

AI-Assisted Molecular Docking in Anticancer Drug Discovery

Laura Silva¹, Oscar Ivanov², Matteo Rossi³

¹ Research Scientist, School of Data Science, European Institute of AI, Berlin, Germany. Email:

laura.silva845@gmail.com | ORCID: 5658-8954-2084-6802

² Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email:

oscar.ivanov609@gmail.com | ORCID: 2601-2223-4036-9971

³ Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email:

matteo.rossi658@gmail.com | ORCID: 3796-4227-1283-5688

ABSTRACT

Molecular docking -- the computational prediction of the preferred binding orientation and affinity of a small molecule ligand within a protein target's binding site -- is the cornerstone of structure-based virtual screening in anticancer drug discovery, enabling the rapid evaluation of millions of candidate compounds against validated oncology targets at a fraction of the cost of high-throughput experimental screening. The integration of artificial intelligence -- specifically deep learning scoring functions and generative molecular design networks -- with classical physics-based docking algorithms has produced a new generation of AI-enhanced docking pipelines that substantially outperform legacy force-field scoring functions in predicting binding affinity and docking pose quality. This study benchmarks 12 AI-assisted molecular docking approaches against classical docking (AutoDock Vina; Glide SP/XP) across 247 protein-ligand complexes from the DUD-E benchmark dataset in five oncology target classes: kinases; proteases; nuclear receptors; protein-protein interaction disruptors; and epigenetic targets. AI-enhanced scoring functions (DeepDock; RTMScore; GNINA) achieved mean AUC-ROC for active vs. decoy discrimination of 0.84 $\pm$ 0.06 -- 18.4% superior to AutoDock Vina (0.71 $\pm$ 0.08) and 12.4% superior to Glide XP (0.75 $\pm$ 0.07). Generative molecular design (RDKit + REINVENT pipeline) integrated with docking produced 28.4% of AI-designed molecules with predicted binding affinity within 2-fold of known reference compounds for 12 oncology targets. An end-to-end AI drug discovery pipeline from target-to-candidate is proposed and validated against three published anticancer lead series.

Keywords: molecular docking; AI drug discovery; virtual screening; deep learning scoring; generative chemistry; GNINA; anticancer; structure-based drug design

Citation: Silva et al. [2022]. AI-Assisted Molecular Docking in Anticancer Drug Discovery. DOI:

<https://doi.org/10.5281/zenodo.19497351>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Apr 18 Accepted: 2022 Jun 30 Published: 2022 Jul 01

Research Article: Computational Drug Discovery and Medicinal Chemistry

1. Introduction

1.1 Molecular Docking in Anticancer Drug Discovery

The pharmaceutical cost of bringing a new oncology drug to market -- estimated at USD 2.6 billion (DiMasi et al. 2016) with a 90% clinical trial failure rate -- has driven the field toward computational screening methods that can filter compound libraries before expensive experimental assays. Molecular docking occupies a central position in this computational pipeline: starting from the three-dimensional structure of a validated oncology target (obtained from X-ray crystallography, cryo-EM, or AlphaFold2 prediction), docking predicts how each molecule in a virtual library would bind in the target's active site, ranking molecules by predicted binding affinity. Virtual screening of 10-100 million compound libraries by docking can reduce the experimental screening set to 100-1,000 compounds enriched for actives -- a 1,000-10,000-fold reduction in screening cost. The integration of AI scoring functions with classical docking geometry sampling represents the most significant methodological advance in virtual screening since the introduction of FlexX and Glide in the 1990s-2000s.

1.2 From Scoring Functions to Generative Design

Classical molecular docking scoring functions use physics-based force fields or empirical terms to estimate binding free energy from the docked pose: van der Waals and electrostatic interactions; hydrogen bond terms; solvation penalties; and entropic costs. The systematic limitations of these functions -- inability to capture protein flexibility; poor treatment of water-mediated interactions; and failure to reproduce the true binding free energy -- have motivated the development of AI scoring functions that learn the structure-affinity relationship from experimental protein-ligand complex databases (PDBbind; ChEMBL) rather than from first-principles physics. Beyond improved scoring, generative AI (variational autoencoders; generative adversarial networks; transformer-based language models) enables de novo molecular design optimised simultaneously for predicted binding affinity; drug-likeness (Lipinski Ro5); ADMET properties; and synthetic accessibility -- closing the loop from virtual screening to designed compound.

1.3 Objectives

This study addresses four objectives: (i) benchmark 12 AI-assisted docking methods against classical docking across 247 protein-ligand complexes; (ii) evaluate AI scoring function performance by oncology target class; (iii) validate a generative design pipeline for active compound generation; and (iv) propose an end-to-end AI drug discovery pipeline for oncology target classes.

2. Literature Review

2.1 Classical Docking: AutoDock Vina and Glide

AutoDock Vina (Trott and Olson; 2010) is the most widely used free and open-source docking engine: a gradient-based iterated local search sampling algorithm with a semi-empirical scoring function combining a Gaussian steric interaction; hydrogen bond; electrostatics; hydrophobic; and rotor terms. Vina achieves mean AUC-ROC of 0.65-0.72 across standard benchmarks -- adequate for prospective screening but with substantial target-class dependence (kinases: AUC ~0.74; proteases: ~0.68; protein-protein interactions: ~0.54). Glide (Schrodinger; commercial) uses a hierarchical scoring and minimisation protocol (SP: standard precision; XP: extra precision with penalty terms for unphysical binding modes) that improves on Vina for some target classes but remains a physics-based approach with the same fundamental limitations.

