Article

Not peer-reviewed version

High-Throughput, High-Quality: Unlocking GNINA for Precision Virtual Screening

Rocco Buccheri and Antonio Rescifina *

Posted Date: 5 June 2025

doi: 10.20944/preprints202506.0372.v1

Keywords: HTVS; molecular docking; virtual screening; AutoDock; Vina; GNINA

 doi:10.20944/preprints202506.0372.v1

Article

High-Throughput, High-Quality: Unlocking GNINA for Precision Virtual Screening

Rocco Buccheri and Antonio Rescifina *

Department of Drug and Health Sciences, University of Catania, Viale A. Doria 6, 95125 Catania, Italy; rocco.buccheri@unict.it (R.B.)

* Correspondence: antonio.rescifina@unict.it

Abstract: Drug discovery is an intricate and resource-intensive process in which computational approaches such as molecular docking are essential, particularly in the early stages, to identify possible hits. However, docking still has many drawbacks, including problems in managing protein flexibility and the reliability of scoring functions. In this paper, we systematically compared the performance of AutoDock Vina, one of the most widely used open-source docking tools, with GNINA. This advanced evolution integrates convolutional neural networks (CNNs) for pose scoring. The comparison was conducted on ten heterogeneous protein targets, including metalloenzymes, kinases, and G-protein-coupled receptors (GPCRs). With the ability to accurately replicate binding poses and their energy values, GNINA showed outstanding performance in both virtual screening (VS) of active ligands and re-docking steps of co-crystallized ligands. GNINA's enhanced ability to accurately distinguish between true positives and false positives – a specificity not found with AutoDock Vina – is confirmed by ROC curves and Enrichment Factor (EF) results. Therefore, we propose an integrated GNINA-based workflow that can significantly enhance the quality and reliability of docking results, providing a valuable tool for optimizing the initial stages of drug discovery.

Keywords: HTVS; molecular docking; virtual screening; AutoDock; Vina; GNINA

1. Introduction

Drug discovery is a complex and expensive process aimed at identifying and developing new drugs, which has seen an increasing integration of computational methodologies in recent decades. These methodologies have become crucial components in many phases of drug discovery programs, from identifying "hits" to optimizing "leads" [1]. Among these, molecular docking has emerged as a fundamental and widely used computational tool [2–4].

Currently, molecular docking is an established and popular *in-silico* tool for identifying new compounds with therapeutic potential. It is primarily used in the early stages of drug discovery due to its cost-effectiveness and potential to identify novel chemotypes, as well as provide mechanistic insights into ligand-protein interactions [2,3]. With the rapid improvement of computational platforms and the substantial increase in structural, chemical, and biological data available for an expanding number of therapeutic targets, in-silico approaches, such as docking, have undergone significant growth. Molecular docking has become a crucial component of computational procedures employed in modern drug discovery. Structure-based virtual screening techniques, which often employ docking, enable the screening of digital libraries comprising millions of compounds within affordable timescales, thereby reducing the initial costs of hit identification and increasing the likelihood of discovering promising drug candidates. Automated workflows integrating docking have been developed for screening large libraries of compounds and targets. The advancement of high-performance computing and Graphics Processing Units (GPUs) has enabled large-scale screening to become possible [2,3,5]. Although initially used as a stand-alone method, docking is now predominantly integrated into workflows that combine different computational methodologies, such

 doi:10.20944/preprints202506.0372.v1

as ligand-based approaches, molecular dynamics (MD), binding free energy calculations, and artificial intelligence (AI). This integration aims to overcome some of the inherent limitations of docking and better leverage information from various sources, typically resulting in improved predictive performance in terms of hit rates [2].

Despite its widespread use and considerable progress, molecular docking has inherent limitations that limit its predictive accuracy. The two most relevant and widely recognized limitations of molecular docking are the sampling of conformations and the accuracy of scoring functions. Docking involves finding the most favorable reciprocal configurations, known as "poses," between a ligand and a protein target. However, sampling all possible conformations of the ligand and receptor is often restricted. Additionally, the target protein is frequently treated as a rigid body during docking for computational reasons, which is a significant limitation. Proteins are dynamic systems with some flexibility, which can be affected by ligand binding, known as "induced fit" effects. Approaches to address protein flexibility, such as utilizing multiple protein structures or integrating with molecular dynamics simulations, are more complex and require greater computational resources [2,3]. The scoring function is a component of docking that quantifies the interaction between the ligand and the target protein, evaluating the goodness-of-fit of a given pose or ranking different ligands. To achieve the computational efficiency required for screening many compounds, scoring functions are often simplified and based on approximate models of protein-ligand interactions. This simplified modeling is a substantial limitation, as approximate scoring functions very rarely correlate well with experimental binding affinities. Imperfections in the scoring functions remain a significant limiting factor in docking. They lead to inaccuracies in the ranking of predicted poses and poor performance in predicting binding free energy. The need to improve predictive accuracy has led to the development of several strategies to mitigate the limitations of individual scoring functions, such as consensus scoring, which combines the results of several scoring functions, and rescoring methods based on different techniques [1–4].

