, to mitigate a potential bias toward rigid structures with few rotational bonds, we also introduced a flexibility term based on the number of rotational bonds.

To conclude, we would like to point out that only the average RMSD, docking scores, and the number of rotatable bonds are required as input metrics. In this sense, our approach is flexible and can be used with basically any other docking program that produces a docking pose and score. For user convenience, the whole ESSENCE-Dock approach has been implemented in Metascreener v1.1, meaning the consensus rescoring and post-processing can be executed using just a single command.

Enrichment Factors

To validate the protocol, we selected a diverse set of DUD-E targets and computed enrichment factors (EFs) at different thresholds. We compared these EFs with those obtained from individual docking methods (LeadFinder and Gnina) as well as a more basic consensus docking approach using the mean binding energy and mean rank as scoring metrics. The EF (34) for the top 5, 1, 0.5, and 0.1% was calculated using Eq. 4:

$$EF = \frac{activesEF\%/totalEF\%}{AllActives/AllCompounds} \quad (4)$$

The EF compares the fraction of active compounds within a specific top percentage of ranked compounds (ActivesEF%) to the overall distribution of actives in the dataset. This essentially compares the distribution of actives at the top of the ranking compared to random chance. A high EF signifies a higher concentration of true positives (TPs) with few false positives (FPs)

within that top fraction, while a low EF suggests many FPs ranked at the top. This makes it a valuable metric to assess a method's effectiveness in early-stage hit identification.

Output Files

After scoring and ranking all the compounds using Eq. 1, ESSENCE-Dock is able to generate an interactive PyMol (24) session (PSE) file. This PSE file contains the protein structure and the top 50 results (user-configurable) based on the consensus (re)scoring. Each entry in the PSE file displays the three docked structures from LeadFinder, DiffDock, and Gnina, enabling easy visual assessment of binding pose similarity. Additionally, the file includes predictions of protein-ligand interactions as predicted by PLIP (25), offering insights into the potential binding interactions. Furthermore, these PLIP results are also summarized in both JSON and spreadsheet-style CSV files for convenient further analysis. Finally, all of the information that ESSENCE-Dock uses (the individual docking scores, RMSD values, etc.) for all compounds is saved to an output CSV file as well, summarizing all of the key data for every docked compound.

Results and Discussion

To demonstrate the efficacy of ESSENCE-Dock, we applied our workflow to a set of 21 distinct DUD-E targets, representing a diverse range of protein families and biological contexts. After docking and ranking the compounds, the enrichment factors (EF) were calculated at different fractions. EFs are a measure of the method's ability to identify active compounds among the top-ranked candidates, with higher EF values indicating better performance (see Materials and Methods Eq. 4 for the exact EF formula). We evaluated EFs for the individual DUD-E targets at various fractions: the top 5%, top 1%, top 0.5%, and top 0.1%, in order to analyze ESSENCE-Dock’s efficacy in different scenarios. To compare and benchmark ESSENCE-Dock's performance, we also computed the EFs at the same threshold levels using the rankings from the individual docking methods: LeadFinder (LF) and Gnina (GN). We also calculated the EFs for DiffDock by ranking the compounds based on the calculated Confidence metric. However because this was not intended to be used as a virtual screening metric, we refer to Table S1 of the supplementary data. Additionally, we calculated two basic consensus metrics using the individual GN and LF docking calculations: the Mean Binding Energy (MBE), which provides insights into the collective binding affinities, and the Mean Rank (MR), which offers a measure of the ranking consistency across different docking methods. All of these results are summarized in Table 1. The average results are compared using a bar chart in Fig. 4. The same information for the Median and all of the results for the individual DUD-E targets can be found in the Supplementary Data (Figs. S1-23).