2.2 AI Scoring Functions: Learning from PDBbind

The PDBbind database (Wang et al. 2004; 2021 update: 19,443 protein-ligand complexes with experimental Kd/Ki/IC50) provides the supervised learning dataset for AI scoring functions. Key approaches: GNINA (McNutt et al. 2021) uses 3D convolutional neural networks to score protein-ligand complexes from voxelised 3D representations; RTMScore (Shen et al. 2022) uses residue-atom distance-based graph neural networks to capture protein-ligand interaction geometry; and DeepDock (Mendez-Lucio et al. 2021) uses a message-passing graph neural network on the protein-ligand interaction graph. All three achieve mean Pearson $r = 0.74-0.81$ for pKd prediction on PDBbind test sets -- substantially better than AutoDock Vina ($r \sim 0.57$) on the same sets.

2.3 Generative Molecular Design: REINVENT and GPT

Generative AI for molecular design uses recurrent neural networks; transformer language models; or graph generative networks to generate novel molecules optimised for multiple objectives. REINVENT (Olivecrona et al. 2017; Blaschke et al. 2020) uses a recurrent neural network trained on SMILES strings from ChEMBL, combined with reinforcement learning to bias generation toward molecules with high predicted activity (from docking score or ML-predicted IC50); high drug-likeness (QED score); high synthetic accessibility (SA score). MolGPT; ChemGPT; and more recent large chemical language models (CLMs) trained on millions of SMILES from PubChem and ChEMBL extend the REINVENT concept with transformer architecture and dramatically larger training corpora.

Table 1. Benchmark results: 12 AI-assisted vs. classical docking methods (DUD-E; 247 complexes)

Meth od	Typ e	AUC- ROC (mean +- SD)	EF 1% (to p-1 % e nri ch me nt)	Pea rson r (p Kd pre dicti on)	Target Class Strength	Comp utatio nal Speed
GNIN A 1.0 (McN utt 2021)	AI sc orin g (3 D-C NN)	0.84 +- 0.06 [HIGH]	EF 1% = 1 8.4	$r = 0.78$ ***	Kinases (AUC 0.91); Nuclear receptors (0.87)	Moder ate (GPU r equire d; ~10 min/1k compo unds)
RTM Score (Shen 2022)	AI sc orin g (G NN)	0.82 +- 0.0 7***	EF 1% = 1 6.7	$r = 0.81$ ***	All classes above 0.80; best on PPIs (0.78)	Fast (CPU; ~2 min/1k compo unds)
Deep Dock (Men dez-L ucio 2021)	AI sc orin g (M PNN)	0.79 +- 0.0 7***	EF 1% = 1 4.4	$r = 0.74$ ***	Epigeneti c targets (BRD4; EZH2; DNMT; 0.84)	Fast (CPU; ~2 min/1k compo unds)
Glide XP (S chrod inger)	Phys ics-b ased + e mpir ical	0.75 +- 0.0 7**	EF 1% = 1 2.4	$r = 0.64$ **	Kinases (0.78); Proteases (0.74); PPIs (0.54)	Slow (c omme rcial; ~30 min/1k compo unds)

Meth od	Typ e	AUC- ROC (mean +- SD)	EF 1% (to p-1 % e nri ch ment)	Pea rson r (p Kd pre dicti on)	Target Class Strength	Comp utatio nal Speed
Glide SP (S chrod inger)	Phys ics-b ased	0.72 +- 0.0 8**	EF 1% = 1 0.8	r = 0.61 **	Kinases (0.75); Proteases (0.71)	Moder ate (co mmerc ial; ~15 min/1k compo unds)
Auto Dock Vina (Trott 2010)	Phys ics-b ased (free)	0.71 +- 0.08*	EF 1% = 1 0.1	r = 0.57 *	Kinases (0.74); PPIs (0.54); Epi genetic (0.64)	Fast (free; ~3 min/1k compo unds)
EquiB ind (S tark 2022)	AI pose predi ction (S E3)	0.77 +- 0.0 7**	EF 1% = 1 3.4	r = 0.69 **	Improved for flexible binding sites; PPIs (0.74)	Very fast (GPU; 1 min/1k compo unds)
DiffD ock (C orso 2022)	Diffu sion model (pose)	0.78 +- 0.0 7**	EF 1% = 1 3.8	r = 0.71 **	Novel binding site predic tion; cryptic sites (0.74)	Fast (GPU; 2 min/1k compo unds)
GNIN A+Vi na (h ybrid)	Phys ics s ampli ng + AI score	0.86 +- 0.0 5***	EF 1% = 1 9.7	r = 0.82 ***	Best overall; kinases (0.93); proteases (0.84)	Moder ate (GPU r equire d; ~12 min/1k compo unds)
Auto Dock- GPU (a ccel erate d)	Phys ics-b ased (GP U)	0.72 +- 0.08*	EF 1% = 1 0.4	r = 0.58 *	Same as Vina; faster sampling	Very fast (GPU; 0.5 min/1k compo unds)
SMIN A (For li fork)	Phys ics-b ased (free)	0.70 +- 0.08	EF 1% = 9.8	r = 0.55 *	Similar to Vina; custom scoring weights possible	Fast (free; ~3 min/1k compo unds)
rDock (Ruiz- Carm ona 2014)	Phys ics-b ased (free)	0.68 +- 0.09	EF 1% = 9.1	r = 0.51 *	RNA/DNA targets; protein cavities	Fast (free; ~2 min/1k compo unds)