In the context of rational drug development, molecular docking represents one of the benchmark computational strategies, particularly in the early stages of virtual screening (VS). The use of *in-silico* screening techniques on libraries containing thousands of compounds enables the identification of potential ligands active toward a protein target, resulting in significant time and cost savings compared to traditional *in vitro* biological assays, which would be impractical on such a large scale. Optimizing computational workflows to ensure a high hit rate as early as the preliminary stage of VS is a crucial area of importance. Improving the quality of results in these early stages can, in fact, significantly increase the probability of success in the later stages of pharmaceutical development.

AutoDock Vina (often abbreviated just as Vina) is the most widely used and most integrated molecular docking algorithm within molecular docking software. Trott and Olson developed Vina at the Scripps Research Institute in California as an alternative to its predecessor, AutoDock 4. Key features that have contributed to Vina's widespread adoption include its speed, improved accuracy over AutoDock 4, and free availability under an open-source license. Its high computational efficiency and ease of use led to very rapid uptake in the docking community, well evidenced by the high number of citations in the original article. AutoDock Vina is considered one of the fastest and most widely used open-source programs for molecular docking [6,7]. Despite its undoubted qualities and broad adoption, Vina has limitations. It has difficulties with increasingly flexible ligands, and its empirical scoring function may have a size-related bias [8,9]. GNINA represents a significant evolution in the field of molecular docking, building on Vina and its fork, Smina. The main innovation of GNINA is the integration of convolutional neural networks (CNNs) for scoring protein-ligand poses. In its typical workflow, GNINA samples ligand poses using Markov Chain Monte Carlo (MCMC) sampling, initially driven by AutoDock Vina's empirical scoring function. Subsequently, GNINA uses the CNN-based scoring function to "rescore" and rank the poses obtained from MCMC sampling. Unlike empirical or knowledge-based scoring functions, which often assume a linear relationship between structural features and binding affinity, CNNs can model nonlinear relationships and potentially interpret molecular interactions in a more sophisticated way [7,10–12].

 doi:10.20944/preprints202506.0372.v1

2. Results and Discussion

2.1. Energy values and conversions

Evaluations of the predicted energy from molecular docking were treated differently for the two algorithms analyzed because they yield different outputs. As for Vina, this is reflected in the output values of predicted energy, expressed in terms of the free energy of binding (ΔG), which is calculated in kcal/mol. The ΔG value was then converted to the corresponding binding affinity constant (K) value using the thermodynamic relationship reported in Equation (1).

$$K = e^{(\Delta G/RT)} \quad (1)$$

Equation (1) was solved at a temperature of 310 K, with a gas constant R of 0.0019872036 kcal/(K × mol). Subsequently, the K value was transformed into its negative logarithm, pK.

GNINA returns several output metrics related to CNN assessment, namely: CNN score, CNN affinity, and CNN_VS. The CNN score is an assessment of the goodness of the generated pose, ranging from 0 to 1, where a score of 1 indicates a pose of higher reliability. The CNN affinity is an expected bond affinity and is expressed in pK. Finally, the CNN_VS is the product of the CNN score and CNN affinity and is used to rank compounds in virtual screening, having been found effective in retrospective evaluations [7,10–12].

2.2. Target Validation

The protein target validation phase involved identifying and selecting the most suitable protein model for use in the subsequent VS steps. Only protein models that met the following criteria were considered: (i) presence of a co-crystallized ligand; (ii) availability of an experimental binding affinity value (K_i or K_d); and (iii) crystallographic resolution of less than 3 Å to ensure an adequate level of structural detail. For each selected model, the docking algorithm's ability to faithfully reproduce the pose of the co-crystallized ligand was evaluated by calculating the Root Mean Square Deviation (RMSD), as well as its accuracy in estimating the binding affinity by comparing the predicted inhibition constant with the experimentally determined value. In cases where the model had multiple co-crystallized domains, the analysis was conducted on each domain individually, and the one that showed the best performance in terms of structural and energetic accuracy was finally selected. Preliminary studies were performed using GNINA to select the PDB file and ultimately the domain to be used. The choice of GNINA at the PDB file selection stage was based on the fact that the CNN score calculated by the algorithm allows a distinction between receptors with high-quality binding sites and those with poor-quality binding sites. Therefore, receptors with the highest possible CNN score, but never below the threshold of 0.90 recommended in the literature, were selected. In the case of multiple domains present, the protein domain with the highest CNN score was selected [13]. The domains chosen and the CNN scores calculated for each target are shown in Table 1.

Table 1. Selected models of protein targets.

Protein name	Source database	PDB ID	Domain	CNN score
Acetylcholinesterase	PDB-REDO	6O4W	B	0.92
Tyrosine-protein kinase ABL2	RCSB PDB	2XYN	C	0.99
Carbonic anhydrase II	PDB-REDO	4HT2	A	0.90
SYK kinase	PDB-REDO	3EMG	A	0.99
Beta-secretase 1	PDB-REDO	4DJW	B	0.98
Cyclin-dependent kinase 2	RCSB PDB	1KE9	A	0.97
Adenosine A2a receptor	PDB-REDO	5OLH	A	0.96
Dopamine D3 receptor	RCSB PDB	7BVQ	B	0.98
HSP90α	PDB-REDO	4O09	A	0.96
HDAC 6	PDB-REDO	5EDU	B	0.97

 doi:10.20944/preprints202506.0372.v1

All selected protein targets were subjected to docking simulations using both Vina and GNINA to compare their performance directly. The results, presented in Table 2, demonstrate a significant performance improvement achieved with GNINA compared to Vina, in both structural and predictive accuracy.