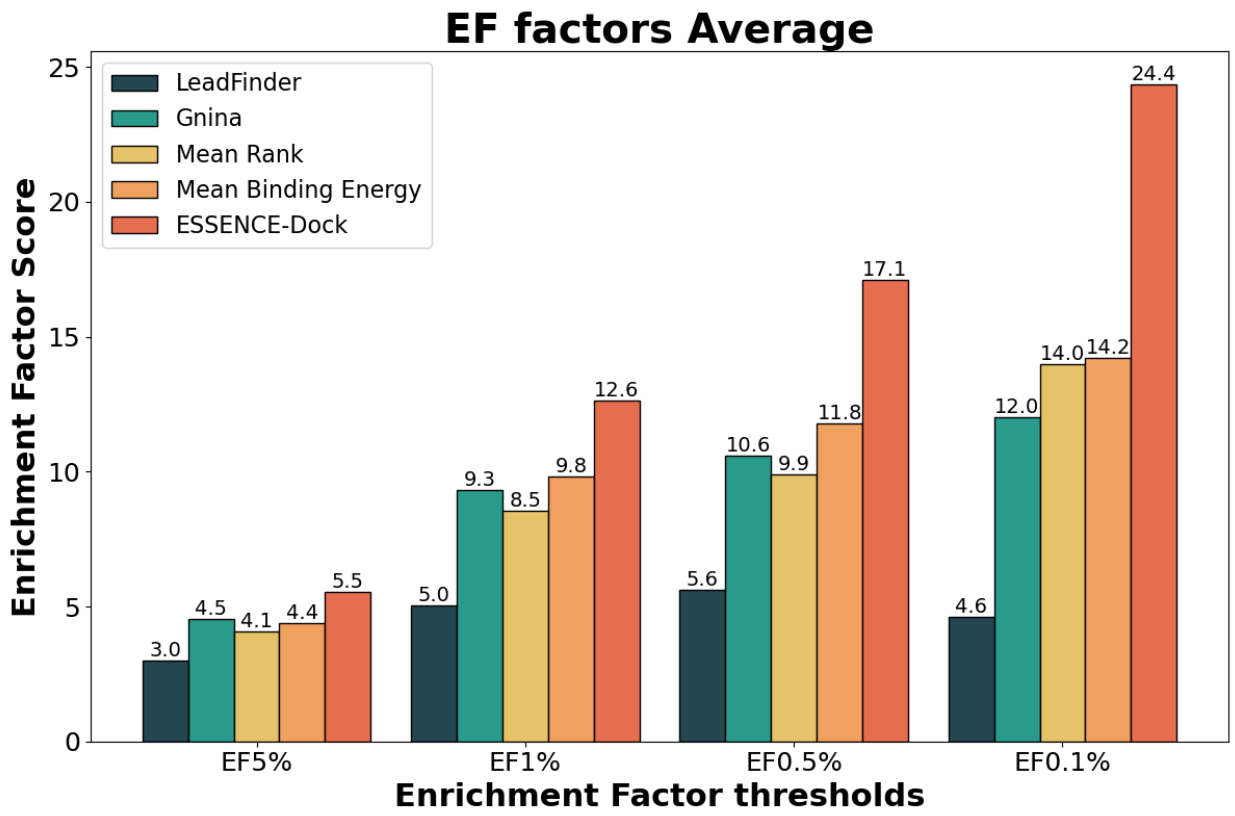

Fig. 4: Bar Chart of Average Enrichment Factors (EF) Across 21 Various DUD-E Targets at Varying EF Thresholds.

This figure provides a comparative analysis of the average Enrichment Factors (EF) across 21 distinct DUD-E targets. The Y-axis represents the average enrichment factor, while the X-axis indicates the corresponding EF threshold values. The color-coded bars correspond to the different ranking methods: LeadFinder, Gnina, Mean Rank, Mean Binding Energy, and ESSENCE-Dock

Table 1: Comparative Performance of ESSENCE-Dock and Reference Docking Methods Across 21 DUD-E Targets at Various Enrichment Factors (EF%). Table 1 compares the performance of ESSENCE-Dock (Cons.) with reference methods (LeadFinder - LF, Gnina - GN, Mean Rank - MR, and Mean Binding Energy - MBE) across 21 DUD-E targets at various Enrichment Factors (EF%). The table presents EF values for top 5%, top 1%, top 0.5%, and top 0.1%. Shaded cells indicate method rankings, with green denoting higher comparative EF and red indicating lower comparative EF.