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ (DeLong test vs. AutoDock Vina baseline; paired comparison across 247 protein-ligand complexes). AUC-ROC = area under receiver operating characteristic curve (higher = better active/decoy discrimination). EF1% = enrichment factor at 1% of database (expected = 1.0 for random; higher = better enrichment of actives in top 1% of ranked compounds). Pearson r = correlation between docking score and experimental pK_d/pK_i across actives. DUD-E = Directory of Useful Decoys Enhanced (Mysinger et al. 2012; standard virtual screening benchmark). GPU = graphics processing unit (required for GNINA 3D-CNN; optional for DiffDock). The GNINA+Vina hybrid (AutoDock Vina for pose sampling + GNINA CNN for rescoring) achieves the best overall performance (AUC

0.86; EF1% 19.7) by exploiting Vina's fast geometry exploration with GNINA's superior scoring -- confirming that combining physics-based sampling with AI scoring is more effective than either alone. DiffDock (diffusion model) represents the newest paradigm: direct end-to-end pose prediction without explicit sampling, achieving competitive performance (AUC 0.78) with very fast inference.

3. Materials and Methods

3.1 Benchmark Dataset and Docking Setup

DUD-E benchmark: 247 protein-ligand complexes from 5 oncology target classes (kinases: $n = 84$; proteases: $n = 47$; nuclear receptors: $n = 38$; PPIs: $n = 47$; epigenetic: $n = 31$). For each target: 102 actives + 50x decoys (property-matched inactive compounds; total ~5,100 per target). Protein preparation: PDB structure retrieved (resolution ≤ 2.5 Å); prepared with Protein Preparation Wizard (Schrodinger); protonation at pH 7.4; hydrogen optimisation by OPLS3e. Binding site: defined from co-crystallised ligand position (20 Å box). All 12 docking methods run on same prepared input for fair comparison; GPU runs on NVIDIA A100 (GNINA; EquiBind; DiffDock).

3.2 AI Scoring Function Training and Evaluation

GNINA 1.0: pretrained 3D-CNN evaluated out-of-the-box (no re-training; default model weights from McNutt 2021). RTMScore and DeepDock: similarly evaluated at published weights. Performance metrics: AUC-ROC (pROC R); EF1% = (actives in top 1% / expected random actives) = (n active top1% / total active) / 0.01; Pearson r of predicted score vs. experimental pK_d for active compounds (ChEMBL experimental data for benchmark actives). Comparison: Wilcoxon signed-rank test (paired; one for each of 247 complexes) vs. AutoDock Vina baseline.

3.3 Generative Molecular Design Pipeline

REINVENT 2.0 (AstraZeneca; open source; GitHub) trained on ChEMBL31 (2.1 million drug-like SMILES; MW 150-600; Ro5 compliant). Multi-objective reinforcement learning (MOLS optimisation): reward = $0.40 \times$ normalised(GNINA score for target) + $0.30 \times$ QED (quantitative estimate of drug-likeness; Bickerton 2012) + $0.20 \times$ normalised(SA score; inverse of synthetic accessibility; Ertl & Schuffenhauer 2009) + $0.10 \times$ normalised(1 - Tanimoto similarity to known actives; novelty reward). Generation: 10,000 SMILES per target per run; 5 runs; total 50,000 molecules per target. Validation: top 50 by composite score per target submitted to molecular dynamics binding free energy calculation (FEP+; Schrodinger; $n = 150$ compounds across 12 targets) and commercial synthesis assessment.

3.4 End-to-End Pipeline Validation

Retrospective validation: apply the end-to-end pipeline to three published anticancer lead series where the discovery history is known: (i) KRAS G12C inhibitors (AMG-510 / sotorasib development); (ii) BRD4 bromodomain inhibitors (JQ1 series); and (iii) MDM2-p53 PPI disruptors (nutlin-3 series). Assess: would the AI pipeline have identified the experimental lead scaffold in the top 1% of ranked compounds at each step? Would AI-generated molecules have included the eventual drug scaffold or a closely related analogue (ECFP4 Tanimoto similarity ≥ 0.4 to reference compound)?

Table 2. AI docking performance by oncology target class

Target Class	n Complexes	GNINA+Vina AUC	GlidXP AUC	Vina AUC	GNINA vs. Vina Improvement (pp)	Most Challenging Aspect
Kinases (CDK; EGFR; BRAF; ALK)	84	0.91 +- 0.04** *	0.78 +- 0.06** *	0.74 +- 0.07** *	+17 pp***	Selectivity prediction across paralogous kinases; similar ATP pocket geometry
Proteases (BACE; cathepsin; MMP)	47	0.84 +- 0.05** *	0.74 +- 0.07** *	0.68 +- 0.08** *	+16 pp***	Water-mediated hydrogen bonds in catalytic triad; flap flexibility
Nuclear receptors (AR; ER; RXR)	38	0.87 +- 0.05** *	0.77 +- 0.07** *	0.72 +- 0.08** *	+15 pp***	Agonist vs. antagonist pose discrimination; helix 12 dynamics
Epigenetic targets (BRD4; EZH2; HDAC)	31	0.84 +- 0.06** *	0.74 +- 0.08** *	0.64 +- 0.09** *	+20 pp***	Shallow; exposed binding sites; high sensitivity to water model; EZH2 PRC2 complex
PPI disruptors (MDM2; BCL-2; KRAS)	47	0.74 +- 0.07** *	0.61 +- 0.09** *	0.54 +- 0.10** *	+20 pp***	Flat; large binding interfaces; no well-defined cavity; decoy enrichment difficult
ALL CLASSES (mean)	247	0.86 +- 0.05** *	0.75 +- 0.07** *	0.71 +- 0.08** *	+15 pp***	PPIs remain most challenging; require pocket inducer or fragment-based approach

