SynGlue: AI-Driven Designer for Clinically Actionable Multi-Target Therapeutics

Saveena Solanki[1,†], Sanjay Kumar Mohanty[1,†], Shiva Satija[1], Sonam Chauhan[1], N.V.M. Rao Bandaru[3], Sandeep Dukare[3], Nirbhay Kumar Tiwari[3], Naveen Kumar R[3], Aravind A B[3], Subhendu Mukherjee [3], Vishakha Gautam[1], Sakshi Arora[1], Suvendu Kumar[1], Subhadeep Duari[1], Arushi Sharma[1], Raidhani Shome[1], Debarka Sengupta[1,2], Chandrasekhar Abbineni[3], Susanta Samajdar[3,#], Gaurav Ahuja[1,2,#]

[1]Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi, 110020, India.

[2]Infosys Centre for AI, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi, 110020, India.

[3]Aurigene Oncology Limited, Electronic City Phase II, Bangalore, India.

[†]Shared First Authors

[#]**Correspondence:**

Gaurav Ahuja; Email: gaurav.ahuja@iiitd.ac.in

Susanta Samajdar; Email: susanta_s@aurigene.com

Running title: SynGlue: AI for PROTAC & Multi-Target Design

Keywords: Generative AI, Fragments, Polypharmacology, Multi-Target, Cheminformatics, PROTACs

ABSTRACT

The rational design of multi-targeting compounds and targeted protein degraders, such as PROTACs, remains a significant challenge in drug discovery. Here, we introduce SynGlue, a modular generative AI platform that integrates fragment-based reasoning with data-driven and structure-guided approaches to design multi-target therapeutics. Its flexible four-module architecture enables the generation of chemically diverse, target-specific, and polypharmacologically optimized compounds. The platform supports user-defined scoring functions and external docking tools, offering an end-to-end pipeline from target structure to synthesizable candidate. Using this platform, we designed and synthesized novel BRD4-VHL PROTACs featuring previously unreported and uniquely structured warheads. *In vitro* profiling confirmed potent nanomolar growth inhibition across multiple cell lines, driven by selective BRD4 degradation and subsequent c-Myc downregulation. *In vivo*, the lead compound established favorable pharmacokinetics, pharmacodynamics, and tolerability profiles. Importantly, this lead PROTAC demonstrated significant tumor burden reduction in mouse xenograft models via its targeted degradation mechanism. In summary, this work validates SynGlue as a powerful generative AI platform for the rapid and effective design of multi-targeting therapeutics with promising preclinical anti-cancer activity.

INTRODUCTION

The drug discovery paradigm is rapidly evolving beyond the traditional "one target, one drug" model, advancing towards strategies that address the inherent complexity of human diseases [1,2]. Multi-targeting molecules have emerged at the forefront of this shift, offering the potential to achieve enhanced therapeutic outcomes by simultaneously targeting multiple nodes within disease networks [3,4]. This polypharmacological approach holds particular promise for complex multifactorial conditions such as cancer, where targeting a single gene/pathway is often insufficient [5–7]. Multi-targeting molecules can act through various mechanisms, including synergistic inhibition [8], induction of protein-protein interactions/proximity [9], and Targeted Protein Degradation (TPD) [10]. Among the most transformative multi-targeting strategies are Proteolysis Targeting Chimeras (PROTACs) [11,12]. PROTACs are heterobifunctional molecules comprising ligands for a target Protein of Interest (POI) and an E3 ubiquitin ligase, connected by a chemical linker [11,12]. This unique architecture enables PROTACs to catalytically induce ubiquitination and subsequent degradation of a variety of proteins that are traditionally deemed "undruggable" by conventional inhibitors. The success of PROTACs has significantly expanded the druggable proteome and opened unprecedented therapeutic avenues [13].

Despite their immense potential, rational design of highly effective and specific PROTACs remains a significant scientific hurdle [14,15]. Success in PROTAC design hinges on the intricate interplay and co-optimization of three components: the POI binder, the E3 ligase ligand, and the connecting linker. These elements collectively promote the formation of a stable and productive ternary complex (POI-PROTAC-E3 ligase), which dictates the efficiency and selectivity of degradation (DC_{50}) [14,15]. Optimizing linker length, composition, and attachment points is particularly challenging, as subtle changes can drastically alter ternary complex geometry and degradation outcomes [16,17]. To solve such multifaceted, complex optimization problems, computational methods are becoming increasingly indispensable. Structure-based methods aim to predict ternary complex formation and binding affinities, while ligand-based and AI-driven techniques leverage existing data to guide design and predict activity [18,19]. Several computational tools have been developed to aid in the design of PROTACs. Platforms like Molecular Operating Environment (MOE) and Internal Coordinate Mechanics (ICM) offer broad modeling capabilities but struggle to accurately predict novel ternary complex conformations.

Specialized tools like PRosettaC focus specifically on ternary complex modeling but can be computationally intensive [20]. Deep learning approaches such as DiffPROTACs show promise for linker design but may be limited by the scope of their training data [21]. Other workflows like PRODE [18] face persistent challenges related to accuracy, computational cost, accessibility, or the ability to generalize beyond PROTACs to other multi-targeting modalities. A critical need remains for a versatile, accurate, and accessible platform that integrates diverse data types and employs advanced AI for the *de novo* design of a wide range of multi-target molecules.

To overcome longstanding challenges in the rational design of multi-target therapeutics, we developed SynGlue, a next-generation computational platform powered by generative AI. SynGlue is purpose-built for the design and optimization of advanced modalities such as PROTACs and other multi-targeting molecules (MTMs). SynGlue directly links molecular design with validated therapeutic efficacy across multiple target classes. It uniquely integrates a vast data-driven knowledge base of ligand-target interactions (MagnetDB) with structure-based design principles within a unified and modular generative framework. This synergistic approach overcomes the aforementioned key limitations, such as limited target-specificity, static libraries, insufficient training data, and disjointed workflows, enabling accurate target prediction, specific binding site identification, and *de novo* designing of compounds with user-defined multi-target capabilities. Unlike existing tools that address isolated aspects of multi-target drug design, SynGlue uniquely integrates the largest curated ligand-target repository (MagnetDB; 6.37 M interactions) with ultra-fast TRIE-based fragment mapping (O(n) search time) and an AI-driven generative engine, delivering end-to-end design capabilities that accelerate the transition from concept to validated therapeutic in weeks rather than months. SynGlue's modular architecture encompasses five core interconnected modules: MagnetDB for curated interaction data, Computator and Annotator for interaction and type-specific analyses, the Generator for AI-driven molecular design, and Prioritizer for activity prediction and multi-stage filtering. All these collectively supporting both data-driven and structure-guided design workflows. SynGlue is a next-generation AI-powered platform that unites large-scale ligand-target data, structural modeling, and generative chemistry to design advanced multi-target therapeutics such as PROTACs. To showcase its real-world potential, we applied SynGlue to the urgent challenge of developing novel BRD4 degraders, key therapeutic agents in oncology, where resistance to existing inhibitors is a growing concern. SynGlue generated and prioritized candidate PROTACs through its integrated AI-driven design pipeline. We demonstrated its utility by implementing a complete discovery pipeline: from the AI-driven *de novo* design of novel BRD4-targeting

PROTACs, through their chemical synthesis and detailed *in vitro* characterization, to comprehensive *in vivo* validation encompassing pharmacokinetics, pharmacodynamics, tolerability, and therapeutic efficacy. Combined, these findings establish SynGlue as a powerful platform for accelerating the discovery and development of innovative multi-targeting therapeutics across oncology and beyond.

RESULTS

SynGlue Offers Rapid and Accurate Multi-Target Mapping

Effective rational design of multi-target therapeutics, including PROTACs, demands a robust computational framework built upon comprehensive, high-quality interaction data. Addressing this need, we developed SynGlue, an advanced platform that uniquely combines data-driven and structure-guided methodologies. The data-driven arm uses MagnetDB, a curated multi-source interaction database with TRIE-based indexing, alongside modules for fragment mapping (Computator), interaction-type classification (Annotator), and AI-driven molecule generation (Generator). The structure-guided arm designs cavity-specific ligands from protein structures and ranks binding warheads by physicochemical properties. A final Prioritizer module, which is common to both data or structure-guided approach, predicts activities such as half-maximal degradation concentration (DC50) and maximum fractional amount of degraded protein (D_{max}) and applies multi-stage filtering to select the most promising PROTACs and multitarget molecules **(Figure 1a)**. The data-driven module operates independently of the target structure, leveraging solely experimental interaction data **(Figure 1a)**, complementing SynGlue's structure-guided approach, which utilizes protein 3D structural information. The cornerstone of this module is MagnetDB, a meticulously curated database built to power multi-target analysis. We aggregated and rigorously filtered data from established sources (ChEMBL [22], BindingDB [23], DrugBank [24], STITCH [25], BioSNAP [26], and Small Molecule Suite [27]), resulting in a high-quality dataset of 6.37 million unique protein-ligand interactions, encompassing 1.94 million distinct compounds and 20,129 protein targets, a distilled and quality-filtered subset derived from 37.55 million raw entries **(Supplementary Figure 1a-b; Supplementary Table 1)**. Notably, in addition to taking protein 3D structure as an input, the structure-guided module of SynGlue offers a high degree of flexibility for the selection of tools for the cavity-specific *de novo* designing of the interaction domains (warheads), such as our recently introduced Gcoupler [28], DeepFrag [29], and MolAICal [30], among others **(Figure 1a)**.

To enable efficient analysis and scalable substructure-based searching across 1.94M compounds, within this vast chemical space, compounds were fragmented using the Retrosynthetic Combinatorial Analysis Procedure (RECAP) algorithm [31], yielding 2.4 million unique terminal fragments **(Figure 1b; Supplementary Figure 1c)**. For rapid and scalable querying of these fragments, we implemented a TRIE (prefix tree) data structure called as MagnetDB TRIE, comprising over 2.4 million branches and 5.7 million nodes **(Supplementary**

Figure 1d), facilitates near-instantaneous identification of terminal fragments within query compounds **(Figure 1c)**. Benchmarking established the TRIE's exceptional efficiency in fragment retrieval, with O(n) time complexity. Empirical time complexity (Big O) closely matched theoretical O(n) performance, significantly outperforming brute-force $O(n^2)$ substructure matching **(Supplementary Figure 1e),** which was orders of magnitude faster (enabling retrieval of 1,000 fragments in just 12 seconds versus ~18 hours) than the standard brute-force search **(Figure 1d, e)**. Crucially, the TRIE implementation preserved essential Simplified Molecular Input Line Entry System (SMILES) chemical fidelity, retaining bond/atom order, chirality, and ensuring chemical fidelity **(Figure 1f)**. Its robustness was confirmed across varying fragment sizes and complexities, with consistently high retrieval precision and reproducibility across multiple runs **(Figure 1g; Supplementary Figure 1f).** Fragment distribution analysis revealed that the majority of fragments clustered within bin 2 (≤50 atoms), balancing structural diversity with computational tractability **(Figure 1h)**. Correspondingly, search time evaluations confirmed that with increasing fragment size, search efficiency remained robust, highlighting the computational scalability of the approach **(Figure 1i)**. The predominance of small terminal fragments (≤10 atoms) not only enhances computational efficiency in our screening workflow but also provides a rational cutoff for query fragment selection in scoring **(Figure 1j)**. Smaller fragments, with reduced structural complexity and fewer conformations, enable faster and more accurate similarity matching. This size limit focuses scoring on tractable, pharmacologically relevant motifs while avoiding artifacts from larger moieties. This efficient indexing and retrieval system enabled rapid mapping of input compounds to relevant targets within MagnetDB, forming the backbone of SynGlue's data-driven generation and annotation workflows. Finally, the data-driven module systematically processes and scores query molecules by integrating multiple databases, SMILES conversion, hierarchical fragment mapping, MCS-based scoring, and final classification into distinct molecular types **(Supplementary Figure 1b).**

SynGlue employs a rigorous dual-stage matching strategy for mapping query compounds to known targets through their terminal fragments **(Figure 1k, Supplementary Figure 1c)**. The first stage involves a rapid exact-match search using the optimized MagnetDB TRIE, which enables efficient identification of terminal fragments. The TRIE structure supports O(n) insert, search, and delete operations **(Supplementary Figure 1e)**. The initial benchmarking results

reveal that fragment search times slightly increase with fragment length and remain consistent across 1,000 random fragments, demonstrating the TRIE's mapping efficiency **(Supplementary Figure 1f-g)**. The comparison across datasets and percentile-based fragment classes indicates that both search and prefix match times remain efficient and consistent, highlighting the approach's robustness for diverse fragment libraries **(Supplementary Figure 1h)**.