	EF5%					EF1%					EF0.5%					EF0.1%				
	LF	GN	MR	MBE	Cons.	LF	GN	MR	MBE	Cons.	LF	GN	MR	MBE	Cons.	LF	GN	MR	MBE	Cons.
ADA	3.9	3.2	4.3	4.1	5.6	1.1	0.0	9.8	5.4	11.9	0.0	0.0	8.5	4.3	21.3	0.0	0.0	9.9	0.0	39.7
AKT1	1.7	3.6	2.7	3.3	1.4	1.7	8.6	4.4	6.5	2.7	2.7	12.2	6.8	8.8	2.7	0.0	26.9	6.7	16.8	3.4
AMPC	0.8	0.8	0.8	0.0	2.9	0.0	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CASP3	2.9	4.2	5.5	5.0	5.0	9.5	4.5	10.6	11.1	9.5	11.2	4.1	13.2	15.2	17.2	5.0	0.0	14.9	24.9	39.8
CP3A4	2.8	1.5	2.4	2.9	2.0	4.7	1.8	2.3	4.1	1.8	8.2	2.3	4.7	7.0	0.0	11.7	5.9	5.9	5.9	0.0
CXCR4	2.0	0.5	1.0	0.5	0.5	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FA7	5.6	8.6	9.0	9.3	7.5	8.7	14.8	20.1	20.9	17.4	7.0	17.4	29.7	24.4	22.7	0.0	0.0	55.8	37.2	55.8
FABP4	0.4	8.1	3.4	5.5	6.4	0.0	21.3	2.1	8.5	10.6	0.0	25.5	0.0	8.5	8.5	0.0	0.0	0.0	0.0	0.0
FPPS	0.0	5.6	0.0	0.0	9.4	0.0	2.4	0.0	0.0	20.1	0.0	0.0	0.0	0.0	28.0	0.0	0.0	0.0	0.0	11.7
GCR	0.7	5.3	2.9	3.9	4.7	0.0	15.8	6.2	13.5	13.9	0.0	21.8	7.0	21.0	21.0	0.0	47.3	7.9	23.7	27.6
GLCM	5.5	1.1	3.7	4.1	5.5	11.0	1.8	3.7	11.0	5.5	11.3	0.0	3.8	15.0	7.5	0.0	0.0	0.0	0.0	0.0
HIVPR	5.1	4.5	7.1	7.0	6.5	5.4	6.2	14.2	11.2	9.7	4.5	6.0	17.6	14.2	12.0	3.8	9.4	26.3	24.4	26.3
HIVRT	2.5	3.3	3.8	3.8	4.2	2.7	3.6	5.6	5.6	8.9	1.2	4.7	7.1	5.9	13.0	0.0	9.0	9.0	6.0	18.0
HS90A	1.8	0.0	0.2	0.5	3.9	2.3	0.0	0.0	0.0	8.0	2.2	0.0	0.0	0.0	13.5	0.0	0.0	0.0	0.0	11.2
ITAL	3.3	3.2	2.9	3.3	3.0	8.0	11.6	11.6	13.1	9.5	13.1	8.7	21.8	20.4	17.5	7.0	7.0	55.6	48.7	48.7
KIF11	1.2	8.3	6.9	7.8	7.4	0.0	21.4	8.6	16.3	24.0	0.0	30.9	6.9	17.2	34.3	0.0	25.7	0.0	0.0	25.7
MK14	2.5	3.3	3.3	3.9	2.4	5.2	9.2	6.6	8.3	5.4	5.9	13.5	9.7	12.8	9.3	1.8	31.5	21.0	29.8	29.8
NRAM	0.6	1.2	3.1	2.0	7.5	0.0	0.0	2.0	0.0	17.3	0.0	0.0	0.0	0.0	29.0	0.0	0.0	0.0	0.0	53.6
PA2GA	2.4	5.3	1.6	1.8	8.9	0.0	6.1	1.0	1.0	25.5	0.0	6.1	2.0	2.0	30.6	0.0	10.6	10.6	10.6	31.8
TRY1	3.0	5.9	6.0	6.1	4.0	3.6	6.5	6.7	8.2	8.2	3.1	7.6	7.6	9.4	12.0	6.8	18.1	9.1	9.1	27.2
WEE1	14.3	17.6	15.3	16.8	17.6	41.8	60.3	61.3	61.3	53.5	47.5	61.3	61.3	61.3	59.3	61.3	61.3	61.3	61.3	61.3
Average	3.0	4.5	4.1	4.4	5.5	5.0	9.3	8.5	9.8	12.6	5.6	10.6	9.9	11.8	17.1	4.6	12.0	14.0	14.2	24.4
Median	2.5	3.6	3.3	3.9	5.0	2.3	6.1	5.6	8.2	9.5	2.2	6.0	6.9	8.8	13.5	0.0	5.9	7.9	6.0	26.3