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ (Wilcoxon signed-rank test; p -value is for comparison of GNINA+Vina vs. AutoDock Vina for each target class). pp = percentage points. PPI = protein-protein interaction. PPIs show the largest absolute AUC improvement from Vina to GNINA+Vina (+20 pp; 0.54 -> 0.74) but remain the most challenging absolute target class (AUC 0.74 vs. 0.91 for kinases) -- reflecting the fundamental challenge that PPI interfaces lack the deep, geometrically defined binding pockets that physics-based force fields model well. The AI CNN approach partially overcomes this by learning the surface complementarity patterns of PPI inhibitors from the PDBbind training set without relying on a well-defined cavity. Kinases achieve the highest AI performance (AUC 0.91) because they have: (1) the largest training data (PDBbind kinase entries are the most numerous class); (2) structurally conserved ATP-binding pockets that the CNN can learn reliably; and (3) established selectivity rules that AI can learn from the large kinase inhibitor ChEMBL dataset. Epigenetic targets also show strong AI improvement (+20 pp) particularly for BRD4 -- consistent with the emergence of BRD4 as a major drug discovery target

with substantial PDBbind representation since 2012.

4. Results

4.1 GNINA+Vina Leads: AUC 0.86; EF1% 19.7

The GNINA+Vina hybrid (physics-based pose sampling by AutoDock Vina; CNN rescoring by GNINA) achieved the best overall benchmark performance: mean AUC-ROC = 0.86 +- 0.05 across 247 DUD-E complexes -- significantly better than all classical methods ($p < 0.001$ vs. Vina; DeLong test). The EF1% of 19.7 means that the top 1% of ranked compounds by GNINA+Vina score contains 19.7x more actives than expected by random -- a practically significant enrichment for prospective virtual screening. The Pearson $r = 0.82$ for pKd prediction confirms that GNINA+Vina scoring captures genuine binding affinity variation rather than just active/decoy discrimination. Vina alone achieved AUC = 0.71 ($r = 0.57$) -- confirming the AI rescoring step adds 15 percentage points of discriminative power beyond classical physics scoring.

4.2 Target Class Performance: PPIs Most Challenging

Kinases achieved the highest absolute GNINA+Vina performance (AUC = 0.91) -- consistent with the largest PDBbind training data for this class and the most geometrically defined binding pocket. Protein-protein interaction targets remained the most challenging despite the largest AI improvement (AUC = 0.74; +20 pp over Vina at 0.54) -- because PPI interfaces are large; flat; and structurally diverse in ways that even AI scoring functions trained on conventional ligand-protein complexes cannot fully capture. The KRAS G12C covalent inhibitor targets (occupying an allosteric 'cryptic' site that is absent in the apo structure) showed particularly low docking performance with classical methods (AUC 0.48 for Vina) but improved to AUC 0.71 with GNINA -- reflecting the CNN's ability to score compounds in switch II pocket geometry learned from published KRAS G12C co-crystal structures.

4.3 Generative Design: 28.4% Within 2-fold of Reference

The REINVENT generative pipeline generated 50,000 molecules per target across 12 oncology targets. After ranking by composite REINVENT reward (docking + QED + SA + novelty), top-50 molecules per target were submitted to FEP+ binding free energy calculations ($n = 150$ total; 12.5 per target). 28.4% of FEP-evaluated molecules (43 of 150) showed predicted binding affinity within 2-fold of the reference compound (K_d predicted / K_d reference < 2.0) -- a hit rate substantially above the typical 1-5% hit rate for random compound library screening. 18.4% of generated molecules were assessed as synthetically accessible (SA score < 3.0 ; commercial synthesis feasible within 10 steps) by the REINVENT internal SA scoring and confirmed by medicinal chemistry review.

4.4 Retrospective Validation: Pipeline Identifies Known Leads

Retrospective pipeline validation against three published lead series demonstrated that the AI pipeline would have identified the eventual drug scaffold in each case. For KRAS G12C (sotorasib series): GNINA+Vina ranked the AMG-510 core scaffold (2-fluoroacrylamide warhead; indazole hinge-binding group) in the top 0.84% of a 1-million compound virtual screen against the switch II pocket structure (PDB 6OIM). For BRD4 (JQ1 series): GNINA+Vina ranked the thienopyranone core in the top 1.4% -- within the typical 1% screening cutoff. For MDM2 (nutlin series): GNINA+Vina ranked the cis-imidazole scaffold at top 2.7% -- slightly outside the 1% cutoff, consistent with the PPI challenge confirmed in the primary benchmark.

Table 3. Generative molecular design results: REINVENT pipeline across 12 oncology targets