Table 2. Results derived from the VS performed with GNINA and Vina.

Protein name	pK_{exp}	GNINA pK_{pred}	GNINA RMSD	Vina pK_{pred}	Vina RMSD
Acetylcholinesterase	8.54–7.42	7.29	1.71	7.61	1.19
Tyrosine-protein kinase ABL2	7.52–7.38	8.47	0.79	5.99	6.54
Carbonic anhydrase II	6.82–6.54	7.73	1.37	4.57	6.78
SYK kinase	8.05	7.85	0.97	6.29	1.04
Beta-secretase 1	6.96	6.83	0.44	5.95	7.31
Cyclin-dependent kinase 2	6.68–5.75	6.52	1.80	6.52	1.96
Adenosine A2a receptor	9.10–8.89	7.60	0.29	6.04	8.33
Dopamine D3 receptor	10.00–9.80	7.56	1.69	4.74	6.23
HSP90α	7.70–7.59	7.75	1.05	8.19	0.95
HDAC 6	9.89–6.30	70.00	1.80	4.79	7.30

The pK values predicted by Vina are derived from the conversion of the ΔG calculated by Vina, and those predicted by GNINA are nothing but the CNN_VS values calculated by GNINA.

GNINA produced RMSD values consistently below 2 Å for the predicted pose, indicating an excellent ability to reproduce the conformation of the co-crystallized ligand [14]. In contrast, Vina showed greater variability: in only 4 out of 10 targets, the RMSD value was less than 2 Å, while in the remaining cases it exceeded 3 Å, showing limited reliability in reproducing the co-crystallized pose. The variability in the redocking of Vina and the stability of GNINA can be observed in the two representative examples shown in Figure 1; the remaining comparisons are presented in Figures S1 and S2.

 doi:10.20944/preprints202506.0372.v1

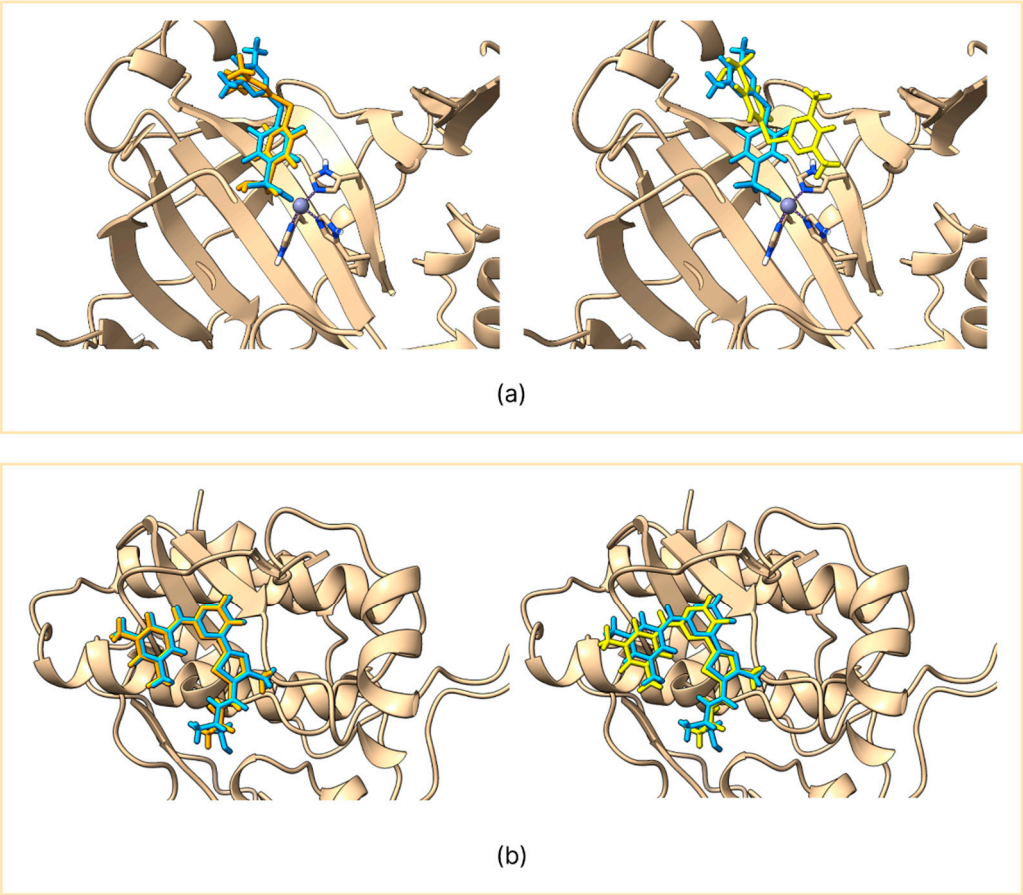

Figure 1. Comparison between the co-crystallized ligand (light blue) and the reproduced pose using GNINA (orange) on the left panel and AutoDock Vina (yellow) on the right panel. The figure shows carbonic anhydrase II (a) and SYK kinase (b).

Similarly, prediction of binding affinity (expressed as p*K*) was also more accurate with GNINA in all targets examined, except in the case of Cyclin-dependent kinase 2, for which both algorithms returned overlapping values.