As illustrated in Table 1, our ESSENCE-Dock method often outperformed the other methods, especially at the smaller EF fractions of EF0.5% and EF0.1%. In the case of EF0.1%, ESSENCE-Dock was only surpassed by another method in five cases: AKT1, CP3A4, GCR, ITAL, and MK14. Even when our approach performed worse, it often still showed notable enrichment. Only in the cases of AKT1 and CP3A4 did ESSENCE-Dock substantially underperform other methods. For the other three DUD-E targets, ESSENCE-Dock's performance remained competitive with high EFs of 27.6, 48.7, and 29.8 for GCR, ITAL, and MK14, respectively. In some other cases such as AMPC, CXCR4, FABP4, and GLCM, none of the methods manage to identify actives in the top 0.5% and 0.1% of ranked compounds, resulting in EFs of 0. However, an important note is that these specific datasets are relatively small (ranging from only ~2900 to ~3850 compounds). This means that at these top enrichments, only a few compounds are considered, which can partly explain why all of the methods score substantially worse here.

The individual EFs were averaged for each method and EF threshold, as summarized in Fig. 4. This figure clearly demonstrates that our approach achieved the highest average enrichment scores across all evaluated thresholds. While the improvement at EF5% is relatively modest, the improvement becomes more evident as the EF threshold decreases. This is especially the case with EF0.1%, where ESSENCE-Dock has an enrichment factor that is 5 times higher than LeadFinder and 2 times higher than Gnina. Furthermore, even in comparison to the Mean Rank and Mean Binding Energy consensus methods, ESSENCE-Dock is still nearly twice as effective.

Fig. 5 shows histograms comparing the ranking performance of the individual docking methods (GN and LF) and the ESSENCE-Dock approach for three distinct scenarios of varying performance: NRAM (A), CASP3 (B), and CP3A4 (C). In the NRAM scenario, ESSENCE-Dock outperforms both individual docking methods substantially. For CASP3, the individual methods already exhibit strong independent performance, and combining the results using the ESSENCE-Dock consensus approach further enhances the overall ranking. Conversely, CP3A4 demonstrates a scenario with limited consensus between methods, resulting in lower performance by ESSENCE-Dock. These specific examples are explored in greater detail below. Other targets could have been chosen to be featured here, and similar figures for all of the tested DUD-E datasets are provided in the Supplementary Data (Figs. S24-44). Each figure within this comparison features two separate histograms: one representing the distribution of active compounds (in blue) and the other depicting the distribution of decoy compounds (in orange). Of note are the varying quantities of active and decoy compounds which lead to separate Y-axes for each group. This distinction is caused by the significantly larger number of decoys compared to active compounds. For all docking methods, lower (more negative) scores on the X-axis indicate a better result.