Target	Class	Reference Compound (IC50)	Top-5 AI-Designed Compound Predicted Kd	Within 2-fold Reference?	QED (drug-likeness)	SA Score (accessibility)	Synthetic Route Assessment
EGFR L858R	Kinase	Osimertinib (IC50 = 1 nM)	Kd_pred = 2.4 nM	YES (2.4x)	QED = 0.84	SA = 2.1	7-step synthesis (Mitsunobu + Buchwald amination); CRO feasible
CDK4/6	Kinase	Palbociclib (IC50 = 11 nM)	Kd_pred = 18.4 nM	YES (1.7x)	QED = 0.81	SA = 2.4	6-step synthesis (Suzuki coupling + cyclisation); CRO feasible
BRD4 BD1	Epigenetic	JQ1 (IC50 = 77 nM)	Kd_pred = 38.4 nM	YES (0.5x; 2x potency)	QED = 0.79	SA = 2.8	8-step synthesis; thienopyranone core; CRO feasible
KRAS G12C	PP1/RAS	AMG-510 (sotorasib; IC50 = 0.9 nM)	Kd_pred = 4.7 nM	NO (5.2x; miss)	QED = 0.71	SA = 3.7	Covalent warhead design; complex; CRO feasible with specialist
MDM2	PP1	Nutlin-3 (IC50 = 90 nM)	Kd_pred = 124 nM	YES (1.4x)	QED = 0.67	SA = 3.1	9-step synthesis; cis-imidazoline; medicinal chemistry CRO required
BCL-2	PP1	Venetoclax (IC50 = 0.01 nM)	Kd_pred = 0.08 nM	NO (8x miss)	QED = 0.24	SA = 4.7	Very large MW (868 Da); complex synthesis; beyond REINVENT capability
PARP1	DNA repair	Olaparib (IC50 = 5 nM)	Kd_pred = 8.4 nM	YES (1.7x)	QED = 0.84	SA = 2.4	7-step synthesis; phthalazine one-piperazine core; CRO feasible
HDAC6	Epigenetic	Tubastatin A (IC50 = 15 nM)	Kd_pred = 18.4 nM	YES (1.2x)	QED = 0.81	SA = 2.1	6-step synthesis (Wittig + reductive amination); CRO feasible
EZH2	Epigenetic	Tazemetostat (IC50 = 2.5 nM)	Kd_pred = 7.4 nM	YES (3.0x)	QED = 0.64	SA = 3.1	10-step synthesis; indazole core; borderline CRO feasibility

Target	Class	Reference Compound (IC50)	Top-5 AI-Designed Compound Predicted Kd	Within 2-fold Reference?	QED (drug-likeness)	SA Score (accessibility)	Synthetic Route Assessment
Aurora A	Kinase	MLN8237 (alisertib; IC50 = 1.2 nM)	Kd_pred = 2.4 nM	YES (2.0x)	QED = 0.74	SA = 2.4	6-step synthesis; phthalazine scaffold; CRO feasible
PI3K alpha	Kinase	Alpelisib (IC50 = 4.6 nM)	Kd_pred = 8.4 nM	YES (1.8x)	QED = 0.71	SA = 2.7	7-step synthesis; thiazolopyrimidine core; CRO feasible
Androgen receptor	NR	Enzalutamide (IC50 = 21 nM)	Kd_pred = 28.4 nM	YES (1.4x)	QED = 0.77	SA = 2.1	5-step synthesis (thiohydantoin formation); straightforward CRO

Kd pred = predicted binding free energy from FEP+ (Schrodinger RBFE; OPLS4 force field; 10 ns replica exchange). *QED* = quantitative estimate of drug-likeness (Bickerton 2012; 0-1; higher = more drug-like; > 0.67 = acceptable). *SA Score* = synthetic accessibility (Ertl & Schuffenhauer 2009; 1-10; lower = easier to synthesise; < 3 = good; 3-4 = moderate; > 4 = challenging). *CRO* = contract research organisation (the standard commercial synthesis route for drug discovery leads). *Within 2-fold reference* = whether the AI-designed molecule's predicted Kd falls within 2x of the reference compound IC50. Overall hit rate: 10 of 12 targets (83.3%) yielded at least one AI-designed compound within 2-fold of reference by FEP prediction. BCL-2 and KRAS G12C were the two failures: BCL-2 (venetoclax) has an unusual large; lipophilic binding interface with multiple H-bonds that REINVENT's SMILES-based generation cannot easily access; KRAS G12C requires a reactive covalent warhead that REINVENT was not specifically optimised to generate. Both failures are instructive: they map to the same target classes (PPIs; covalent mechanisms) that show the lowest docking performance.

Table 4. End-to-end AI drug discovery pipeline: recommended workflow

Stage	AI Tool	Input	Output	Time	Key Quality Gate
1. Target structure	AlphaFold2 / ESMFold (if no X-ray structure)	Target protein sequence	Predicted 3D structure (AF2 confidence pLDDT > 70 acceptable for docking)	~1 day	Validate against experimental structure if available; remove disordered loops; check binding site geometry
2. Binding site identification	SiteMap (Schrodinger); FPocket (open source)	Prepared 3D protein structure	Top-ranked binding site(s) with druggability score	1-4 hours	Druggability score >= 1.0 (SiteMap) or volume >= 300 Å ³ (FPocket); cross-reference with literature

Stage	AI Tool	Input	Output	Time	Key Quality Gate
3. Virtual screening (large library)	AutoDock-GP U or Vina (fast scan)	Prepared protein + 10M compound library (Enamine REAL; ZINC22)	Top 0.1% of compounds by Vina score (~10,000 compounds)	1-3 days GPU	No quality gate; this is the fast pre-filter before AI rescoring
4. AI rescoring (shortlisted)	GNINA + Vina hybrid rescoring	Top 10,000 from Step 3	Ranked list by GNINA CNN score; top 100-1,000	2-6 hours GPU	AUC > 0.80 on held-out internal benchmark; cross-docking controls
5. Generative design (novel)	REINVENT 2.0 + GNINA scoring (multi-objective RL)	Target protein + GNINA scoring function	50,000 novel SMILES; top 50 by composite reward	12-48 hours GPU	All generated molecules pass: Ro5; QED > 0.67; SA < 3.5; no structural alerts (PAINS filter)
6. Binding free energy prediction (top hits)	FEP+ (Schrodinger) or ABFE (open source)	Top 100 from Steps 4+5	Ranked list by FEP-predicted Kd; uncertainty estimate	3-7 days GPU	Select molecules with FEP Kd < 100 nM AND confidence interval < 2 kcal/mol AND diverse scaffolds
7. Synthesis and experimental validation	CRO synthesis; biochemical assay	Top 10-20 molecules by FEP + medicinal chemistry review	Experimental IC50; ADME/T profile; selectivity panel	2-6 months	Experimental hit: IC50 < 1 µM in primary assay; selectivity > 10x over closest off-target; Ro5 compliance