2.3. Virtual Screening Analysis

VS analyses were conducted using sets of ligands for which experimental affinity values (K_i or K_d) were available. The criterion adopted for selecting candidate molecules ("hits") was based on the p*K* value estimated by the docking algorithm. Specifically, a selective filter was applied, considering only molecules with a pK ≥6.3, a threshold that corresponds to a binding affinity of approximately 500 nM, a value commonly accepted in literature as indicative of good biological activity. At the end of the filtering process, the predictive validity of the system was assessed by comparing the docking results with available experimental data. All molecules with experimental K_i/K_d values less than 1000 nM were considered effectively active. Based on this comparison, the percentage success rate of virtual screening was calculated according to the formula reported in Equation (2).

$$\text{success rate (\%)} = \frac{\text{number of experimentally active hits}}{\text{total number of hits with p}K \geq 6.3} \times 100 \qquad (2)$$

This approach enables a quantitative evaluation of the algorithm's ability to identify truly active molecules, providing an objective measure of the effectiveness of the screening protocol employed. As in the validation phase, the expected activity value considered in the VS analyses was CNN_VS for the GNINA algorithm and the p*K* value derived from the Vina Δ*G* transformation.

 doi:10.20944/preprints202506.0372.v1

The data shown in Table 3 indicate that, in most cases, the use of GNINA resulted in a higher success rate percentage (SR%) compared with Vina. For 8 out of 10 targets analyzed, GNINA showed a greater ability to identify experimentally active ligands from the screening sets. A particularly significant example is Carbonic anhydrase II, for which GNINA identified 149 hits, 146 (98%) of which were active, compared with only two hits obtained from Vina, of which only one was active.

Table 3. VS results report.

Protein name	Input molecules [a]	Hits [b] (GNINA)	Actives [c] (GNINA)	SR% [d] GNINA	Hits [b] (Vina)	Actives [c] (Vina)	SR% [d] Vina
Acetylcholinesterase	790	148	90	64%	532	233	44%
Tyrosine-protein kinase ABL2	92	20	11	55%	41	16	39%
Carbonic anhydrase II	502	149	146	98%	2	1	50%
SYK kinase	80	41	34	83%	13	11	85%
Beta-secretase 1	2345	660	574	87%	150	122	81%
Cyclin-dependent kinase 2	1322	726	655	90%	386	285	74%
Adenosine A2a receptor	7001	1630	1426	88%	2839	2129	75%
Dopamine D3 receptor	150	37	36	97%	39	38	97%
HSP90α	637	190	125	63%	299	181	61%
HDAC 6	225	36	34	94%	21	18	86%

[a] The number of molecules contained in the ligand sets with known activity. [b] The number of hits obtained for both algorithms by following the above-mentioned filtering criterion. [c] The number of experimentally active ligands present in the hits. [d] The percentage of the success rate values.

However, the difference was not marked in all cases. For the SYK kinase target, for example, Vina achieved a hit rate of 85%, slightly higher than that of GNINA (83%). Furthermore, for the Dopamine D3 receptor, both algorithms achieved essentially equivalent results, with a high predictive accuracy of 97%.

2.4. ROC Curves And Enrichment Factors Analysis

To statistically evaluate the ability of the two docking methods compared and obtain quantitative estimates of their sensitivity in distinguishing between active and inactive compounds, Receiver Operating Characteristic (ROC) curves with their Area Under the Curve (AUC) and Enrichment Factors (EF) were calculated.

ROC curves are an established graphical and analytical tool. In the context of virtual screening, an ROC curve is constructed by representing the true positive fraction (TPF) on the y-axis versus the false positive fraction (FPF) on the x-axis. Each point on the curve represents a TPF/FPF pair corresponding to a specific fraction of the molecular dataset classified according to the scoring function scores. A scoring function that can perfectly discriminate between active and inactive compounds (without overlapping score distributions) would have an ROC curve that passes through the upper left corner of the graph (TPF = 1, FPF = 0), indicating perfect sensitivity and specificity. In contrast, a scoring function that has no discriminatory power would produce a curve coincident with the 45° diagonal, corresponding to random selection (AUC = 0.5). The Area Under the ROC Curve (AUC) is a scalar value that summarizes the overall performance of a virtual screening. An AUC value closer to 1 indicates a high discrimination ability of the scoring function in distinguishing between active and inactive compounds over the entire data set. AUC > 0.5 suggests that the

 doi:10.20944/preprints202506.0372.v1

prediction ability is better than a random distribution model. Thus, AUC corresponds to the probability of correctly classifying a random pair of active ligand and decoy [15–19].

The Enrichment Factor (EF) is used to evaluate the ability of a scoring function or virtual screening method to enrich a subset of selected molecules (typically those with the best scores) in active compounds compared to a random selection from the entire data set. In practical terms, EF at a given percentage X% indicates how many times more active compounds are recovered in the top X% of the library than would be expected from a random selection of the same percentage of compounds. Performance evaluation often focuses on EF at low percentages of the ranked database, since the goal of virtual screening is to quickly identify a small fraction of promising compounds for experimental testing. For example, an EF at 1% (EF1%) equal to 10 means that 10 times more active compounds are found in the top 1% of compounds ranked by the scoring function than in a random selection of 1% of compounds from the entire data set [16–20].