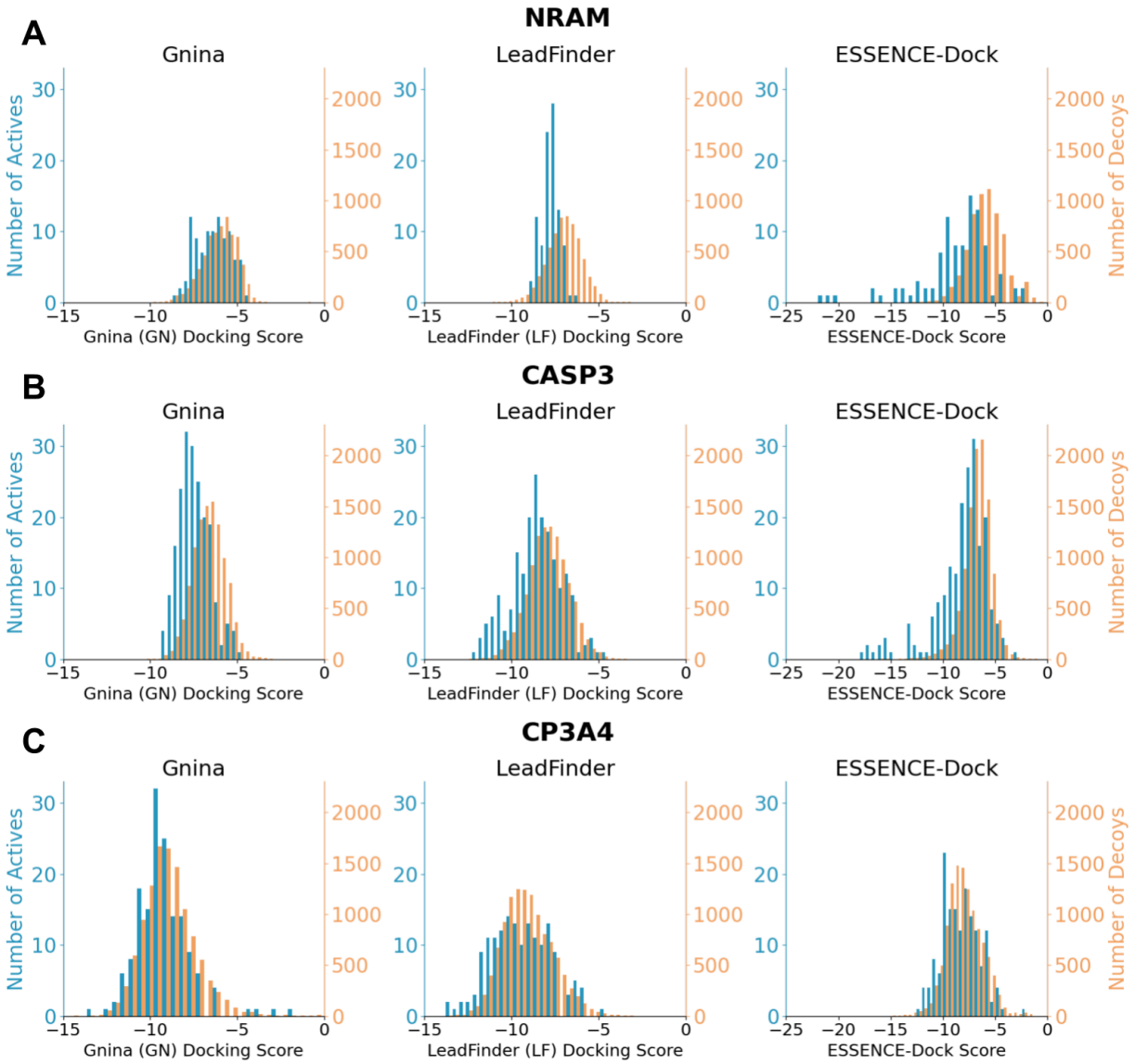

Fig. 5: Comparative Docking Performance of Gnina, LeadFinder, and ESSENCE-Dock for NRAM, CASP3, and CP3A4 from the DUD-E Dataset.
This figure presents the distributions of actives (depicted in blue) and decoys (represented in orange) based on their respective docking scores for the NRAM (A), CASP3 (B), and CP3A4 (C) datasets included in the DUD-E database. Note that the figure employs differently scaled axes to accommodate the varying quantity of actives (left axis) compared to decoys (right axis). In all cases, a lower score (more negative) indicates a better-ranked sample.

In the example of NRAM (Fig. 5A), the individual docking methods GN and LF don't excel at assigning better docking scores to the active compounds compared to the decoys. This is reflected in the overlapping distributions between the actives and decoys in these cases. In contrast, the assigned ESSENCE-Dock scores produce a distinct cluster of active compounds distributed within the range of -25 to around -15, clearly separated from the decoy compounds. These trends align with the quantitative data presented in Table 1, where our ESSENCE-Dock

method at EF0.5% and EF0.1% for NRAM demonstrates substantial enrichment, while the other methods yield enrichment factors of zero.

In the case of CASP3 (Fig. 5B), the individual docking methods GN and LF demonstrate better performance compared to the NRAM case, successfully ranking a higher number of actives near the top based on their respective docking scores. Still, ESSENCE-Dock is able to outperform the other methods using the consensus approach. Again, this is reflected in Table 1, where at EF0.5% and especially EF0.1%, ESSENCE-Dock shows a higher EF than the individual docking methods, as well as the more simplistic MR and MBE consensus results. Interestingly, again there is a distinct cluster of active compounds present between the -20 and -15 scores.