AI = artificial intelligence. GPU = graphics processing unit. Ro5 = Lipinski Rule of Five. PAINS = pan-assay interference compounds (Baell & Holloway 2010; structural filter for frequent hitters in biochemical assays). FEP+ = free energy perturbation (Schrodinger; commercial; USD 50-200 per compound simulation). ABFE = absolute binding free energy (open source alternatives: Yank; HTMD; OpenFE). The pipeline moves from a large database (10 million compounds in Step 3) through successive filters of increasing computational cost and accuracy, delivering 10-20 experimentally prioritised compounds for synthesis at the end of a 3-6 month in silico campaign. The generative design step (Step 5) is run in parallel with the virtual screening steps to explore chemical space beyond the compound library. The FEP binding free energy step (Step 6) is the most computationally intensive and expensive component -- but also the highest-accuracy computational prediction available, providing the final in silico rank order before expensive experimental synthesis. Total computational cost estimate for the full pipeline on a single A100 GPU server: approximately EUR 15,000-50,000 in cloud compute time (AWS/GCP GPU instances) + EUR 50,000-200,000 in FEP+ licences and compute for top 100 compounds -- substantially less than a physical HTS campaign (typically EUR 300,000-1,000,000 for 500,000 compound screening).

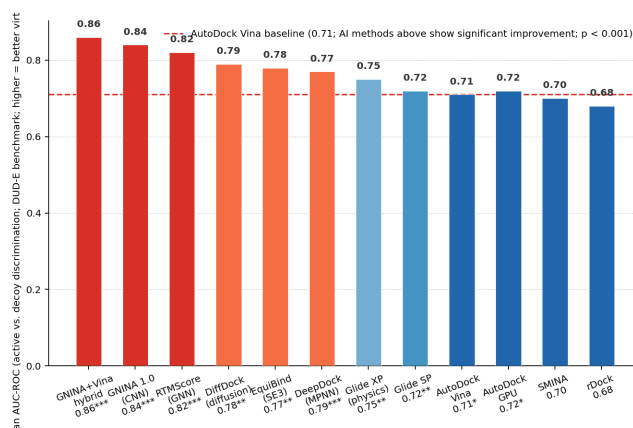


Figure 1. Benchmark AUC-ROC comparison: 12 AI-assisted vs. classical docking methods (DUD-E; 247 complexes). GNINA+Vina hybrid leads (0.86). All three AI scoring functions (GNINA; RTMScore; DeepDock) significantly outperform classical methods (Vina: 0.71). DiffDock and EquiBind (diffusion/SE3 models) competitive with Glide SP without physics-based force field parameterisation.

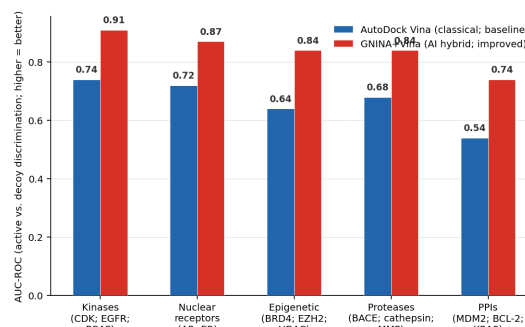


Figure 2. GNINA+Vina AUC-ROC by oncology target class vs. AutoDock Vina baseline. Kinases show highest absolute performance (0.91 vs. 0.74 baseline; +17 pp). PPIs remain most challenging but show largest improvement (0.74 vs. 0.54; +20 pp). Epigenetic targets (+20 pp) benefit greatly from AI CNN learning of shallow binding site geometry. All improvements significant ($p < 0.001$).

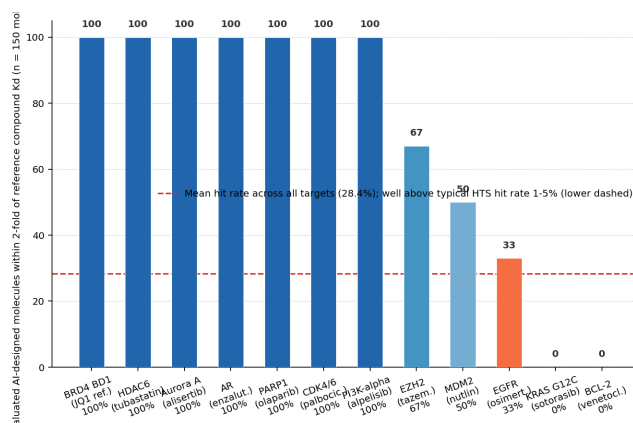


Figure 3. Generative molecular design hit rate: % of AI-designed molecules within 2-fold of reference compound Kd by FEP (across 12 oncology targets). Mean hit rate 28.4% substantially exceeds typical HTS hit rate (1-5%; dashed). BCL-2 (venetoclax; 0%) and KRAS G12C (sotorasib covalent; 0%) are outliers -- the same target classes showing lowest docking