This is followed by a secondary, more granular similarity assessment leveraging both chemical fingerprints, such as ECFP4 [32], MACCS keys, and Maximum Common Substructure (MCS) analysis (using the FMCS algorithm [33]). Stringent similarity thresholds were applied at this stage to ensure only high-confidence fragment-target associations are retained **(Figure 1k)**. The systematic selection of terminal fragments from multiple breakdown levels of the parent molecule ensures a comprehensive coverage of structural diversity, leading to more accurate fragment-based search and mapping **(Supplementary Figure 1i)**. This layered approach yields high-resolution mappings annotated with functional group information and MCS-based visualizations of shared chemical features. To further enhance interpretability, we introduced a novel classification framework that categorizes compounds based on their mapped interaction profiles: type 1 (monovalent, monotarget), type 2 (monovalent, multitarget), type 3 (multivalent, monotarget), and type 4 (multivalent, multitarget) **(Figure 1l)**. This systematic classification, enabled by integrating fingerprinting and MCS analysis within the SynGlue framework, facilitates the accurate characterization of diverse multi-target compounds, including structurally complex molecules like PROTACs. In summary, the combination of data-driven and structure-guided approaches, coupled with a precise dual-stage thresholding strategy, provides a robust and scalable foundation within SynGlue for the accurate identification, analysis, classification, and design of multi-targeting molecules.

SynGlue Accurately Identifies and Classifies Multi-Target Compounds

To rigorously evaluate the performance and accuracy of SynGlue's data-driven module, we conducted a comprehensive benchmarking with the known target prediction tools. First, we assembled a benchmark (validation) dataset comprising 6,935 diverse compounds with reported multi-targeting profiles. This set included known drugs from DrugCentral (DC; 33.7%), known PROTACs from PROTACpedia/PROTACDB [34] (PROTAC*; 64.5%), and other chemically induced dimerization (CID) compounds, molecular glues, Autophagy-targeting chimeras (AUTAC), and multitarget molecules (MTMs) from literature (Others; 1.8%), ensuring broad structural and functional diversity **(Figure 2a; Supplementary Figure 2a-b; Supplementary**

Table 3). We mapped and validated the SynGlue-predicted targets against this comprehensive ground truth dataset **(Supplementary Figure 2c).** The ground truth targets were cross-verified against MagnetDB, confirming the presence of 856 DrugCentral, 107 PROTAC*, and 65 Other targets in the database **(Supplementary Figure 2d).** Principal Component Analysis (PCA) based on chemical descriptors confirmed structural heterogeneity, particularly between the DrugCentral and PROTAC* compound clusters **(Figure 2b)**. Functional group analysis highlighted enrichments specific to PROTACs, such as ether and amine groups, reinforcing their distinct chemical signatures **(Figure 2c)**. We assessed SynGlue's ability to correctly identify known ligand-target interactions within this dataset using hit ratio and miss ratio metrics **(Figure 2d).** The data-driven module consistently demonstrated exceptional performance, achieving a median hit ratio near 1.0 and a miss ratio approaching zero across the datasets, exemplified by the stallimycin-ESR1 interaction (hit) and broprimine-XDH interaction (miss) **(Figures 2e-f)**. SynGlue's consistently high accuracy across DrugCentral, PROTAC*, and other datasets is reflected in its robust and reliable target prediction; for example, it targets SLC6A2, FRAP, and DNMT1 **(Supplementary Figure 2e)**. This performance underscores both the quality of SynGlue's algorithms and their high accuracy in recognizing known interactions, reflecting the comprehensiveness of the underlying MagnetDB. Beyond simple hit identification, SynGlue successfully applied its classification framework to categorize 4,621 compounds (66.7%) into the four polypharmacological types (types 1-4), showcasing its utility in characterizing compound interaction profiles **(Figure 2g, h; Supplementary Table 2).** The framework proved particularly effective in classifying structurally complex entities, including PROTACs and multitarget molecules (MTM) **(Figure 2i)**. Notably, SynGlue demonstrated particular sensitivity in handling structurally complex molecules, although fragmentation limitations exist for highly rigid macrocyclic scaffolds like Cyclosporine.

To further benchmark, we compared SynGlue's capabilities to established target prediction tools: LigAdvisor [35], SwissTargetPrediction [36], and SuperPred [37] using this curated validation dataset. We evaluated each tool based on hit and miss ratios across cluster-representative molecules derived via t-SNE embedding **(Figure 2j)**. Next, we evaluated target overlap within the benchmark dataset to ensure a fair and robust comparison of SynGlue's performance against existing prediction tools **(Figure 2j)**. A key differentiator is SynGlue's significantly larger underlying dataset; MagnetDB contains substantially more ligand entries (e.g., ~2.8M vs. ~28k-376k) and target entries (e.g., ~20k vs. ~2.3k-18k) than the databases used by the compared tools **(Figure 2k-m).** Due to limitations in batch processing capabilities of web server-based

tools such as LigAdvisor, SuperPred, and SwissTargetPrediction, we performed the comparison using a representative subset of 100 molecules selected from the benchmark dataset, ensuring coverage across chemical space and polypharmacological types **(Supplementary Figure 2f-h)**. While the compared tools often exhibited biased predictions or diminished accuracy for multi-target compounds, SynGlue maintained balanced and high-confidence mappings with significantly higher hit ratios, particularly for multi-target compounds incorporating flexible linkers or complex scaffolds, cases where conventional tools frequently showed reduced accuracy or higher false negative rates **(Figure 2o, Supplementary Table 4).** Together, these benchmarking results validate SynGlue's data-driven module's high accuracy and functional versatility in identifying known ligand-target interactions and classifying multi-target compounds. Furthermore, comparative analysis highlights SynGlue's advantages stemming from its comprehensive data foundation and robust algorithms, positioning it as a powerful tool for multi-target compound analysis.

SynGlue Enables Designing of PROTACs and Multi-Target Molecules

To evaluate SynGlue's design capabilities, we extended its application from analyzing known compounds to the *de novo* design of novel multi-targeting therapeutics. To illustrate the workflow, we first focused on PROTAC design targeting the human Androgen Receptor (AR), a clinically validated oncology target with abundant validation data **(Figure 3a, b)**. The integrated PROTAC design workflow combines warhead, E3 ligase, and linker selection while ensuring structural diversity and synthetic feasibility **(Supplementary Figure 3a)**. Additional characterization of atom count distributions and linker property analyses offers a comprehensive view of component flexibility and structural diversity and aids in understanding PROTAC components and their properties for rational design **(Supplementary Figure 3b-h)**. Employing the data-driven approach, SynGlue identified 87 potential AR-binding warheads based on interaction data from MagnetDB and physicochemical properties, leading to the selection of a validated candidate for warhead (CHEMBL146794) **(Figure 3b; Supplementary Figure 5a; Supplementary Table 5)**. The Generator module then produced potential PROTAC candidates by coupling this warhead to a known XIAP E3 ligase ligand **(Figure 3b)**. Concurrently, we also used a structure-guided design workflow utilizing the AR crystal structure (PDB ID: 1E3G) as input in the recently developed Gcoupler tool [28] to design 404 *de novo* cavity-specific ligands, illustrating the complementary strength of SynGlue's dual workflows **(Figure 3b; Supplementary Figure 5b; Supplementary Table 5)**. To finalize the warhead selection, we employed the Gcoupler Authenticator module, which evaluated candidate ligands based on their

molecular-level interaction profiles. The selected warhead was then used by SynGlue's Generator module by incorporating optimized linker parameters and the same XIAP E3 ligase, generating a distinct set of PROTAC candidates. Comparative analysis revealed that the data-driven approach yielded analogs closely resembling experimentally validated AR PROTACs with high chemical similarity (Tanimoto similarity scores). In contrast, the structure-guided approach produced more chemically diverse candidates, exploring underrepresented and novel chemical space while still retaining relevance to known structures. This highlights the complementary strengths of SynGlue's dual modalities for PROTAC design for expanding ligand diversity and novelty. Beyond PROTACs, SynGlue's integrated platform is adept at designing other MTMs **(Figure 3c).** Leveraging the data-driven module, MagnetDB identifies suitable warhead fragments for multiple desired targets based on high interaction coverage (≥75%). Next, we demonstrated SynGlue's capacity to design non-degrading multitarget molecules. For the EGFR/CDK1 dual-target case, SynGlue identified validated warheads for EGFR (CHEMBL939) and CDK1 (CHEMBL4442620) and subsequently designed candidate MTMs predicted to retain binding compatibility with both targets via their respective terminal fragments **(Figure 3c)**. A similar successful outcome was achieved designing MTMs for the PRKX/DRD1 pair using identified warheads (CHEMBL2152768 and CHEMBL1201356) **(Figure 3c; Supplementary Table 6).** These examples underscore SynGlue's versatility in designing diverse MTM chemotypes, as highlighted by the distributions of their linker properties and chemical taxonomy analyses **(Supplementary Figure 3i-k)**.