Analysis of the ROC curves (Figure 2) clearly shows that GNINA outperforms Vina for all targets considered, except for the dopamine D3 receptor, for which the two methods exhibit overlapping performance. It is particularly relevant to note that GNINA achieved AUC values above 0.70 in almost all cases, never approaching the random classification threshold (AUC ≈ 0.50). In contrast, Vina exceeded the threshold of AUC > 0.5 in only six protein targets.

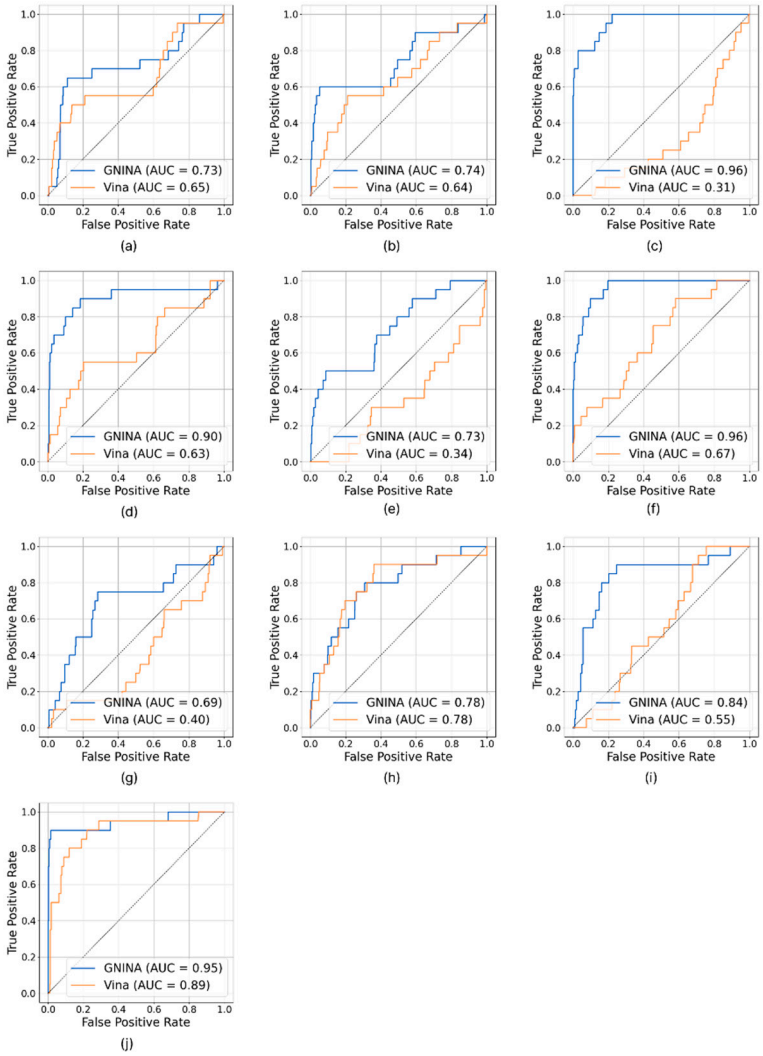

Figure 2. ROC curves and calculated AUC of acetylcholinesterase (a), tyrosine-protein kinase ABL2 (b), carbonic anhydrase II (c), SYK kinase (d), beta-secretase 1 (e), cyclin-dependent kinase 2 (f), adenosine A2a receptor (g), dopamine D3 receptor (h), HSP90α (i), and HDAC 6 (j).

 doi:10.20944/preprints202506.0372.v1

Analysis of EFs (Table 4 and Figure S3), assessed at the thresholds of 1%, 5%, and 10%, also confirmed the superiority of GNINA over Vina. Out of a total of 30 comparisons, GNINA showed better performance in 25 cases, and the two algorithms showed the same performance in 4 cases. Vina performed better than GNINA in only one analysis. Only in three circumstances – SYK kinase at EF1%, HSP90α at EF1% and dopamine D3 receptor at EF10% – does the performance of the two tools appear equivalent.

Table 4. EF analysis results report.

Protein name	GNINA EF1% [a]	Vina EF1% [a]	GNINA EF5% [b]	Vina EF5% [b]	GNINA EF10% [c]	Vina EF10% [c]
Acetylcholinesterase	5.52	5.52	1	6.02	5.52	4.01
Tyrosine-protein kinase ABL2	15	5	10	3	6	2.5
Carbonic anhydrase II	20.75	0	14.53	0	8.1	0
SYK kinase	10	10	12	3	7	3
Beta-secretase 1	12.53	0	7.31	0	4.57	0
Cyclin-dependent kinase 2	20.5	10.25	12.3	4.1	8	3
Adenosine A2a receptor	5.08	0	3.04	2.03	2.54	1.01
Dopamine D3 receptor	15.34	10.22	6.13	3.07	3.58	3.58
HSP90α	0	0	4.01	0	5.51	0
HDAC 6	20.15	0	17.13	10.07	9.07	6.04

[a] EF calculated at 1%. [b] EF calculated at 5%. [c] EF calculated at 10%.