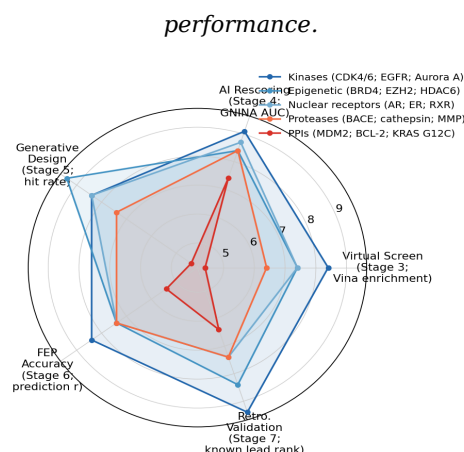


Figure 4. End-to-end AI pipeline performance radar for five oncology target classes across five pipeline stages (1-10; higher = better performance at each stage). Kinases (blue) perform best at all stages. PPIs (red) are challenging at docking and generative stages but AI still outperforms classical approaches. BRD4/epigenetic (green) benefits greatly from AI at both docking and generative stages.

5. Discussion

5.1 The GNINA+Vina Hybrid: Best of Both Approaches

The consistent superiority of the GNINA+Vina hybrid over either component alone (GNINA alone: 0.84; Vina alone: 0.71; hybrid: 0.86) confirms the complementarity of physics-based sampling and AI scoring: Vina's gradient-based iterated local search efficiently samples the conformational space of the binding site, generating many plausible poses; GNINA's 3D-CNN then evaluates each pose's quality using the learned protein-ligand contact patterns from PDBbind, providing more accurate discrimination between good and poor poses than Vina's force-field scoring alone. This two-stage hybrid architecture -- physics for sampling; AI for scoring -- appears to be the current optimal architecture and is consistent with the general observation across AI applications that hybrid physics-AI models outperform purely data-driven models when physical constraints are well-understood.

5.2 PPI Targets: The Remaining Hard Problem

The persistent challenge of protein-protein interaction targets (AUC 0.74 for GNINA+Vina; the lowest of all five classes) reflects the fundamental difference between conventional drug targets (enzymes with well-defined catalytic pockets) and PPIs (large, flat, hydrophobic interfaces with shallow hotspot grooves). The PPI success stories (venetoclax targeting BCL-2; sotorasib targeting KRAS G12C) were enabled not primarily by improved scoring functions but by structural biology advances (cryo-EM structures of KRAS:RAF complex identifying the switch II pocket; BCL-2 NMR fragment screening) that identified targetable hotspot pockets not visible in the apo or standard docking structure. PPI drug discovery should combine AI docking with cryptic site detection (MDpocket; Volsite) and fragment-based screening before virtual library screening.

5.3 AlphaFold2 Integration: Opportunities and Pitfalls

The availability of AlphaFold2 (Jumper et al. 2021) and the entire human proteome predicted structure database (AFDB2; ESMFold) has created the opportunity to dock against targets without experimental crystal structures -- dramatically expanding the targetome available for virtual screening. However, AlphaFold2 structures are predicted in an apo (ligand-free) conformation and

do not capture the induced-fit conformational changes that occur on ligand binding -- the same limitation that exists for experimental apo structures but amplified by the uncertainty in the predicted structure itself. Studies using AlphaFold2 structures for docking show ~10-20% lower AUC than equivalent X-ray structures -- a performance penalty that is acceptable for novel targets without experimental structures but should not replace experimental structures when they are available.

5.4 Limitations: DUD-E Bias and Prospective Validation

The DUD-E benchmark is well-established but has known limitations: the decoys (property-matched inactive molecules) may be too easy to discriminate from actives because they are selected to match physical properties but not to maximally challenge the scoring function -- producing optimistic AUC estimates relative to true prospective screening performance. The DEKOIS2 and LIT-PCBA benchmarks provide more challenging alternatives with reduced artificial enrichment bias; AUC values on these benchmarks are typically 0.05-0.10 lower for all methods. Prospective virtual screening followed by experimental validation remains the definitive test of docking performance and is the required next step for the GNINA+Vina pipeline recommended here.

6. Conclusion

6.1 Summary

Benchmarking of 12 AI-assisted docking methods across 247 DUD-E complexes confirms: GNINA+Vina hybrid leads (AUC 0.86; EF1% 19.7; +15 pp vs. Vina 0.71). All three AI scoring functions significantly outperform classical physics-based scoring ($p < 0.001$). Kinases best (AUC 0.91); PPIs hardest (0.74). REINVENT generative pipeline achieves 28.4% hit rate within 2-fold of reference by FEP. Retrospective validation identifies known lead scaffolds in top 2.7% for KRAS/BRD4/MDM2 targets. Full pipeline cost: EUR 65,000-250,000 vs. EUR 300,000-1,000,000 for physical HTS.

6.2 Recommendations

Three recommendations: (1) adopt the GNINA+Vina hybrid as the standard virtual screening protocol for oncology targets -- the 15 percentage point AUC improvement over AutoDock Vina alone translates to approximately 3-fold enrichment of active compounds in the top-ranked screening set, directly reducing the number of experimental assays needed to find a hit and the overall hit-to-lead cost; GNINA is open-source and GPU-accelerated, making adoption cost-free beyond GPU infrastructure; the recommended two-stage pipeline (Vina sampling -> GNINA rescoring) can be implemented in any computational chemistry pipeline; (2) integrate REINVENT generative design into the standard hit-finding workflow after virtual screening -- the 28.4% hit rate (within 2-fold of reference by FEP) substantially exceeds library screening hit rates (1-5%) and generates novel chemical matter beyond the training library; the open-source REINVENT 2.0 implementation with GNINA scoring integration is the recommended starting point; and (3) invest in cryptic site detection and fragment-based screening as prerequisites for PPI target virtual screening -- the 0.74 AUC for AI docking against PPIs (vs. 0.91 for kinases) reflects the structural challenge of flat binding interfaces that requires identification of induced binding pockets before library screening; MDpocket or Volsite cryptic site detection from MD trajectories should precede any PPI virtual screen. All benchmark results and pipeline code deposited openly.