To bridge computational design with experimental validation, we applied SynGlue to design PROTACs targeting Bromodomain-Containing Protein 4 (BRD4), a prominent oncology target, for degradation *via* recruiting the von Hippel-Lindau (VHL) E3 ligase. Both the data-driven (using MagnetDB-selected warheads) and structure-guided (using PDB ID: 7REK and Gcoupler) approaches were employed, followed by PROTAC design using SynGlue's Generator module **(Figure 3d)**. We commenced the design of BRD4 targeting PROTACs by interrogating MagnetDB for chemical fragments associated with BRD4 and retained an initial cohort of 1,498 warhead candidates that exhibited a stringent matching threshold (>75%). To ensure a high degree of target specificity, these fragments were subjected to cross-referencing against analogous fragments associated with the closely related Bromodomain isoforms BRD2, BRD3, and BRD9 **(Figure 3e)**. This comparative analysis facilitated the exclusion of 14 promiscuous

fragments lacking exclusive association with BRD4, thereby establishing a refined library of BRD4-selective warhead entities. Subsequently, these entities underwent rigorous filtering based on their predicted physicochemical property ranges that are likely to possess drug-like characteristics. Specifically, molecules exhibiting calculated logP values within the 0-5 range, a criterion associated with acceptable aqueous solubility and cell membrane permeability, were prioritized (n=326). Within this pre-selected set, molecules demonstrating a Quantitative Estimate of Drug-likeness (QED) scores ≥ 0.5 were retained, culminating in a subset of 74 high-probability drug-like warhead candidates. Further refinement based on synthetic accessibility involved the selection of molecular entities possessing Synthetic Accessibility (SA) scores ≤ 5, which further narrowed the candidate pool to 43 synthetically tractable warheads. To ascertain structural novelty, these MagnetDB-derived BRD4 warheads were juxtaposed against a curated compendium of 42 known BRD-targeting warheads (spanning BRD2, BRD3, BRD4, and BRD9) extracted from PROTACDB 3.0. Employing ECFP4-based Tanimoto similarity metrics, only candidate warheads exhibiting similarity scores below a stringent threshold of 0.5, indicative of significant structural divergence, were retained (n=22) **(Figure 3e, f)**. Finally, we applied hierarchical clustering and dendrogram analysis to assess chemical diversity and define fragment families to improve viability; the selected warheads were further optimized using RDKit by evaluating detailed atom-level descriptors and reactive substructures **(Figure 3g)**. RDKit-guided optimization ensured that the final set of BRD4 warheads not only retained their novelty and drug-like properties but were also more synthetically accessible and structurally robust. These optimized warheads formed the foundation for downstream PROTAC design and development. To ensure biological relevance, top-scoring warheads were docked into the BRD4 bromodomain structure (PDB: 3MXF), and their binding energies were evaluated **(Figure 3h-i; Supplementary Figure 5c)**. Warheads showing favorable docking scores and stable cavity engagement were shortlisted based on interaction profiles with key residues (ASN140(A),

MET149, LEU92, TYR97), guiding the selection of structurally compatible and synthetically accessible candidates **(Figure 3j; Supplementary Figure 6a-k)**.

To prioritize the resulting candidates for resource-intensive downstream validation, we implemented Prioritizer module a multi-parameter strategy that balanced binding energy, structural novelty, pharmacokinetic properties, and synthetic feasibility. A linker-based multiclass model was developed and rigorously evaluated to predict the potency of PROTAC with varied linkers **(Supplementary Figure 4a-n)**. Building upon this, we also developed attention-based regression models utilizing publicly available PROTAC activity data for predicting DC_{50} (half-maximal degradation concentration) and D_{max} (maximum degradation) values using PROTAC SMILES as inputs. Both models showed acceptable predictive performance based on $R^2 > 0.8$ and MAE < 0.3 in the cross-validation metrics **(Figure 3k-l; Supplementary Figure 5d-h)**. We further refined the designed PROTACs selection by assessing key physicochemical filtering based on drug-likeness metrics, including molecular weight, clogP, and halogenation, heteroatoms, TPSA, and solubility **(Supplementary Figure 5j)**. This rigorous filtering process led to the selection of two promising isomeric BRD4-specific PROTAC candidates (Compound 1, *aka* AU-14555), and (Compound 2, *aka* AU-14849), predicted to have distinct degradation potentials despite marginal structural differences, for synthesis and experimental testing, underscoring the model's resolution in capturing nuanced activity differences **(Figure 3m; Supplementary Figure4f; Supplementary Table 6)**. To experimentally validate our top-prioritized, aforementioned SynGlue-designed BRD4 PROTACs, we synthesized the lead candidates: Compound 1, Compound 2, and the inactive stereoisomer of Compound 1. We confirmed their structures using ^{1}H and ^{13}C Nuclear Magnetic Resonance, their molecular masses using Liquid Chromatography-Mass Spectrometry, and their molecular formulas using High-Resolution Mass Spectrometry **(Supplementary Figure 7a-i)**. Furthermore, our High-Performance Liquid Chromatography analysis confirmed high chemical purities for these synthesized compounds, indicating the reliability of our synthetic routes. Both the physicochemical and the experimentally elucidated pharmacokinetic profiles of two representative PROTACs highlighted variations in potency, permeability, plasma protein binding, and *in vivo* pharmacokinetics. Notably, Compound 1 exhibited superior bioavailability and target engagement compared to Compound 2, underscoring the impact of linker properties on PROTAC performance **(Supplementary Figure 5k)**. Collectively, these case studies demonstrate SynGlue's robust capability to perform *de novo* design of both PROTACs and MTMs using integrated data-driven and structure-guided generative workflows. Furthermore, its

dual design strategies, which are data-driven and structure-guided, combined with an AI-based prioritization pipeline, highlight SynGlue's utility in efficiently exploring chemical space and identifying promising multi-targeting candidates for subsequent experimental validation.

SynGlue Design Translates to Potent *in vitro* Activity and *in vivo* Anti-Tumor Efficacy.

To evaluate the BRD4-mediated anti-cancer properties of our SynGlue-designed PROTACs, we first performed anti-proliferative assays. Compound 1 exhibited potent antiproliferative activity across diverse cancer cell lines, achieving 50% growth inhibition at nanomolar concentrations (1.1 nM in MV-4-11, 3 nM in VCaP), whereas Compound 2 showed generally lower potency **(Figure 4a)**. Underscoring Compound 1's specificity, its inactive stereoisomer displayed significantly reduced antiproliferative potency in VCaP cells, achieving 50% growth inhibition at 1406 nM versus 4 nM for Compound 1 **(Supplementary Figure 8a-b)**. We then assessed protein degradation using Western Blot analyses. Both Compound 1 and Compound 2 demonstrated potent, dose-dependent degradation of BRD4 in prostate cancer cell lines (VCaP and LNCaP) after 6 hours; Compound 1 also degraded BRD2 and BRD3 in these VCaP and LNCaP cells **(Figure 4b-c)**. Notably, in HeLa cells, Compound 1 showed selective degradation of BRD4 over BRD2 and BRD3 isoforms **(Figure 4c)**. In MV-4-11 cells, Compound 1 induced potent BRD4 degradation, achieving 50% degradation at a concentration of 0.193 nM and outperforming Compound 2 (50% degradation at 0.228 nM) **(Supplementary Figure 8c)**. This degradation by Compound 1 was confirmed to be ubiquitin-proteasome pathway-mediated, as co-treatment with the proteasome inhibitor MLN2238 rescued BRD4 levels **(Figure 4d)**. Consistent with its reduced antiproliferative effect, the inactive isomer showed minimal to no BRD4 degradation in MV-4-11 cells **(Figure 4e; Supplementary Figure 8a)**. These top-target compounds also possessed favorable physicochemical properties, including reasonable aqueous solubility (4.53 µM) **(Figure 3m)**. Next, based on its promising *in vitro* profile, characterized by potent and specific activity of Compound 1, which was clearly differentiated from its inactive stereoisomer **(Supplementary Figure 8a-c).** The results showed potent degradation by Compound 1, underscoring its specificity, its inactive stereoisomer displayed significantly reduced antiproliferative potency, achieving 50% growth inhibition at 1406 nM versus 4 nM for Compound 1 in VCaP cells and 275 nM versus 2 nM for Compound 1 in MV-4-11 cells. Next, based on its promising *in vitro* profile, we advanced Compound 1 for *in vivo* evaluation. Our intravenous pharmacokinetic studies in CD-1 mice revealed a dose-proportional exposure (measured as area under the curve) across a 5-15 mg/kg range with a plasma half-life

of approximately 0.83 hours **(Figure 4f; Supplementary Figure 8d; Supplementary Table 9)**. We next performed tolerability assessment and found that Compound 1 was well-tolerated, as we observed no treatment-related adverse effects up to 10 mg/kg/day in a 14-day intravenous maximum tolerated dose study in CD-1 mice **(Figure 4g; Supplementary Table 9 for MTD data)**. Consistent with this, no significant body weight changes or adverse clinical signs were noted during subsequent efficacy studies **(Figure 4h)**.

Next, we utilized the MV-4-11 cell line-derived xenograft mouse model for the *in vivo* validation of the anti-tumor activity of Compound 1 **(Figure 4i)**. Our results suggest statistically significant, dose-dependent tumor growth inhibition, achieving 47% TGI at 5 mg/kg (q.d.), 120% TGI at 10 mg/kg (q.d.), and 75% TGI at 10 mg/kg (q48h) **(Figures 4j-l; Supplementary Figure 8e; Supplementary Table 9**. Administering Compound 1 at 10 mg/kg every day proved to be the optimal regimen. This schedule achieved a 2.2-fold higher concentration in the tumor compared to the plasma, demonstrating superior tumor targeting. It successfully maintained high tumor levels (246 ± 35 ng/g) while reducing overall systemic exposure (110 ± 19 ng/g). This indicates a favorable pharmacokinetic profile that maximizes efficacy on the tumor while minimizing the risk of off-target side effects **(Figure 4m)**. Furthermore, the anti-tumor activity strongly correlated with *in vivo* BRD4 degradation and downregulation of downstream c-Myc levels in excised tumor tissues **(Figure 4n)**. Sustained target engagement was confirmed by significant BRD4 degradation in tumor lysates up to 24 hours post-treatment **(Supplementary Table 9)**. In summary, our comprehensive preclinical evaluation demonstrates that the SynGlue-designed PROTAC, Compound 1, is a potent and selective BRD4 degrader. We found it to possess favorable drug-like properties, pharmacokinetics, and in vivo tolerability, all of which contribute to its significant, mechanism-driven anti-tumor efficacy in a preclinical cancer model. We believe these findings validate SynGlue's ability to guide the discovery of functionally active multi-target molecules with therapeutic potential.

DISCUSSION

The development of multi-targeting therapeutics, particularly via targeted protein degradation strategies such as PROTACs, represents a significant advancement in medicinal chemistry and pharmacology [4,10,11,14]. These modalities hold the potential to address complex diseases more effectively than conventional single-target drugs by modulating multiple signaling pathways or eliminating disease-driving proteins. However, the rational design of potent, selective, and drug-like multi-target molecules remains a substantial bottleneck due to the intricate structure-activity relationships involved and the expansive chemical space, especially in the context of linker optimization and ternary complex formation [4,14,16]. Therefore, bridging the gap between computational design and experimental success is paramount.

In this study, we introduced SynGlue, a novel next-generation generative AI platform tailored to overcome core challenges of multi-targeting drug discovery. SynGlue distinguishes itself through its synergistic integration of large-scale, ligand-target interaction data curated within the comprehensive MagnetDB with structure-guided design principles within a unified generative framework. This combination enables SynGlue to predict target interactions, optimize molecular structures, and facilitate the *de novo* design of synthesis-ready multi-target compounds. We demonstrated the robustness of this foundation through meticulous evaluation and benchmarking. The strength of SynGlue lies in its meticulously curated large database, MagnetDB, which provides broad coverage of 6.37 million experimentally validated protein-ligand pairs. This resource, coupled with a highly efficient TRIE-based architecture, enables rapid and highly efficient fragment mapping. The dual-stage matching strategy begins with an exact-match search, followed by a detailed similarity assessment using chemical fingerprints and Maximum Common Substructure (MCS) analysis, ensuring accurate target identification and fragment association, and significantly superior hit rates compared to conventional target prediction tools. Additionally, type-specific annotations and medical chemistry-aware functional group analysis are incorporated using custom SMARTS-based rules. This architecture underpins SynGlue’s core advantage: transitioning from retrospective analysis to the *de novo* design of multi-target compounds, enabling the exploration of a vast chemical space with high accuracy. Performance benchmarks confirmed SynGlue's exceptional accuracy in identifying known interactions within diverse chemical datasets, significantly outperforming other methods, due to its extensive data foundation and highly refined and robust algorithms.