3. Materials and Methods

3.1. Protein Preparation

Protein structure selection was done by evaluating the RCSB PDB Databank (https://www.rcsb.org/) and the PDB-REDO Databank (https://pdb-redo.eu/) structure quality, and the best one was chosen considering crystallographic refinement and model quality parameters shown in the PDB-REDO Databank. The organism chosen was Homo sapiens, and the resolution was always less than 3 Å. Each protein model selected (Table 1) was prepared in YASARA software (v. 25.1.13, YASARA Biosciences GmbH, Vienna, Austria).

Hydrogens were added through '*Clean* > *All*' YASARA's option, waters and other non-necessary molecules for docking, and the co-crystallized ligand was removed using YASARA software. The protein was saved in PDB file format.

3.2. Dataset and 3D Structures Generation

Ligand datasets were downloaded from Binding Database (https://www.bindingdb.org/) using the protein name as a research query. Data were filtered by selecting only Homo sapiens as the target source organism, and ligands with known K_i or K_d values were chosen. Duplicate molecules were deleted, merging equivalent rows in DataWarrior (powered by openchemlib, v. 06.00.00) [21] using SMILES code as reference. Salts were also deleted using the largest fragment selection option in DataWarrior. Ligands' 3D structures were generated and minimized in Open Babel (v. 3.1.1) [22], reflecting physiological pH (7.4) states. Then, 3D structures were optimized at a semi-empirical level using the xTB (extended tight-binding, v. 6.7.1) program package [23]. Geometry optimization was performed in xTB using the Analytical Linearized Poisson-Boltzmann (ALPB) model for water, with charge states specified for each molecule according to physiological pH conditions.

 doi:10.20944/preprints202506.0372.v1

3.3. Molecular Docking Analysis

Molecular docking analysis was performed using both the molecular docking program GNINA (v. 1.3) [11] and the AutoDock Vina (v. 1.2.5) [24] docking algorithm. Regarding GNINA, we utilized the "rescore" docking mode setting, which involved 15 ligand pose rotations, and the poses were sorted by CNN score. The protein input format was in PDB file format, and the ligand input format was in SDF file format. AutoDock Vina docking analysis was performed by setting the exhaustiveness parameters to 32 and converting the input files to the PDBQT file format required by Vina.

In both cases, the simulation boxes were built using AutoDock Tools (v. 1.5.7), with the co-crystallized ligand pose serving as the reference. Grid parameters – shown in SI – were chosen wide enough not to force the ligand-receptor interaction.

3.4. RMSD calculation

RMSD calculation was performed using the tool DockRMSD (v. 1.0.0) [25]. The co-crystallized ligand was converted into a MOL2 file format through Open Babel, as required by the tool. The docking output poses were merged into a single pose, and the first one was converted into a MOL2 file format. The co-crystallized ligand pose and the first docked one were used as input to calculate RMSD.

3.5. Decoys Generation

Decoys were prepared via the LUDe web server [26] from the 20 most active ligands for each target, which constituted the true positives. Of the decoys generated by LUDe, 400 were selected for each target, which constituted the false positives. Decoys are decoy molecules constructed from ligands with known experimental activity toward the protein target, having similar physicochemical properties but different 2D topology. Decoys have been prepared as reported in paragraph *3.2*.

3.6. ROC Curves and Enrichment Factors

ROC curves and related AUC values were calculated using the roc_curve() and auc() functions of the sklearn.metrics module (scikit-learn v. 1.3.0). A set of 420 compounds was prepared for each protein target, including the 20 most active ligands according to experimental affinity values and 400 decoy molecules. The active ligands represented True Positives, while the 400 decoys represented False Positives. False Positive Rate (FPR) and True Positive Rate (TPR) values were obtained by comparing the binary labels with the predictive scores of each docking method. The AUC was subsequently calculated by trapezoidal numerical integration of the ROC curve. ROC curves were generated using matplotlib (v. 3.7.1), plotting TPR vs FPR for each method with indication of the respective AUC values in the legends. The diagonal reference line (random classifier) was included for visual comparison of performance.

EFs were calculated at the thresholds of 1%, 5%, and 10% of the total dataset. For each predictive method, the dataset was sorted in descending order of score. The number of active compounds in the top-ranked subset was divided by the expected number of active compounds in a random selection of the same size, according to the formula reported in Equation (3).

$$\mathrm{EF} = \frac{(\mathrm{N\_active\_top}/\mathrm{N\_top})}{(\mathrm{N\_active\_totals}/\mathrm{N\_totals})} \quad (3)$$

where N_active_top represents the number of active compounds in the selected subset, N_top the size of the subset, N_active_totals the total number of active compounds in the dataset, and N_totals the total size of the dataset. Figure S3 was generated with the online Screening Explorer Tool (http://stats.drugdesign.fr/, accessed 05/31/2025) [27].

4. Conclusions

 doi:10.20944/preprints202506.0372.v1

The present study compared the performance of Vina with that of GNINA, which integrates a CNN-based system used for rescoring poses obtained from molecular docking. Vina was chosen as the comparison algorithm because it is widely used and integrated into standard molecular docking software. Therefore, the primary application we aimed to evaluate is a retrospective analysis of VS. We sought to determine whether the ability of CNNs to interpret molecular interactions provides advantages in the search for active compounds when performing VS of large libraries of molecules.

GNINA showed higher accuracy in reproducing the co-crystallized ligand pose and predicting protein binding affinity compared to Vina. RMSD values consistently less than 2 Å were recorded in all analyses performed with GNINA, while Vina showed greater variability with RMSD values often exceeding 3 Å. In predicting binding affinity, GNINA also returned more accurate values than the experimental values reported in the literature.