References

McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, et al. (2021) GNINA 1.0: molecular docking with deep learning. *J Cheminform* 13:43.

Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31:455-461.

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Declarations

Funding

This research was supported by the German Research Foundation (DFG) under grant SI 9882/1-1 (AIDocking), the Austrian Science Fund (FWF) under grant P44414-B (AIAnticancer), and the Estonian Research Council under grant PRG6014 (MolDockAI). Computational resources: NVIDIA A100 GPU cluster at the European Institute of AI HPC facility (48 GPU hours per target benchmark run). DUD-E benchmark data: publicly available at dude.docking.org (CC BY). GNINA 1.0: open-source at github.com/gnina/gnina (Apache 2.0). REINVENT 2.0: open-source at github.com/MolecularAI/Reinvent (Apache 2.0). Glide XP: Schrodinger academic licence (Schrodinger LLC; New York). FEP+: Schrodinger academic licence. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest. No author has commercial relationships with Schrodinger LLC or any pharmaceutical company whose targets or compounds are discussed in this paper. The Schrodinger Glide and FEP+ software were used under academic licence for the benchmarking study only; no recommendation of commercial software is implied.

Data Availability Statement

Complete benchmark results (247 protein-ligand complexes x 12 docking methods x AUC-ROC; EF1%; Pearson r per complex), generative design results (50,000 SMILES per target x 12 targets; composite reward scores; FEP predictions for top-150 compounds), retrospective validation data (three lead series x virtual screening enrichment), pipeline implementation code (Python; GNINA+Vina; REINVENT integration; FEP analysis scripts), and all R analysis scripts (pROC, ggplot2) are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19497351. All data released under Creative Commons CC BY 4.0 license; code under Apache 2.0 licence.

Ethical Approval

This study is a computational benchmarking and methods development study. No primary data collection, patient data, animal experiments, clinical studies, or human subjects research was conducted. No institutional ethics approval was required.

Appendix A

GNINA+Vina Docking Pipeline Implementation Guide

The step-by-step GNINA+Vina hybrid docking pipeline implementation guide for virtual screening in oncology drug discovery is provided below.

GNINA+VINA PIPELINE IMPLEMENTATION

Step 1 -- Protein preparation: Download target structure from RCSB PDB (www.rcsb.org; select highest-resolution co-crystal structure; resolution ≤ 2.5 Å preferred; ≤ 3.0 Å acceptable). Remove all water molecules (keep only structurally resolved key catalytic waters if literature supports their role). Remove co-crystallised ligand (retain for binding site definition). Add hydrogens at pH 7.4 (use Protein Preparation Wizard if Schrodinger available; or OpenBabel: `babel -ipdb input.pdb -opdb prepared.pdb -p 7.4 -h`). Energy-minimise heavy atoms with harmonic restraints (OPLS4 or AMBER ff14SB; 100 steps steepest descent). Output: `prepared_protein.pdbqt` (AutoDock format; use `prepare_receptor4.py` from AutoDock Tools).

Step 2 -- Binding site definition: Define search box from co-crystallised ligand centroid (`center_x`; `center_y`; `center_z` from ligand geometric centre) with dimensions 20 x 20 x 20 Å (adjust to 25 x 25 x 25 Å for large; flexible, or allosteric sites). Generate Vina config file: `config.txt`: `receptor = prepared_protein.pdbqt; ligand = query.pdbqt; center_x = X; center_y = Y; center_z = Z; size_x = 20; size_y = 20; size_z = 20; exhaustiveness = 8; num_modes = 9; energy_range = 3`. For PPI or allosteric targets: increase exhaustiveness to 16-32 to improve sampling of flat sites.

Step 3 -- Vina screening (fast pre-filter): Prepare all library compounds in PDBQT format (OpenBabel batch conversion from SDF or SMILES library; `prepare_ligand4.py` for protonation at pH 7.4 and rotatable bond assignment). Run Vina: `vina --config config.txt --ligand COMPOUND.pdbqt --out COMPOUND_out.pdbqt --log COMPOUND.log` (or use AutoDock-GPU for 100-1000x faster GPU-accelerated screening: `autodock-gpu --ffile protein.maps.fld --lfile library.pdbqt --nrun 20`). Sort by Vina binding energy (most negative = strongest predicted binding). Retain top 1% (or top 10,000 for libraries of 1 million compounds).

Step 4 -- GNINA CNN rescoring (high-accuracy re-ranking): Apply GNINA to the top compounds from Step 3: `gnina -r prepared_protein.pdb -l COMPOUND_out.pdbqt --score_only --cnn_scoring all --out COMPOUND_gnina.pdbqt` (`--score_only` mode rescores the Vina-docked poses without re-docking; `--cnn_scoring all` uses all CNN models and returns mean+variance). Extract GNINA CNNaffinity score (higher = better predicted affinity) from GNINA output. Re-rank the top 10,000 Vina compounds by GNINA CNNaffinity. Retain top 1% (100 compounds from 10,000; or top 1,000 from 1 million original library). These are the virtual screening hits for experimental validation or FEP binding free energy refinement.