Finally, Fig. 5C shows the example of CP3A4 from the DUD-E database. This is a case where ESSENCE-Dock performs worse compared to the other methods, as can also be seen in the EF0.1% results for CP3A4 in Table 1. Even the individual docking methods, GN and LF provide better enrichment than ESSENCE-Dock in this specific example. While these results for ESSENCE-Dock on the CP3A4 dataset seem disappointing, they are also informative. The observation that ESSENCE-Dock assigns lower scores to nearly all compounds in this dataset indicates its effectiveness in detecting a lack of consensus, which, in turn, suggests that the results may be unreliable.

In this context, the ESSENCE-Dock score can also serve as a confidence metric for the consensus results. When the score is relatively bad (e.g., >-15), it typically indicates that ESSENCE-Dock detects low consensus. This suggests that there is substantial disagreement among the individual docking methods, and the results should be interpreted with caution. This is corroborated by the findings for most other datasets where ESSENCE-Dock performs poorly (AMPC, CP3A4, CXCR4, FABP4, and GLCM), which typically showed low ESSENCE-Dock scores (>-15), even for the top-ranked compounds, indicating low confidence in these results. We believe that the ability of ESSENCE-Dock scores to act as a quality measure for the docking results is a valuable feature. Specifically, during virtual screening campaigns using ESSENCE-Dock, a certain score cutoff (e.g., -15 or -20) could be applied. Compounds scoring below the threshold would be rejected, allowing researchers to effectively filter and prioritize compounds, reducing false positives and focusing on candidates with a more robust consensus.

Finally, we briefly discuss the general runtimes of both the individual docking runs and the ESSENCE-Dock calculations themselves. More detailed information for three DUD-E examples MK14, TRY1 and WEE1 are provided in the supplementary data: Fig. S45 for the individual docking times and Fig. S46 for the consensus time. These datasets vary in size, and can thus give a good indication for the approximate runtimes across various scenarios.

In general, the Gnina calculations took the most amount of time. This is also clear from Fig. S45, and can probably be attributed to the fact that these calculations were run using only CPUs, while Gnina's CNN scoring algorithm has been optimized to be run on a GPU. The Gnina docking calculations were assigned 4 CPUs per job, which was able to accelerate the initial generation and optimization of binding poses compared to running it with only 1 CPU. However,

the final step of CNN scoring took up the most amount of time, and seemed to be unchanged regardless of the amount of assigned CPUs. performing the Gnina calculations using GPUs would accelerate the CNN scoring and reduce overall runtime substantially.

Additionally, DiffDock and LeadFinder took about the same amount of wall-clock time to complete in most cases. However, even though their runtime was similar, DiffDock and LeadFinder don't have the same computational cost: DiffDock jobs were assigned 4 CPU cores, while LeadFinder jobs ran using only one. Note that DiffDock was also designed to be run on a GPU, but when it is not, it can still take advantage of multiple CPU cores if they are available. Finally, LeadFinder was developed to be run on CPUs and thus is the most optimized. This translated into the lowest runtimes most of the time. A lot of these general trends can be clearly seen in Fig. S45.

Specific details about the ESSENCE-Dock consensus runtimes are presented in Fig. S46. During this process, the individual docking calculations for every compound are processed and integrated by computing their final ESSENCE-Dock score. The time to generate a PyMol Session file with PLIP scores is not included here. The consensus calculations are performed using a Python script, which also makes use of subprocess calls to obrms and obrotamer to calculate the RMSD and retrieve the number of rotatable bonds. The time spent in subprocess calls is not included in the user time. Instead, it is listed as system time and therefore we include all 3 metrics as returned from the Unix "time" command. ESSENCE-Dock consensus calculations also scale linearly, as is also evident from Fig. S46. This is corroborated by R^2 value of 0.99 for system-time and 1.00 for user- and real-time. The total runtime can get more significant if the datasets are really large, but processing small- to medium-sized libraries is still very manageable, only taking up about 15 minutes for a small dataset (like the WEE1 example; ~6250 compounds) and about 70 minutes for larger ones (like the MK14 example; ~36500 compounds).