We successfully applied SynGlue integrated modules to design chemically diverse and functionally relevant candidates across multiple modalities, including PROTACs targeting AR and BRD4 and MTMs directed at EGFR/CDK1 and PRKX/DRD1. Crucially, the platform demonstrates a complete pipeline from AI-driven design to preclinical validation. Using SynGlue, we designed novel BRD4-VHL PROTACs, prioritized candidates using a combination of AI-based activity prediction and molecular modeling, docking, and physicochemical filtering; we prioritized a lead BRD4-targeting PROTAC and performed its preclinical validations using an assay of *in vitro*, biochemical, and *in vivo* assays.

The existing computational tools for PROTAC/MTM design focus on specific aspects, such as ternary complex modeling (e.g., PRosettaC [20]), linker designing (e.g., DiffPROTACs [21]), or target prediction from single molecules (e.g., SwissTargetPrediction [36], LigAdvisor [35], and SuperPred [37]). While valuable, these tools may lack the integrated data scope, the generative versatility across different multi-targeting modalities, or the demonstrated end-to-end validation culminating in *in vivo* efficacy. SynGlue's ability to leverage both large-scale interaction data and structural information within a generative AI framework allows for a more holistic design process, balancing chemical relevance with novelty and predictive power. The successful prioritization strategy further highlights its utility in navigating the vast chemical space and focusing experimental efforts on the most promising candidates. SynGlue offers a cohesive, multi-layered pipeline capable of producing and prioritizing validated multi-target degraders. Despite its demonstrated successes, certain limitations remain. SynGlue's data-driven components rely on the availability and quality of public interaction data within MagnetDB. While the platform handles diverse chemotypes well, highly rigid or unusual scaffolds can pose challenges for the current fragmentation and matching algorithms. The accuracy of the AI-driven activity predictions depends on the scope and quality of the training data available for specific endpoints like DC_{50}/D_{max}. Future work will focus on expanding MagnetDB, refining the fragmentation algorithms, incorporating more sophisticated structural modeling techniques (e.g., advanced ternary complex prediction), refining fragmentation methods for macrocyclic systems, and continuously improving the predictive models through active learning cycles incorporating new experimental data.

In conclusion, SynGlue represents a transformative platform for the rational design of multi-targeting therapeutics. By integrating extensive large-scale biological resources, efficient algorithms, fragment-aware molecular designing, structure-guided constraints, and AI-based activity prediction with a validated end-to-end pipeline from *in silico* modeling to *in vivo* efficacy,

SynGlue provides a scalable framework for rapidly identifying and optimizing degrader compounds.

MATERIAL AND METHODS

Construction and Preparation of the MagnetDB Interaction Database

MagnetDB was constructed by systematically aggregating and curating experimentally validated protein-ligand interaction data from six major public repositories: DrugBank [24] (v5.1.10), BindingDB [23] (downloaded 28-May-2023), ChEMBL[22] (v33, released 31-May-2023), STITCH[25] (v5.0), BioSNAP - ChG-InterDecagon, ChG-Miner, ChG-TargetDecagon (August 2018 release)[26], and the Small Molecule Suite[27]. Database-specific extraction pipelines were employed for each database: XML parsing for DrugBank; SDF file processing for BindingDB; API-based species-restricted querying for ChEMBL; extraction from pre-compiled interaction networks for BioSNAP and STITCH; and mapping within the Small Molecule Suite. This initial collection yielded a total of 37.55 million interaction records. The aggregated dataset underwent rigorous preprocessing to ensure data quality, consistency, and relevance. Ligands were standardized to canonical SMILES format using OpenBabel (v3.0.0), and protein identifiers were converted to UniProt accession numbers wherever possible. Redundant, inconsistent, or ambiguous entries were resolved using a combination of automated mapping and manual curation. This process harmonized Compound names, target names, and assay descriptions across different databases, ultimately consolidating entries based on unique ligand-target pairs. The dataset was initially consolidated to 35.21 million entries after basic identifier normalization. A critical filtering step was applied to retain only high-confidence, experimentally validated direct binding interactions, excluding predicted or indirect associations. To ensure species specificity, data were filtered for organisms *Homo sapiens* (NCBI Taxonomy ID: 9606), *Mus musculus* (NCBI ID: 10090), and *Saccharomyces cerevisiae* (NCBI ID: 4932). The dataset was reduced to 21.39 million species-specific entries. High-confidence pairs were retained, and duplicates were removed, yielding the final curated MagnetDB, which comprises 6.37 million unique, experimentally validated, protein-ligand interactions sourced from the aforementioned databases. This dataset includes 1.94 million unique chemical compounds interacting with 20,129 unique protein targets. Each entry in MagnetDB is annotated with essential information, including a unique interaction ID, ligand canonical SMILES, ligand name, target UniProt ID, target name, source database(s), assay information (where available), and species.

To enable scalable and chemically precise fragment mapping, we implemented a custom TRIE (prefix tree) data structure indexed by reversed canonical SMILES strings with a terminal symbol ($) to denote fragment endpoints. All compounds were systematically fragmented using

the Retrosynthetic Combinatorial Analysis Procedure (RECAP) algorithm is implemented via the RDKit (v2023.09.02). RECAP applies chemically relevant cleavage rules to identify synthetically accessible substructures. We focused on terminal fragments generated through RECAP decomposition by constructing hierarchical fragmentation maps and retaining only one cleavage event per level to ensure each terminal fragment represents a unique, meaningful substructure with a clear point of attachment, marked by an asterisk (*). This approach ensured the preservation of terminal sites relevant for ligand-target interaction mapping while excluding non-terminal or multiple cleaved fragments. Terminal fragments were specifically targeted because they are most likely to retain the binding motifs and anchoring functionalities critical for direct protein-ligand interactions, thus ensuring the biological relevance of the dataset while avoiding the chemical ambiguities introduced by non-terminal or multiple cleaved fragments. Terminal fragments generated from RECAP decomposition were inserted into the TRIE and linked to corresponding MagnetDB identifiers via a hash map. The TRIE design preserves stereochemistry, bond order, and atom connectivity, allowing chemically meaningful retrieval. Canonical SMILES ensured that for any chemical structure, a unique and deterministic SMILES string is generated, reflecting the exact 3D molecular structure, including stereocenters and cis/trans isomerism within a 1D string representation. A TRIE data structure was implemented in Python using optimization to enable efficient searching of these fragments based on their canonical SMILES strings, overcoming the limitations of computationally expensive brute-force substructure searching. The TRIE supports efficient prefix-based and exact-match queries with O(n) complexity, offering significant performance advantages over conventional substructure searches. Approximately 2.4 million fragments from the curated MagnetDB were yielded, and the full TRIE structure was constructed in ~3.5 minutes on a standard CPU (256G System memory, Intel(R) Core(TM) i9-10900X CPU). For downstream analysis, both the TRIE and fragment-Magnet ID mappings were serialized using Python’s pickle module. This implementation enables rapid, structure-aware fragment matching across large chemical space within the SynGlue platform.

TRIE Performance Benchmarking

The TRIE data structure implemented in SynGlue was rigorously benchmarked to evaluate its efficiency and scalability, focusing on insertion time, search time, and prefix matching time. Two distinct fragment sets were employed for testing: (a) A random subset of 1,000 terminal fragments from MagnetDB, and (b) Fragments representing the 10th and 90th percentiles of structural diversity across MagnetDB's constituent sources. For each dataset, average insertion

and search times were recorded, and retrieval accuracy was evaluated using retrieval rate metrics. Empirical time complexity (Big O notation) was compared against theoretical expectations and brute-force substructure matching. Performance was further assessed across varying fragment lengths to evaluate structural scalability **(Figure 1g-j)**. Performance was further evaluated across varying fragment lengths, confirming the TRIE's scalability.

SynGlue Data-Driven Analysis Workflow

SynGlue employs a fragment-based analysis pipeline to rapidly identify and classify potential protein targets from input query molecules. This data-driven approach decomposes molecules into fragments and maps them to a curated target-fragment index for efficient target association. To analyze query molecules (input as SMILES), identify potential protein targets via fragment mapping against MagnetDB, and classify their polypharmacological profile, facilitated rapid and scalable classification of compounds. The SynGlue pipeline begins with canonicalization of SMILES via OpenBabel (v3.0.0) and RECAP-based fragmentation using RDKit (v2023.09.02), then decomposition of molecules into terminal fragments that are mapped with the MagnetDB TRIE for exact match retrieval, enabling fast, exact fragment matching and leveraging the curated depth of MagnetDB for reliable target associations. To ensure robust validation and enhance confidence in predicted interactions, the pipeline integrates multi-modal similarity scoring that combines ECFP4 (1024-bit) and RDKit topological fingerprints (2048-bit) alongside Tanimoto and Overlap coefficients. Additionally, it incorporates Maximum Common Substructure (MCS) analysis using the rdFMCS algorithm (RDKit version v2023.09.02) to identify the largest shared substructure between query fragments and database entries. These layered similarity metrics ensure robust identification of biologically meaningful matches. The inclusion of stringent filters such as : >10 atoms per query molecule; MCS overlap threshold ≥ 25% query and ≥ 75% database fragment; and the exclusion of trivial/monatomic fragments (e.g., *OH, *O). These layered criteria improved the biological and chemical relevance of

matched fragments with high similarity scores and added confidence. High-confidence matches are functionally annotated. Based on fragment-to-target mappings and ligand mapped fragments, molecules are systematically classified into four polypharmacological categories: type 1 (monovalent, monotarget), type 2 (monovalent, multitarget), type 3 (multivalent, monotarget), or type 4 (multivalent, multitarget) **(Figure 1l)**. This classification framework offers immediate insights into molecular mechanisms that are particularly valuable for prioritization strategies, especially for multi-targeted drug design. Final output includes matched fragments, associated targets, similarity scores, functional annotations, and classification metadata, which collectively inform downstream prioritization and design workflows. Overall, SynGlue's integration of efficient fragment search, rigorous similarity assessment, and structured classification enables informed, high-throughput decision-making in multi-ligand therapeutic design.

Performance Evaluation and Comparative Benchmarking of Analysis Workflow

The SynGlue data-driven workflow was applied to a benchmark dataset of 6,935 compounds with known targets. Predicted targets were compared against ground truth targets (excluding known targets not present in MagnetDB for exact match analysis). Accuracy was quantified using hit rate and miss rate.

$$\text{Hit Rate} = \frac{TP}{TT}; \quad \text{Miss Rate} = \frac{FN}{TT}$$

where;
TP = True Positives
FN = False Negatives
TT = Total Known Targets

Performance was assessed across different compound categories (Drug Central, PROTAC, and Others) **(Figure 2e)**. SynGlue was further benchmarked against LigAdvisor, SwissTargetPrediction, and SuperPred using a representative subset of 100 structurally diverse molecules (selected via ECFP4 clustering) **(Supplementary Figure S2-d** for clustering

methodology). Target predictions were obtained from all tools for this subset. These metrics enabled direct comparison of performance across all tools.