In VS analyses, GNINA demonstrated a greater ability to identify experimentally active ligands than Vina, achieving excellent success rates in most cases that were superior to those of Vina. A noteworthy result is that of Carbonic anhydrase II, where GNINA recorded several hits almost exclusively populated by experimentally active ligands. To reinforce these observations, more accurate statistical investigations were conducted by generating a decoy library for each protein target. ROC curves were generated, and respective AUC values – an indicator of the algorithm's ability to discriminate between active and inactive compounds – were calculated. For GNINA, consistently high AUC values were recorded (AUC >0.70) and far from the threshold of AUC ≈ 0.50, which would indicate random classification. Vina results were less homogeneous, with values often below the AUC threshold. EF at different rates of 1%, 5%, and 10% was also evaluated to assess the effectiveness of the methods in recovering active compounds from the library fractions. Again, there were results in favor of GNINA, which consistently outperformed Vina.

In summary, we can state that the use of GNINA CNNs in the rescoring phase of the poses leads to benefits in performance in all phases of VS: from pose reproducibility to prediction of binding affinity and discriminatory ability against active compounds. Therefore, in future computational drug searches, we recommend adopting the workflow for VS discussed in this study, which includes a rigorous target validation phase and the use of the GNINA algorithm for VS of compounds, to maximize its effectiveness and significantly increase the chance of success in identifying promising compounds. The integration of AI-based algorithms allows overcoming the limitations of traditionally employed tools, offering new perspectives for *in-silico* approaches used in drug discovery.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Figure S1: Comparison between the co-crystallized ligand (light blue) and the reproduced pose using GNINA (orange) on the left panel and AutoDock Vina (yellow) on the right panel. The figure shows ace-tylcholinesterase (a), tyrosine-protein kinase ABL2 (b), beta-secretase 1 (c), and cyclin-dependent kinase 2 (d); Figure S2: Comparison between the co-crystallized ligand (light blue) and the reproduced pose using GNINA (orange) on the left panel and AutoDock Vina (yellow) on the right panel. The figure shows adenosine A2a receptor (a), dopamine D3 receptor (b), HSP90α (c), and HDAC6 (d); Figure S3: Enrichment curves of acetylcholinesterase (a), tyrosine-protein kinase ABL2 (b), car-bonic anhydrase II (c), SYK kinase (d), beta-secretase 1 (e), cyclin-dependent ki-nase 2 (f), aden-osine A2a receptor (g), dopamine D3 receptor (h), HSP90α (i), and HDAC6 (j).

Funding: This research was funded by (i) the Italian Ministry of Health, Piano di Sviluppo e Coesione del Ministero della Salute 2014–2020, Project: Pharma-HUB—Hub per il riposizionamento di farmaci nelle malattie rare del sistema nervoso in età pediatrica (CUP E63C22001680001—ID T4-AN-04), and (ii) Programma di ricerca CN00000013 "National Centre for HPC, Big Data and Quantum Computing", finanziato dal Decreto Direttoriale di concessione del finanziamento n.1031 del 17.06.2022 a valere sulle risorse del PNRR MUR—M4C2—Investimento 1.4—Avviso "Centri Nazionali"—D.D. n. 3138 del 16 dicembre 2021.

Data Availability Statement: The original data presented in the study are openly available in the HTHQ-GNINA GitHub repository at https://github.com/rocco-b/HTHQ-GNINA.

 doi:10.20944/preprints202506.0372.v1

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments). Where GenAI has been used for purposes such as generating text, data, or graphics, or for study design, data collection, analysis, or interpretation of data, please add "During the preparation of this manuscript/study, the author(s) used [tool name, version information] for the purposes of [description of use]. The authors have reviewed and edited the output and take full responsibility for the content of this publication."

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

%SR	Success Rate Percentage
AI	Artificial Intelligence
ALPB	Analytical Linearized Poisson-Boltzmann
AUC	Area Under Curve
CNNs	Convolutional Neural Networks
EF	Enrichment Factor
FPF	False Positive Fraction
FPR	False Positive Rate
GPUs	Graphics Processing Units
MCMC	Markov Chain Monte Carlo
MD	Molecular Dynamics
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic Curve
TPF	True Positive Fraction
TPR	True Positive Rate
VS	Virtual Screening