From these results, it is clear that our novel ESSENCE-Dock approach is a powerful consensus docking technique that is able to provide substantial enrichment across various families of proteins. On top of that, it also provides helpful output files, including a PSE file of the top results along with their predicted PLIP interactions. The predicted protein-ligand interactions are also summarized in a useful JSON and CSV file.

Conclusions and Outlook

We have presented ESSENCE-Dock, an easy-to-use consensus docking workflow that focuses on providing a high enrichment of active compounds. In its current implementation, it combines three different docking methods (DiffDock, Gnina, and LeadFinder) and considers the predicted binding pose similarities, docking scores, and ligand flexibility. Additionally, the top results can be visualized in an interactive PyMol session with PLIP-predicted protein-ligand interactions. The predicted protein-ligand interactions are also provided in useful output formats for post-processing. ESSENCE-Dock can be run on HPC clusters using Singularity and Slurm, making it feasible to apply the workflow on larger virtual libraries.

Future work could explore the use of new docking methods, as ESSENCE-Dock is a flexible framework, which can easily be made to work with other docking methods. When new, better-performing docking methods are developed, they can be quickly and easily integrated into the workflow. Additionally, it could be interesting to train a machine-learning model with all of this input data. This would allow us to move to an ML-based scoring function, which could be more robust and outperform the current method, leading to even better overall results.

Data and Software Availability

DiffDockHPC (Software for the DiffDock calculations): https://github.com/Jnelen/DiffDockHPC
Metascreener (Software for the Gnina and LeadFinder docking calculations as well as the ESSENCE-Dock consensus workflow): https://github.com/bio-hpc/metascreener
The DUD-E docking data and ESSENCE-Dock consensus results: https://zenodo.org/doi/10.5281/zenodo.10025839

Funding and Acknowledgments

J.N. was funded by Cátedra Villapharma-UCAM. M.C.-B is a predoctoral student funded by Plan Propio de Investigación, UCAM. Supercomputing resources in this work have been supported by the Plataforma Andaluza de Bioinformática of the University of Málaga (https://www.scbi.uma.es/site/), by the supercomputing infrastructure of the NLHPC (ECM-02, Powered@NLHPC), and by the Extremadura Research Centre for Advanced Technologies (CETA−CIEMAT), funded by the European Regional Development Fund (ERDF). CETA−CIEMAT belongs to CIEMAT and the Government of Spain.

References

1. Schlander M, Hernandez-Villafuerte K, Cheng CY, Mestre-Ferrandiz J, Baumann M. How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment. PharmacoEconomics. 2021 Nov;39(11):1243–69.
2. Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Computational Methods in Drug Discovery. Pharmacol Rev. 2014 Jan;66(1):334–95.
3. Gloriam DE. Bigger is better in virtual drug screens. Nature. 2019 Feb;566(7743):193–4.
4. Pinzi L, Rastelli G. Molecular Docking: Shifting Paradigms in Drug Discovery. Int J Mol Sci. 2019 Sep 4;20(18):4331.
5. Blanes-Mira C, Fernández-Aguado P, de Andrés-López J, Fernández-Carvajal A, Ferrer-Montiel A, Fernández-Ballester G. Comprehensive Survey of Consensus Docking for High-Throughput Virtual Screening. Molecules. 2022 Dec 25;28(1):175.
6. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–61.
7. Koes DR, Baumgartner MP, Camacho CJ. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. J Chem Inf Model. 2013 Aug