SynGlue Generative Design Workflow

The SynGlue Generator, for *de novo* PROTAC and MTM design, utilizes the Link-INVENT architecture (encoder-decoder RNN with LSTM units: embedding size 256; 3 hidden layers of 512 units), trained via reinforcement learning. A warhead prioritization module filters MagnetDB fragments based on physicochemical and structural constraints (MW 200-500 Da; LogP 0-5; TPSA 20-120 $Å^2$; rotatable bonds 0-10; rings 0-5; aromatic proportion <0.45; SA ≤ 6; Csp^3 0.3-0.8) and required reactive groups (Cl, Br, I, NH_2). This ensures that only realistic, drug-like warheads with appropriate reactive moieties are considered for downstream design. Candidate warheads are then mapped to compatible E3 ligases (from PROTAC-DB) based on similar property filtering and known compatibility based on property-matched filtering from PROTAC-DB, typically yielding 5-30 pairs per input warhead **(Supplementary Figure 3a-h)**. For compatible warhead-E3 pairs, the SynGlue Generator designs linkers, optimizing a multi-objective scoring function (a weighted geometric mean of linkers that balances hydrogen bond donors (HBDs), ring content, and graph length). Although explicit quantitative results for the number of generated molecules and final score distributions are detailed in the Results section **(Figure 3a-c, Table S4)**, this scoring framework was empirically chosen to ensure designed linkers possess biologically relevant features such as optimal flexibility, balanced hydrophobicity/hydrophilicity, and conformational adaptability to fit within the spatial constraints of the ternary complex. These features directly impact ternary complex stability and pharmacokinetics, making them critical for high-quality, biologically meaningful designs. High-scoring molecules (score ≥ 0.8) are annotated in SynGlue with a unique identifier and stepwise design metadata, including scaffold, SMILES, computed properties (molecular weight, clogP,

TPSA, Halogen Count, Heteroatom Count, LogS ESOL), and total score in the final output. SynGlue does not explicitly penalize redundancy in generated linkers but relies on inherent diversity from reinforcement learning exploration. Generative runs typically produce 2000-4000 PROTACs per warhead-E3 pair, with ~10 minutes runtime on a 12-core CPU (128 GB RAM). SynGlue was developed with a modular architecture, enabling future integration of additional generative or predictive modules as plug-and-play components. For instance, generative constraints such as ternary complex energy prediction, proteome-wide off-target modeling, or disease-context filtering based on expression data can be incorporated into the scoring pipeline. This design allows SynGlue to evolve into a general-purpose engine for the conditional generation of multi-functional therapeutics, adaptable to diverse target classes modalities (e.g., molecular glues, bifunctional degraders), and therapeutic areas beyond oncology.

SynGlue Data Driven Used for Multitarget Molecules

SynGlue's data-driven design module streamlines the creation of multitarget molecules (MTMs) by leveraging MagnetDB-derived fragments that demonstrate a high degree of matching (fragment coverage over 75%) to target binding sites, ensuring strong baseline affinity. These candidate warheads are further refined through a stringent filtering process incorporating critical drug-likeness and developability parameters, including molecular weight (200-500 Da), logP (0-5), topological polar surface area (TPSA, 20-120 $Å^2$), number of rotatable bonds (0-10), aromatic ring content (<0.45), synthetic accessibility (SA ≤ 6), and Csp^3 fraction (0.3-0.8). This comprehensive profile ensures that the selected warheads possess optimal solubility, permeability, and synthetic feasibility while maintaining a balanced conformational flexibility essential for engaging multiple binding sites simultaneously. After the target warhead selection,

SynGlue's Generator module is used for the design and optimization of linkers to maintain the ideal spatial arrangement, chemical stability, and flexibility needed for robust multitarget binding. This integrated approach maximizes cooperative interactions and enhances the functional synergy of the resulting MTMs. The effectiveness of this workflow has been demonstrated in case studies such as CDK1-EGFR and PRKX-DRD1 multitarget molecules **(Figure 3c)**, where SynGlue's data-driven methodology enabled rapid prototyping of MTMs with tailored therapeutic profiles by combining predictive analytics with precise structural alignment.

SynGlue Structure-Guided Design Workflow

For targets that lack sufficient data coverage in MagnetDB, such as novel or less-studied protein targets, SynGlue employs a complementary structure-guided workflow. This approach leverages curated or predicted 3D protein structures (from RCSB PDB or AlphaFold) processed using standard protein preparation pipelines, followed by Gcoupler to identify relevant pharmacophores in druggable binding pockets and ligand design. The structure-guided design is particularly valuable in these data-sparse scenarios because it leverages three-dimensional pharmacophore complementarity to pinpoint targetable features and binding hot spots that might be invisible to purely ligand-based approaches. These pharmacophores guide the *de novo* design, leading to the selection of compatible warheads based on structural complementarity and properties (molecular weight, LogP, halogenation, heteroatoms, and TPSA). Selected warheads are paired with compatible E3 ligases (as in the data-driven workflow). The SynGlue Generator designs linkers that respect the geometric and steric constraints of the identified pockets, ensuring that final candidates are not only synthetically feasible but also structurally compatible with the target's unique binding landscape**.** Final candidates can be refined through molecular docking (AutoDock Vina v1.2) and geometric analyses (RMSD ≤ 2.0 Å from pharmacophore sites) to confirm proper binding orientations and energy. This workflow was applied in the AR-XIAP and BRD4-VHL case studies using RCSB PDB and Gcoupler-based pharmacophore models to generate high-confidence multitarget candidates. **(Figure 4b; Supplementary Table 4,5).** Final molecules are output as SMILES, accompanied by design and scoring metadata for downstream prioritization.

Predictive Prioritization of Designed Compounds

To systematically rank designed BRD4-VHL PROTACs, predictive models for degradation potency were developed: (1) A regression-based attention learning model estimates DC_{50} and D_{max} values using molecular representations and structural embeddings. (2) A linker-based classification model predicts the DC_{50} activity class (high, medium, low) and categorizes linker potency based on the DC_{50} quartiles. The regression framework employed a component-aware attention model, wherein GROVER-derived graph embeddings (4800-dimensional) were independently generated for each PROTAC component warhead, linker, and E3 ligase. An attention mechanism was implemented to learn component-wise weights reflecting the relative contribution of each substructure to overall degradation efficacy. These attention-weighted embeddings were used with a separate GROVER embedding of the full PROTAC molecule and passed through a deep feedforward regressor consisting of three dense layers (512, 256, 128 units) with ReLU activations and dropout (rate = 0.2). The model was trained using mean squared error loss with the Adam optimizer over 200 epochs with early stopping. For comparative analysis, we also trained multiple traditional regressors, including Random Forest, Gradient Boosting, AdaBoost, Ridge, Lasso, ElasticNet, Support Vector Regression, k-Nearest Neighbors, XGBoost, Multilayer Perceptron, Stochastic Gradient Descent, and Decision Trees, using the same GROVER embeddings as input. All regressors were implemented using scikit-learn (v1.3.0), XGBoost (v1.7+), and LightGBM (v3.3+), and evaluated using ten-fold cross-validation with R^2, MAE, and RMSE as performance metrics.

To complement continuous prediction and enable rapid ranking of linker designs, we constructed a multiclass classification model for potency class prediction, using a curated dataset of 2,320 PROTACs with experimentally validated DC□□ values. Linkers were grouped into four potency classes (low, medium-low, medium-high, and high) based on DC□□ quartile thresholds (Q1 = 6.28 ng/mL, Q2 = 54.96 ng/mL, Q3 = 585 ng/mL). Feature extraction was performed using five orthogonal descriptor sets: bioactivity-based Signaturizer (v1.1.14) [38], graph-based GROVER (https://github.com/tencent-ailab/grover) [39], physicochemical Mordred (1.2.0) [40], image-based ImageMol (https://github.com/ChengF-Lab/ImageMol) [41], and language model-based embeddings from ChemBERTa (ChemBERTa-77M-MLM) [42], trained as a multi-class classifier **(Supplementary Figure 4a)**. After data preprocessing and an 80:20 stratified train-test split, feature selection was performed using the Boruta algorithm to identify the most informative variables. A suite of twelve classifiers, including SVM, Logistic Regression, kNN, Random Forest, Extra Trees, SGD, Gradient Boosting, XGBoost, Gaussian Naive Bayes,

Multilayer Perceptron, AdaBoost, and Decision Trees, were trained and evaluated using one-vs-all AUC-ROC, accuracy, and F1-score. Hyperparameter tuning was conducted via GridSearchCV from the scikit-learn (v1.5.2) library, and final models were validated using 10-fold cross-validation. Class probabilities were used to prioritize linkers predicted to fall into high-potency classes with high confidence, supporting targeted design decisions. **(Supplementary Figure 4b-k).**

Chemical Space Analysis of Designed Compounds

The chemical diversity of designed PROTACs was evaluated relative to known PROTACs. ECFP6 fingerprints (radius 3, 2048 bits) were computed (RDKit-2019.03.1). Principal Component Analysis (PCA) was performed on the binary fingerprint matrix using scikit-learn (1.5.2) to visualize and compare the distribution of SynGlue-designed compounds versus known compounds to assess novelty and coverage.

Novel BRD4 Warhead Discovery Pipeline

To identify potent, selective, and structurally novel warheads for BRD4, we implemented a multi-step filtering pipeline combining data-driven selection with stringent physicochemical and synthetic feasibility criteria. We began by retrieving BRD4-associated fragments from MagnetDB and retained those with a high matching percentage (>75% fragment coverage). To ensure specificity, we included related bromodomain targets such as BRD2, BRD3, and BRD9, and removed any overlapping fragments (Tanimoto similarity ≥ 0.9 using ECFP4) were removed, thereby isolating BRD4-specific candidates. The filtered warheads were then evaluated for favorable drug-like properties, retaining those with logP values between 0 and 5 (indicating balanced solubility and permeability), QED scores ≥ 0.5 (suggesting drug-likeness), and Synthetic Accessibility (SA) scores ≤ 5 (indicating ease of synthesis). To further ensure novelty, we sourced known warheads for BRD family targets from PROTACDB (3.0) and computed ECFP4-based Tanimoto similarity against our MagnetDB-derived BRD4 candidates **(Figure 3e)**.

Only those fragments with similarity scores below 0.5 were retained, ensuring that the final set represented structurally novel BRD4 warheads. Finally, we performed hierarchical clustering and dendrogram analysis to visualize chemical diversity and identify distinct warhead families, from which we selected the most promising novel candidates **(Figure 3h-g)**. These candidates were further optimized by assessing synthetic feasibility using RDKit-based synthetic accessibility scoring and structure optimization.

Chemical Synthesis and Characterization

Compound 1 (AU-14555), Compound 2 (AU-14849), and an Inactive Isomer (AU-16914, or negative control were synthesized via multi-step procedures involving peptide couplings (HATU/DIPEA), protecting group manipulations (Boc deprotection with TFA), and purification by CombiFlash® chromatography. Detailed procedures and intermediate steps are provided[43] **(Supplementary Note 1)**. All non-aqueous reactions were performed under nitrogen in oven-dried glassware. Reaction progress was monitored by TLC (Merck silica gel 60 F254, UV 254 nm). Reagents and solvents were commercial grade. Structural elucidation used NMR (^{1}H at 400 MHz, ^{13}C at 101 MHz, DMSO-d□, TMS internal standard for ^{1}H). LC-MS used an Agilent 1100-LC/MSD VL (0.1% formic acid in ACN or water/MeOH). HRMS used a Thermo Orbitrap Q-Exactive Plus LC-MS/MS (ESI). Final purity was confirmed by HPLC (Compound 1: 98.37%; Compound 2: 98.26%; Inactive Isomer: 97.69%). The inactive isomer of Compound 1 (AU-16914) was synthesized as a stereochemical negative control to confirm that the observed biological activity is driven by the specific configuration of the active compound.