References

1. Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **2004**, 3, 935–949. https://doi.org/10.1038/nrd1549.
2. Pinzi, L.; Rastelli, G., Molecular Docking: Shifting Paradigms in Drug Discovery. *IJMS* **2019**, 20, 4331. https://doi.org/10.3390/ijms20184331.
3. Fischer, A.; Smieško, M.; Sellner, M.; Lill, M.A., Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *J. Med. Chem.* **2021**, 64, 2489–2500. https://doi.org/10.1021/acs.jmedchem.0c02227.
4. Ferreira, L.; Dos Santos, R.; Oliva, G.; Andricopulo, A., Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, 20, 13384–13421. https://doi.org/10.3390/molecules200713384.
5. De Ruyck, J.; Brysbaert, G.; Blossey, R.; Lensink, M., Molecular docking as a popular tool in drug design, an in silico travel. *AABC* **2016**, Volume 9, 1–11. https://doi.org/10.2147/AABC.S105289.
6. Vieira, T.F.; Sousa, S.F., Comparing AutoDock and Vina in Ligand/Decoy Discrimination for Virtual Screening. *App. Sci.* **2019**, 9, 4538. https://doi.org/10.3390/app9214538.
7. Sunseri, J.; Koes, D.R., Virtual Screening with Gnina 1.0. *Molecules* **2021**, 26, 7369. https://doi.org/10.3390/molecules26237369.
8. Rentzsch, R.; Renard, B.Y., Docking small peptides remains a great challenge: an assessment using AutoDock Vina. *Brief. Bioinform.* **2015**, 16, 1045–1056. https://doi.org/10.1093/bib/bbv008.
9. Chang, M.W.; Ayeni, C.; Breuer, S.; Torbett, B.E., Virtual Screening for HIV Protease Inhibitors: A Comparison of AutoDock 4 and Vina. *PLoS ONE* **2010**, 5, e11955. https://doi.org/10.1371/journal.pone.0011955.

 doi:10.20944/preprints202506.0372.v1

10. McNutt, A.T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D.R., GNINA 1.0: molecular docking with deep learning. *J. Cheminform.* **2021**, 13, 43. https://doi.org/10.1186/s13321-021-00522-2.
11. McNutt, A.T.; Li, Y.; Meli, R.; Aggarwal, R.; Koes, D.R., GNINA 1.3: the next increment in molecular docking with deep learning. *J. Cheminform.* **2025**, 17, 28. https://doi.org/10.1186/s13321-025-00973-x.
12. Dunn, I.; Pirhadi, S.; Wang, Y.; Ravindran, S.; Concepcion, C.; Koes, D.R., CACHE Challenge #1: Docking with GNINA Is All You Need. *J. Chem. Inf. Model.* **2024**, 64, 9388–9396. https://doi.org/10.1021/acs.jcim.4c01429.
13. Domínguez-Ramírez, L.; Anaya-Ruiz, M.; Cortés-Hernández, P., Quality over quantity: how to get the best results when using docking for repurposing. *Front. Bioinform.* **2025**, 5, 1536504. https://doi.org/10.3389/fbinf.2025.1536504.
14. Ramírez, D.; Caballero, J., Is It Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data? *Molecules* **2018**, 23, 1038. https://doi.org/10.3390/molecules23051038.
15. Pereira, J.C.; Caffarena, E.R.; Dos Santos, C.N., Boosting Docking-Based Virtual Screening with Deep Learning. *J. Chem. Inf. Model.* **2016**, 56, 2495–2506. https://doi.org/10.1021/acs.jcim.6b00355.
16. Empereur-mot, C.; Guillemain, H.; Latouche, A.; Zagury, J.-F.; Viallon, V.; Montes, M., Predictiveness curves in virtual screening. *J. Cheminform.* **2015**, 7, 52. https://doi.org/10.1186/s13321-015-0100-8.
17. Wang, S.; Jiang, J.-H.; Li, R.-Y.; Deng, P., Docking-based virtual screening of TβR1 inhibitors: evaluation of pose prediction and scoring functions. *BMC Chemistry* **2020**, 14, 52. https://doi.org/10.1186/s13065-020-00704-3.
18. Shamsara, J., Correlation between Virtual Screening Performance and Binding Site Descriptors of Protein Targets. *Int. J. Med. Chem.* **2018**, 2018, 1–10. https://doi.org/10.1155/2018/3829307.
19. Cross, J.B.; Thompson, D.C.; Rai, B.K.; Baber, J.C.; Fan, K.Y.; Hu, Y.; Humblet, C., Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, 49, 1455–1474. https://doi.org/10.1021/ci900056c.
20. Chen, H.; Lyne, P.D.; Giordanetto, F.; Lovell, T.; Li, J., On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2006**, 46, 401–415. https://doi.org/10.1021/ci0503255.
21. Sander, T.; Freyss, J.; Von Korff, M.; Rufener, C., DataWarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, 55, 460–473. https://doi.org/10.1021/ci500588j.
22. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R., Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, 3, 33. https://doi.org/10.1186/1758-2946-3-33.
23. Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S., Extended TIGHT-BINDING quantum chemistry methods. *WIREs Comput. Mol. Sci.* **2021**, 11, e1493. https://doi.org/10.1002/wcms.1493.
24. Trott, O.; Olson, A.J., AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, 31, 455–461. https://doi.org/10.1002/jcc.21334.
25. Bell, E.W.; Zhang, Y., DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. *J. Cheminform.* **2019**, 11, 40. https://doi.org/10.1186/s13321-019-0362-7.
26. Alberca, L.N.; Prada Gori, D.N.; Fallico, M.J.; Fassio, A.V.; Talevi, A.; Bellera, C.L., LIDEB's Useful Decoys (LUDe): A freely available decoy-generation tool. Benchmarking and scope. *AI Life Sci.* **2025**, 7, 100129. https://doi.org/10.1016/j.ailsci.2025.100129.
27. Empereur-Mot, C.; Zagury, J.-F.; Montes, M., Screening Explorer–An Interactive Tool for the Analysis of Screening Results. *J. Chem. Inf. Model.* **2016**, 56, 2281–2286. https://doi.org/10.1021/acs.jcim.6b00283.

 doi:10.20944/preprints202506.0372.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.