26;53(8):1893–904.
8. McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, et al. GNINA 1.0: molecular docking with deep learning. J Cheminformatics. 2021 Jun 9;13(1):43.
9. Stroganov OV, Novikov FN, Stroylov VS, Kulkov V, Chilov GG. Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening. J Chem Inf Model. 2008 Dec;48(12):2371–85.
10. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J Med Chem. 2004 Mar 1;47(7):1739–49.
11. Torres PHM, Sodero ACR, Jofily P, Silva-Jr FP. Key Topics in Molecular Docking for Drug Design. Int J Mol Sci. 2019 Sep 15;20(18):4574.
12. Tuccinardi T, Poli G, Romboli V, Giordano A, Martinelli A. Extensive Consensus Docking Evaluation for Ligand Pose Prediction and Virtual Screening Studies. J Chem Inf Model. 2014 Oct 27;54(10):2980–6.
13. Kukol A. Consensus virtual screening approaches to predict protein ligands. Eur J Med Chem. 2011 Sep;46(9):4661–4.
14. Huang N, Shoichet BK, Irwin JJ. Benchmarking Sets for Molecular Docking. J Med Chem. 2006 Nov 1;49(23):6789–801.
15. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009 Dec;30(16):2785–91.
16. Houston DR, Walkinshaw MD. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. J Chem Inf Model. 2013 Feb 25;53(2):384–90.
17. Ochoa R, Palacio-Rodriguez K, Clemente CM, Adler NS. dockECR: Open consensus docking and ranking protocol for virtual screening of small molecules. J Mol Graph Model. 2021 Dec 1;109:108023.
18. Morris CJ, Stern JA, Stark B, Christopherson M, Della Corte D. MILCDock: Machine Learning Enhanced Consensus Docking for Virtual Screening in Drug Discovery. J Chem Inf Model. 2022 Nov 7;acs.jcim.2c00705.
19. Palacio-Rodríguez K, Lans I, Cavasotto CN, Cossio P. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. Sci Rep. 2019 Mar 26;9(1):5142.
20. Zhang N, Zhao H. Enriching screening libraries with bioactive fragment space. Bioorg Med Chem Lett. 2016 Aug 1;26(15):3594–7.
21. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, et al. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. PLOS Comput Biol. 2014 Apr 10;10(4):e1003571.
22. Mysinger MM, Carchia M, Irwin JohnJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. J Med Chem. 2012 Jul 26;55(14):6582–94.
23. Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking [Internet]. arXiv; 2022 [cited 2023 Feb 1]. Available from: http://arxiv.org/abs/2210.01776
24. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.
25. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein–ligand interaction profiler. Nucleic Acids Res. 2015 Jul 7;43(Web Server issue):W443.
26. RDKit: Open-source cheminformatics. [Internet]. RDKit; 2023 [cited 2023 Jul 25]. Available from: https://github.com/rdkit/rdkit
27. Molconvert | Chemaxon Docs [Internet]. [cited 2023 Jul 25]. Available from: https://docs.chemaxon.com/display/docs/molconvert.md
28. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and

performance of MMFF94. J Comput Chem. 1996;17(5–6):490–519.
29. Schrödinger Release 2020-4: Maestro, Schrödinger, LLC, New York, NY, 2020.
30. Roos K, Wu C, Damm W, Reboul M, Stevenson JM, Lu C, et al. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. J Chem Theory Comput. 2019 Mar 12;15(3):1863–74.
31. Novikov FN, Stroylov VS, Zeifman AA, Stroganov OV, Kulkov V, Chilov GG. Lead Finder docking and virtual screening evaluation with Astex and DUD test sets. J Comput Aided Mol Des. 2012 Jun 1;26(6):725–35.
32. bio-hpc/metascreener: Collection of scripts that integrates docking, virtual screening, similarity and molecular modeling programs. [Internet]. [cited 2023 Dec 5]. Available from: https://github.com/bio-hpc/metascreener/tree/main
33. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminformatics. 2011 Oct 7;3(1):33.
34. Krüger DM, Evers A. Comparison of Structure‐ and Ligand‐Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. ChemMedChem. 2010 Jan 4;5(1):148–58.

Abbreviations

ADT: AutoDock Tools
CNN: Convolutional Neural Network
Cons.: Consensus
CPU: Central Processing Unit
CSV: Comma Separated Values (file)
DD: DiffDock
DUD(-E): Directory of Useful Decoys(-Enhanced)
ECR: Exponential Consensus Ranking
EF: Enrichment Factor
GN: Gnina
GPU: Graphics Processing Unit
HTS: High-throughput Screening
HPC: High-Performance Computing
LF: LeadFinder
MBE: Mean Binding Energy
ML: Machine Learning
MR: Mean Rank
PDB: Protein Data Bank
PLIP: Protein-Ligand Interaction Profiler
PSE: PyMol Session (file)
RAM: Random Access Memory
RB: Rotatable Bonds
RMSD: Root Mean Square Deviation
VS: Virtual Screening

For Table of Contents Use Only