Cell culture

VCaP (Vertebral-Cancer of the Prostate), LNCaP (Lymph Node Carcinoma of the Prostate), HeLa (cells of Henrietta Lacks), and MV-4-11 (MonoMac Variant 4-11 cells) cell lines were purchased from the American Type Culture Collection (ATCC, USA). VCaP cells were maintained in Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 10% (v/v) fetal bovine serum (FBS); LNCaP cells were cultured in RPMI-1640 medium supplemented with 10% FBS; MV-4-11 cells were maintained in Iscove's Modified Dulbecco's Medium with 10% FBS; HeLa and MCF7 cells were maintained in Eagle's Minimum Essential Medium with 10% FBS. All the cells were cultured at 37□°C in a humidified atmosphere with 5% CO_2.

Proliferation assay

LNCaP, VCaP, HeLa cells, and MV-4-11 were seeded in 96-well black polystyrene microplates (Corning, Cat #CLS3603) and treated with Compound 1 (C1) and Compound 2 (C2) at a final DMSO concentration of 0.1%. Cell viability was assessed using the CellTiter-Glo® Luminescent Cell Viability Assay. The GI□□ values were calculated based on luminescence measurements, indicating the concentration at which 50% growth inhibition was observed. Western blot analysis of BRD4, BRD3, and BRD2 expression was performed in VCaP, LNCaP, HeLa, and MV-4-11 cell lines following treatment with compounds C1 and C2 across a concentration range (0.1 nM to 1000 nM).

Western Blotting

For Western blotting, cell lysates were prepared in RIPA buffer (ThermoFisher, #89900) supplemented with a protease inhibitor cocktail (Sigma-Aldrich, #P8340), and total protein was quantified using a Pierce BCA Assay (ThermoFisher, #23225). An equal quantity of protein from each sample was separated by 8% SDS-PAGE and transferred to polyvinylidene difluoride (PVDF) membranes. Membranes were probed with primary antibodies specific for BRD2 (Cell Signaling Technology [CST], #5848), BRD3 (Santa Cruz Biotechnology, #sc81202), BRD4 (CST, #13440), and c-Myc (CST, #5605). β-Actin (CST, #3700S) or β-Tubulin (CST, #86298) were used as loading controls. For detection, membranes were incubated with either HRP-conjugated secondary antibodies (CST, #7074 and #7076) followed by enhanced chemiluminescence development, or with near-infrared IRDye® secondary antibodies (LI-COR, #926-68070 and #926-32211) before being imaged on an Odyssey Imaging System (LI-COR Biosciences)

In Vivo Pharmacokinetics (PK) and Pharmacology Studies

Pharmacokinetics studies used male CD-1 mice (7-9 weeks). For IV (3 mg/kg): Compound 1 was formulated by dissolving in 2% v/v DMA + 20% w/v hydroxy beta cyclodextrin in water (pH ~7). For IV dose-escalation (5, 10, 15 mg/kg): Compound 1 dissolved in 10% v/v PEG-400 + 3% w/v Cremophor EL + 87% v/v PBS. Dose administration was done via the tail vein (IV). Blood was collected via the submandibular vein at specified time points (0.083–24 h). Plasma is isolated and stored at -80°C. Compound 1 was quantified using a fit-for-purpose LC-MS/MS method using Carbamazepine as internal standard (calibration range 5–1000 ng/mL). PK parameters calculated using non-compartmental analysis (Phoenix WinNonlin). Pharmacology

studies used an M-V-4-11 cell line-derived xenograft (CDX) model in male athymic nude mice. Cells were injected SC (15 × 10□ cells in HBSS: ECM gel). Treatment was initiated at tumor volumes of ~150–200 mm^3. Mice (n=10/group) received Compound 1 IV (5 mg/kg QD, 10 mg/kg QD, or 10 mg/kg QAD) for 17 days. Tumor volumes (TV = length × $width^2$ × 0.5) were measured thrice weekly, and body weights were measured daily. Tumor growth inhibition (TGI%) calculated: %TGI = 100 – ((ΔT/ΔC) × 100). For mechanism studies, tumor tissues were harvested, snap-frozen, pulverized, lysed (CST lysis buffer), and analyzed via Western blot for BRD2, BRD3, BRD4, c-Myc, and β-actin. A 14-day MTD study used male CD-1 mice (n=6/group), which received Compound 1 IV at 2.5, 5, and 10 mg/kg/day (formulation: 10% PEG-400, 3% Cremophor EL, 87% PBS), monitored for clinical signs, and measured body weight. Hematology, biochemistry, and histopathology were performed post-study.

Statistical analysis

All statistical analyses were performed using the statistical packages provided within the R and Python programming frameworks, along with Past-4 and GraphPad PRISM software. The p-value cutoff for significance was 0.05. Significance levels are * < 0.05, ** < 0.01, *** < 0.001, and **** < 0.0001. Statistical analysis was performed using GraphPad Prism (v10.2.3). Tumor volume comparisons were analyzed using the Brown-Forsythe and Welch ANOVA test to account for unequal variances. Data are presented as mean ± standard error of the mean (SEM), and significance was defined at P□<□0.05. No statistical method was used to predetermine the sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. No statistical methods were used to predetermine sample sizes. Data distribution was assumed to be normal, but this was not formally tested. Data collection and analysis were not performed blind to the conditions of the experiments.

Data Availability

The source code of SynGlue is available on the GitHub project: https://github.com/the-ahuja-lab/SynGlue. The validation datasets of PROTACs(PROTAC-DB 3.0, PROTACpedia), DrugCentral, and Literature are accessible in the Supplementary tables. The study utilized data from DrugBank 24 (v5.1.10), BindingDB (downloaded 28-May-2023), ChEMBL (v33, released 31-May-2023), STITCH (v5.0), BioSNAP - ChG-InterDecagon, ChG-Miner, ChG-TargetDecagon (August 2018 release), and the Small Molecule Suite (Harvard, 2019), and PROTAC-DB 3.0.

Code Availability

A Python package for SynGlue is provided via pip: https://test.pypi.org/project/SynGlue/. SynGlue is free for academic institutions; however, a commercial license key is mandatory for commercial usage. The source code of SynGlue is available on the project GitHub page: https://github.com/the-ahuja-lab/SynGlue.

Conflict of Interest

Gaurav Ahuja serves as a consultant to Aurigene Oncology Limited and is a co-founder of InfraQR Solutions LLP; these affiliations are disclosed as potential competing interests. All other authors declare that they have no competing interests.

Acknowledgments

The authors thank the IT-HelpDesk team of IIIT-Delhi for assisting with the computational resources. We thank all the members of the Ahuja lab for their intellectual contributions at various stages of this project. The Ahuja lab is supported by the Ramalingaswami Re-entry Fellowship (BT/HRD/35/02/2006), a re-entry scheme of the Department of Biotechnology, Ministry of Science & Technology, Government of India, and an intramural Start-up grant from Indraprastha Institute of Information Technology-Delhi. Saveena Solanki, Sanjay Kumar Mohanty, and Vishakha Gautam are supported by the INSPIRE fellowship from the Department of Science & Technology, India.

Author Contributions

G.A. designed the SynGlue Computational Workflow, initiated the computational part of the project, and supervised the overall study. S.S. (Susanta Samajdar) designed, planned, and ensured the smooth execution of all experimental validations. S.S. (Saveena Solanki) spearheaded the SynGlue workflow and designed and coded the entire codebase. S.K.M. assisted in code refinement and implementation. S.S. (Shiva Satija) and S.C. assisted with SynGlue pipeline validation, testing, and MagnetDB data compilation. N.V.M.R.B., S.D. (Sandeep Dukare), N.K.T., N.K.R., and A.A.B. performed all the experiments under the guidance of S.S. (Susanta Samajdar). V.G., S.A., S.K., S.D. (Subhadeep Duari), A.S., and R.S. assisted in SynGlue testing and manuscript editing. D.S. assisted with algorithm implementation

and manuscript editing. C.A. assisted in all wet lab experiment execution, data compilation, and figure making. All authors have read and approved the final manuscript.

Figure Legends

Figure 1: SynGlue's data-driven architecture and MagnetDB performance.

(a) Schematic detailing SynGlue's dual workflow architecture. The data-driven path (top) utilizes the MagnetDB interaction database, TRIE-based Computator for fragment mapping, Annotator for polypharmacology type classification, culminating into the Generator for AI-driven molecule design. The structure-guided path (bottom) starts with a target protein's PDB structure, proceeds through de novo warhead design and performs pharmacological filtering, before combining prioritized warheads with selected E3 ligases into the Generator for *de novo* synthesis of multi-targeting molecules (MTMs) like PROTACs. The Prioritizer module predicts compound potency (e.g., DC□□, D□□□) and applies multi-parameter filtering to select the most promising candidates for experimental validation. **(b)** Conceptual representation of the steps performed to implement the Retrosynthetic Combinatorial Analysis Procedure (RECAP) algorithm to design chemically valid terminal fragments from input molecules, enabling efficient indexing and searching within the SynGlue workflow. **(c)** Schematic representation depicting the exemplary workflow with Cisapride as a query molecule to illustrate the fast search approach for mapping the terminal fragments in the MagnetDB using the TRIE-based approach. **(d)** Bar plot depicting the computational time in seconds required for constructing the TRIE data structure using the terminal fragments of compounds present in the MagnetDB repeated over five runs (mean□±□S.E). **(e)** Bar plots comparing the computational time required to retrieve 1,000 representative fragments using the optimized TRIE structure versus a standard brute-force sequential search. **(f)** Scheme depicting the essential chemical information (e.g., bond order, atom configuration, ring closures, chirality, aromaticity) encoded in the canonical SMILES is preserved during TRIE construction and node representation, ensuring chemically accurate matching. **(g)** Radar plot illustrating the high and consistent retrieval rate (precision) of fragments across multiple independent test runs. **(h)** Bar plots depicting the retrieval precision and performance across fragment complexity bins. **(i)** Scatter plot mapping fragment atom count against search time, demonstrating robust performance of SynGlue data-driven approach across varying fragment sizes. **(j)** A density plot showing the distribution of terminal fragment atom counts within the dataset. This distribution highlights key trends in the structural composition of terminal fragments, which play a critical role in TRIE's search functionality. **(k)**

Schematic outlining the workflow of Dual-stage fragment matching initiated after an exact fragment match is found via the TRIE search. Details the subsequent refinement stage involving similarity scoring (using ECFP4/RDKit fingerprints and Tanimoto/Overlap coefficients) and Maximum Common Substructure (MCS) analysis, along with the application of thresholds to ensure high-confidence fragment-target associations. **(l)** Graphical representation illustrating the SynGlue framework for classifying query compounds into four distinct polypharmacological categories (type 1: Monovalent/Monotarget; type 2: Monovalent/Multitarget; type 3: Multivalent/Monotarget; type 4: Multivalent/Multitarget) based on the number and target associations of their mapped terminal fragments.

Figure 2: Validation of SynGlue's prediction accuracy and classification across the benchmark dataset.

(a) Compilation of validation datasets visualized as a donut plot, showcasing contributions from DrugCentral, PROTAC* (*PROTACpedia and PROTACDB), and chemical-induced dimerization studies. The three main classes: Drug Central, PROTAC, and Others, are highlighted for subsequent comparative analysis. **(b)** Principal Component Analysis (PCA) plot of atom pair descriptors, differentiating the three compound classes. **(c)** Heatmap displaying the distribution of functional groups across the Drug Central, PROTAC, and Others classes. **(d)** Formulae for calculating the hit and miss ratios for compounds within the validation datasets. **(e)** Violin plots comparing the distributions of hit and miss ratios across the Drug Central, PROTAC, and Others classes, illustrating their central tendencies and variability in predictive accuracy. **(f)** A density plot visualizing the overall distribution of hit and miss ratios across all three compound classes. **(g)** Schematic diagram illustrating the SynGlue framework for classifying query compounds into four polypharmacological categories based on the number and target associations of their mapped terminal fragments: type 1 (Monovalent/Monotarget), type 2 (Monovalent/Multitarget), type 3 (Multivalent/Monotarget), and type 4 (Multivalent/Multitarget). **(h)** Bar plot quantifying the number of molecules belonging to each polypharmacological type (type 1-4) within the Drug Central, PROTAC, and Others validation classes. **(i)** Representative molecules from the Drug Central, PROTAC, and Others classes, illustrating examples of types 1, 2, 3, and 4 polypharmacological profiles. **(j)** Venn diagram depicting the overlap of molecular targets among the Drug Central, PROTAC, and Others datasets, revealing shared and unique target spaces. **(k)** Bar plots showing the total number of unique ligands utilized by SynGlue, LigAdvisor, SuperPred, and SwissTarget for predicting multi-targeting molecules. **(l)** Bar plots showing the

total number of unique target proteins utilized by SynGlue, LigAdvisor, SuperPred, and SwissTarget for predicting multi-targeting molecules. **(m)** Alluvial plot illustrating data flow from prediction tools (LigAdvisor, SuperPred, SwissTarget) through different data sources and compound/target types. **(n)** Graphical representation outlining the key steps in the comparative performance evaluation of SynGlue, LigAdvisor, SuperPred, and SwissTarget. **(o)** Comparative histogram plot of hit ratios for SwissTarget, LigAdvisor, SuperPred, and MetaboGlue, along with their corresponding hit rates.

Figure 3: SynGlue enables de novo design of PROTACs and multitarget molecules.

(a) SynGlue's integrative workflows for de novo design of PROTACs and multitarget molecules (MTMs), combining data-driven and structure-guided modules for comprehensive design and linker optimization. **(b)** A comparative analysis showcases Androgen Receptor (AR)-XIAP (X-linked inhibitor of apoptosis protein) PROTACs designed using SynGlue's data-driven (left) and structure-guided (right) approaches, with a Principal Component Analysis (PCA) plot below illustrating the enhanced chemical diversity achieved by SynGlue while preserving predicted target relevance. **(c)** Examples of bifunctional, multi-targeting molecules designed using the SynGlue workflow are provided, including MTMs targeting Epidermal Growth Factor Receptor (EGFR) and Cyclin-Dependent Kinase 1 (CDK1) (linked warheads CHEMBL939 and 4442620) and MTMs designed for Protein Kinase C-related kinase X (PRKX) and Dopamine Receptor D1 (DRD1) (using warheads CHEMBL2152768 and CHEMBL1201356). **(d)** Similarly, a comparative visualization displays Bromodomain-containing protein 4 (BRD4) PROTACs designed using SynGlue's data-driven (left) and structure-guided (right) approaches, accompanied by a PCA plot below indicating the greater chemical diversity of the SynGlue-designed AR PROTACs while maintaining predicted target relevance. **(e)** Multi-stage warhead filtering and prioritization, incorporating BRD analog selection, physicochemical filtering (MW, LogP, TPSA, rotatable bonds, aromatic fraction, SA, Csp^3), novelty assessment, and final warhead selection. **(f)** Distribution of Tanimoto similarity scores for selected novel warheads relative to known BRD family warheads. **(g)** Dendrogram depicting chemical diversity and clustering of selected warheads, illustrating structural novelty and diversity of the SynGlue-optimized warhead library. **(h)** A flow diagram further details the molecular docking strategy employed for validating or prioritizing the designed warheads, utilizing the experimentally determined structure of Bromodomain-containing protein 4 (BRD4) (Protein Data Bank identifier: 3MXF). **(i)** Representative chemical structures of prioritized warheads, highlighting their

structural diversity and novelty. A tabular representation provides the chemical structures of SynGlue-designed warheads against the BRD4 protein. **(j)** Interaction map of an optimized warhead within BRD4's binding cavity, showing key protein-ligand interactions. **(k)** Predictive modeling workflow incorporating component-wise and linker-specific features, with models trained for DC_{50} and D_{max} regression as well as linker-specific multi-class classification using scikit-learn (v1.3.0) and cross-validation. **(l)** Performance metrics for the predictive models (R^2, MAE, MSE, RMSE) and density plots for predicted D_{max} and log-transformed DC_{50} distributions, confirming model robustness. **(m)** The chemical structures of two BRD4-targeting PROTACs designed using the SynGlue workflow are shown, accompanied by a table detailing comparative physicochemical profiles and predicted biological activity metrics of these two selected BRD4-specific SynGlue-designed PROTACs.

Figure 4: Preclinical Validation of SynGlue-Designed BRD4 PROTACs.

(a) A bar plot displays the Growth Inhibition 50 (GI_{50}) values in nanomolar (nM) for SynGlue-designed PROTACs, Compound 1 and Compound 2, as determined by CellTiter-Glo® assay across various cancer cell lines, including MV-4-11, VCaP, LNCaP, and MCF7. **(b)** A schematic representation outlines the experimental strategy employed to assess the in vitro Bromodomain-containing protein 4 (BRD4) protein degradation potential of the SynGlue-designed PROTACs in VCaP, LNCaP, and HeLa cell lines following compound treatment for a specified duration (e.g., 6 hours). **(c)** Western blots illustrating the dose-dependent degradation of BRD4, BRD2, and BRD3 proteins in VCaP, LNCaP, and HeLa cell lines treated with increasing concentrations (0.1 nM to 100 nM) of Compound 1 and Compound 2 over 6 hours, with β-Actin or β-Tubulin serving as loading controls; the selective BRD4 degradation by Compound 1 is particularly evident in HeLa cells. **(d)** A Western blot demonstrating the proteasomal degradation of Compound 1-mediated BRD4 degradation, showing that pre-treatment of HeLa cells with a proteasome inhibitor (e.g., MG132 or Bortezomib) followed by Compound 1 treatment rescues BRD4 protein levels compared to treatment with Compound 1 alone. **(e)** A Western blot illustrates the requirement for a functional PROTAC structure for degradation, as treatment of cells (e.g., HeLa) with an inactive isomer of Compound 1 (AU-16914) at equivalent concentrations fails to induce BRD4 degradation compared to the active Compound 1. **(f)** Schematic of the in vivo pharmacokinetic (PK) study design, outlining single intravenous (IV) administration of Compound 1 in male CD-1 mice at 5, 10, and 15 mg/kg, with the corresponding plasma concentration time profiles (ng/mL versus hours) shown in line plots to determine PK parameters (C_{max}, AUC, half-life). **(g)** Line plots showing the maximum tolerated

doses (MTD) and PK profiles of Compound 1, with consistent body weight maintenance across 2.5, 5, and 10 mg/kg/day IV doses over the 14-day administration period, indicating good tolerability. **(h)** Tolerability assessment of Compound 1 in vivo, with line plots tracking average body weight (% change) of male CD-1 mice during the administration and treatment periods, confirming the absence of significant systemic toxicity at tested doses. **(i)** A schematic representation outlines the in vivo anti-tumor efficacy study design using a relevant xenograft model (e.g., MV-4-11 Cell Line Derived Xenograft (CDX) in athymic nude mice), including treatment groups (vehicle control, Compound 1 at 10 mg/kg daily or alternate day IV, treatment duration (e.g., 17 days), and primary endpoints (tumor volume, body weight). **(j)** Line plots tracking tumor volume (mm^3, mean ± SEM) in treated mice, demonstrating robust tumor growth inhibition by Compound 1 relative to vehicle control, with dose-dependent efficacy. **(k)** A scatterplot summarizes the effect of Compound 1 treatment on tumor volume (cubic millimeters (mm^3), Mean ± Standard Error of the Mean (SEM)) at the study endpoint in the xenograft model, comparing vehicle control, Compound 1 treatment groups, with statistical significance indicated. **(l)** Line plots of average body weight changes during the in vivo efficacy study, confirming tolerability across all treatment groups. Experimental setup for evaluating the efficacy of compound 1 in mice: Male athymic nude mice (7 to 8 weeks old, 10 mice per group) were inoculated with MV4-11 cells (ATCC) at a density of 15 million cells per mouse in 200 μL (1:1 HBSS & ECM gel). Treatment was initiated 8 days after inoculation, when the mean tumor volume reached approximately 170 mm^3, and continued for 17 days. Statistical analysis was performed using GraphPad Prism (version 10.2.3), applying Brown-Forsythe and Welch ANOVA tests. **(m)** Table summarizing the measured concentrations of Compound 1 in plasma and tumor tissues at selected post-treatment time points in the efficacy study, demonstrating sustained intratumoral exposure and a favorable pharmacokinetic–pharmacodynamic (PK-PD) correlation. Compound 1 exposure was assessed in plasma samples ($n = 9$-10/group) and tumor tissues ($n = 8$-10/group) at the 1-hour time point following the end of the efficacy study. **(n)** Representative Western blots show the protein levels of BRD4, BRD2, BRD3, and the downstream target c-Myc in tumor tissue lysates harvested from mice at baseline (pre-treatment) and at early time points (1 and 4 hours) post-treatment with Compound 1, demonstrating pharmacodynamic target engagement and pathway modulation. Representative Western blots show the protein levels of BRD4, BRD2, BRD3, and c-Myc in tumor tissue lysates harvested from mice at later time points (6, 8 and 24 hours) post-treatment with Compound 1, illustrating the duration of pharmacodynamic effects.

Supplementary Figure Legends

Supplementary Figure 1: MagnetDB Fragmentation, TRIE Characteristics, and Workflow Schematic

(a) A comparison table of five molecular databases presents the total number of unique compounds, the total number of fragments designed, and the subset of fragments identified as terminal fragments, highlighting the structural depth available for mapping. **(b)** The workflow of the data-driven module is shown, outlining the steps from Simplified Molecular Input Line Entry System (SMILES) input to final fragment type classification, including a schematic diagram detailing the sequential processes of RECAP (Retrosynthetic Combinatorial Analysis Procedure) fragmentation, Trie indexing, filtering, scoring (using Tanimoto coefficient and Maximum Common Substructure (MCS)), and type assignment. **(c)** The distribution of terminal fragments across various databases is depicted, with an inner pie chart illustrating the contribution of selected pharmacophore-containing fragments from each source. **(d)** A Trie structure summary table presents key characteristics of the constructed Trie structure, including total branches, nodes, build time, average frequency, and search efficiency (approximately 99.98%), validating the scalability of the infrastructure. **(e)** The Trie operation complexity is described, detailing the time and space complexity of Trie-based operations (insert, search, delete) used for fragment handling, emphasizing its efficiency in large search space searches and matches. **(f)** Scatter plot showing the relationship between fragment length and search time for 1000 test fragments, validating the computational performance of Trie-based search algorithms. **(g)** Line plot of average search time (in milliseconds) across 1000 test fragments, demonstrating consistent efficiency and highlighting any outliers. **(h)** Performance comparison plots, presented as box plots, compare the average fragment search time (in milliseconds) across databases (left) and the average prefix match time (right), indicating variability in how quickly fragment prefixes are matched in different sources (STITCH, BindingDB, DrugBank, BioSNAP, ChEMBL, SMS), illustrating variation in search performance. **(i)** The fragmentation hierarchy of Claspedine is illustrated as an example of a parent molecule undergoing RECAP fragmentation into first- and second-level terminal fragments, with selected fragments (for analysis) and filtered non-selected fragments indicated.

Supplementary Figure 2: Benchmarking SynGlue: Dataset Characterization and Performance Validation

(a) A donut plot shows the distribution of targets across the dataset, highlighting the proportion of validated targets categorized from PROTACs (Proteolysis-Targeting Chimeras) and other sources (chemical-induced dimerization, scientific literature, and drugs). **(b)** A summary table lists the counts of ligands and terminal fragments across curated datasets such as DrugCentral, PROTACpedia, and literature-derived sources. **(c)** A workflow schematic illustrates SynGlue's stepwise validation process, including target mapping, annotation, and calculating hit and miss ratios. **(d)** A Venn diagram depicts the fragment overlap between MagnetDB and external datasets for DrugCentral, PROTACs, and Others. **(e)** Density plots of hit and miss ratios for selected example ligands from DrugCentral, PROTACs, and Others are shown, illustrating performance separation. **(f)** A t-distributed Stochastic Neighbor Embedding (t-SNE) visualization displays 10 molecular clusters based on the structural similarity of validated molecules. **(g)** A t-SNE plot colored by chemical types (1-4) illustrates the type-wise heterogeneity in chemical space. **(h)** A distribution plot of LogP values (logarithm of the partition coefficient) reflects the lipophilic diversity of the compound dataset.

Supplementary Figure 3: SynGlue PROTAC Generation: Workflow, Parameters, and Physicochemical Property Distributions

(a) The workflow of the SynGlue PROTAC (Proteolysis-Targeting Chimera) design pipeline is illustrated, encompassing steps from warhead and E3 ligase selection to linker optimization, scoring, and prediction. **(b)** A schematic diagram shows the noCof20 score filtering and prioritization logic employed for selecting suitable PROTAC candidates. **(c)** An atom count distribution plot of selected warhead fragments reveals clustering around typical molecular sizes. **(d)** The atom count distribution of E3 ligase ligands displays bimodal trends in fragment length. **(e)** An atom count density distribution is shown for whole PROTAC molecules after assembly. **(f)** Line plots display the statistical distribution of E3 ligase ligand properties, including Hydrogen Bond Donors (HBD), Hydrogen Bond Acceptors (HBA), rotatable bonds, and atomic distances. **(g)** A radar plot compares linker graph-based lengths, ring count, and effective length across the dataset. **(h)** A table lists the parameters for the Recurrent Neural Network (RNN) employed in linker design. **(i)** Density plots visualize various linker features, including HBD, length ratio, number of rings, and effective length. **(j)** A box plot summarizes the linker property distribution across different PROTAC component types. **(k)** A dot matrix shows the frequency of functional group classes across PROTAC components (warheads, linkers, and

E3 ligands), highlighting enriched fragment features in each.

Supplementary Figure 4: Development and Evaluation of Linker-Based PROTAC Classification Models

(a) An overview of the machine learning pipeline for multiclass classification is presented, utilizing linker-based features extracted from PROTAC (Proteolysis-Targeting Chimera) structures and integrating Grover and image-based embeddings; the pipeline encompasses feature selection, hyperparameter tuning, 10-fold cross-validation, and model evaluation across 10 different classifiers. **(b-c)** Performance metrics (Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic 1 Curve (AUC)) are shown for training and testing datasets using Signaturizer features. **(d-e)** Training and testing performance are displayed using ChemBerta embeddings. **(f-g)** Model evaluation results are presented using image-based PROTAC features, showing both training and testing metrics. **(h-i)** Training and testing results are shown using Grover-based graph embeddings. **(j-k)** Model evaluation is provided using Mordred descriptors for training and testing datasets. **(l-m)** Box plots represent the performance distribution over 10-fold cross-validation using gradient boosting for both training and testing datasets. **(n)** Summary statistics of test set predictions (precision, recall, F1-score, MCC, AUC) across all classifiers for the gradient boosting model are presented, highlighting consistency across the metrics.

Supplementary Figure 5: Structure-Guided Cavity Analysis and Attention-Based Regression Model for DC□□ and D_{max} Prediction

(a-b) Heatmaps illustrate the predicted androgen receptor interactions for PROTACs degradation (Proteolysis-Targeting Chimeras) designed using data-driven and structure-guided approaches. **(c)** A table presents the cavity properties (volume, druggability score, and DScores) of Bromodomain-containing protein 4 (BRD4), identifying binding pockets categorized as likely druggable or undruggable. **(d)** A schematic diagram outlines the transformer-based regression model pipeline used for predicting the concentration at half-maximal degradation (DC□□) and maximal degradation (D_{max}), leveraging attention-weighted embeddings derived from the E3 ligase, warhead, and linker components of the PROTACs. **(e)** A box plot displays the performance metrics (Accuracy, Precision, F1-score, Cohen's Kappa coefficient, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve

(AUC) of the regression model across multiple evaluation folds, demonstrating the model's stability. **(f)** Box plots displaying the performance metrics (R^2, MAE, MSE, RMSE) for the regression-based attention learning model predicting maximal degradation (D□□□) of designed PROTACs. **(g)** Box plots showing the performance of the same model in predicting the half-maximal degradation concentration (DC□□), demonstrating robust predictive capability across all metrics. **(h)** Receiver Operating Characteristic (ROC) curves for the multi-class classification model developed to predict DC□□ classes for linker-specific designs, illustrating high area under the curve (AUC-ROC) values across classes (AUC 0.80-0.88), confirming the model's classification accuracy. **(i)** Comparative structural illustrations of designed PROTACs (bottom row) and experimental PROTACs (top row) targeting Von Hippel-Lindau (VHL) E3 ligase, demonstrating both data-driven (right) and structure-guided (left) SynGlue workflows and the conformational diversity captured by the generative design approach. **(j)** Filtering of chemical library based on physicochemical and structural properties. The bar plot shows the number of molecules remaining after each filtering step, starting with molecular weight (40–1066), followed by clogP (0–5), TPSA (0–250), halogen count (≤ 1), heteroatoms (≥ 21), and LogS ESOL (≥ −7). The dataset initially contained 3867 molecules, progressively reduced to 20 molecules that satisfy all filter criteria. **(k)** Summary of key *in vitro* and *in vivo* properties of SynGlue-designed Compound 1 and Compound 2. Physicochemical properties include molecular weight (MW/clogP), BRD4 BD1 IC_{50}, MV-4-11 GI_{50}, VCaP GI_{50}, c-MYC IC_{50}, LNCaP GI_{50}, c-MYC IC_{50}, MCF7 GI_{50}, and aqueous solubility at pH 7.4 (Eq. sol). In vitro stability is summarized by MLM% remaining at 30 and 60 min, plasma protein binding (PPB, % bound), and Caco-2 permeability/efflux data. Pharmacokinetic (PK) parameters for single IV dosing in mice include half-life (t□/□), maximum concentration (C_{max}), area under the curve (AUC), and clearance (Cl), demonstrating differences in drug-like profiles and plasma exposure of the two compounds.

Supplementary Figure 6: Molecular Interaction Analysis of Top Warheads with BRD4 Binding Site

(a-b) Ligplots and interaction maps depicting the interactions between Warhead 1 and the BRD4 active site, illustrating strong anchoring via multiple hydrogen bonds and hydrophobic contacts. **(c-d)** Ligplots and interaction maps depicting the interactions between Warhead 2 and

the BRD4 active site, showing favorable planar alignment with BRD4 residues through π−π stacking and key polar interactions. **(e-f)** Ligplots and interaction maps depicting the interactions between Warhead 3 and the BRD4 active site, demonstrating robust interaction via halogen and hydrogen bonding with critical residues (Asn, Tyr). **(g)** Ligplots and interaction maps depicting the interactions between Warhead 4 and the BRD4 active site, displaying extended interactions with both polar and nonpolar residues, suggesting dual-site anchoring. **(h-j)** Ligplots and interaction maps depicting the interactions between Warhead 5 and the BRD4 active site, maintaining consistent hydrogen bonding networks with high surface contact (corroborated by multiple analyses). **(k)** Ligplots and interaction maps depicting the interactions between Warhead 6 and the BRD4 active site, engaging deeply within the pocket through aromatic stacking and polar contacts, contributing to predicted high affinity. The overall panel illustrates the binding of six selected warheads (Warhead 1 to Warhead 6) to the Bromodomain-containing protein 4 (BRD4) active site.

Supplementary Figure 7: Synthesis, Characterization, and Functional Evaluation of SynGlue-designed PROTACs for BRD4.

(a) High-performance liquid chromatography (HPLC) chromatogram of Compound 1, showing a primary peak at 6.450 min retention time (RT) with an area percentage of 98.37%, confirming its purity. **(b)** High-resolution mass spectrometry (HRMS) data for Compound 1, including the extracted ion chromatogram and the observed molecular ion peaks, confirming its exact mass (1020.4415 Da) and molecular formula (C□□H□□N□O□□S). **(c)** High-resolution mass spectrometry (HRMS) data for Compound 1, including the extracted ion chromatogram and the observed molecular ion peaks, confirming its exact mass (1020.4415 Da) and molecular formula (C□□H□□N□O□□S). **(d)** HPLC chromatogram of Compound 2, with a primary peak at 7.455 min RT and an area percentage of 98.26%, indicative of high compound purity. **(e)** HRMS data for Compound 2, displaying extracted ion chromatograms and molecular ion peaks consistent with its exact mass (1016.4830 Da) and molecular formula (C□□H□□N□O□□S). **(f)** ^{1}H NMR (top) and ^{13}C NMR (bottom) spectra of Compound 2 in DMSO-d□, demonstrating expected chemical shifts and confirming the compound's structural features. **(g)** High-performance liquid chromatography (HPLC) chromatogram of Compound 1 inactive isomer, showing a prominent peak at 6.357 minutes retention time (RT) with an area percentage of 97.69%, confirming its purity. **(h)** High-resolution mass spectrometry (HRMS) spectrum of the inactive isomer of

Compound 1, displaying the observed molecular ion peaks and confirming the exact mass (1020.4415 Da) and the molecular formula (C□□H□□N□O□□S) of the compound. **(i)** Nuclear magnetic resonance (NMR) spectra of Compound 2 in DMSO-d□: the upper panel shows the ^{1}H NMR spectrum, with characteristic proton signals across the chemical shift range, while the lower panel presents the ^{13}C NMR spectrum, depicting distinct carbon chemical shifts, together confirming the structural integrity and chemical identity of Compound 1 inactive isomer.

Supplementary Figure 8: In Vivo Pharmacokinetics, Tolerability, and Pharmacodynamics of SynGlue-designed BRD4 PROTAC.

(a, b) Dose-response curves for cell proliferation inhibition by Compound 1 and its inactive isomer in MV-4-11 cell lines, showing differential potency across varying concentrations (nM range). **(c, d)** Western blot analysis of BRD4 degradation in MV-4-11 cells treated with increasing concentrations (0.001–1000 nM) of Compound 1 and Compound 2 for 16 hours, with tubulin as a loading control. The lower panels show dose–response curves for BRD4 degradation in MV-4-11 cells, with DC□□ values calculated as 0.193 μM for Compound 1 and 0.228 μM for Compound 2.

Supplementary Data

Supplementary Table 1: Tabular representation detailing information about the statistical summary of the curated MagnetDB database, including counts of unique protein-ligand interactions, compounds, protein targets, and terminal fragments, broken down by species and originating source database.

Supplementary Table 2: Tabular representation detailing information about the core modules comprising the SynGlue platform and the key external software packages, libraries, and databases utilized or integrated within the workflow, along with their version numbers where applicable.

Supplementary Table 3: Tabular representation of information about the composition of the benchmarking dataset: CompoundID, their SMILES, and class.

Supplementary Table 4: Tabular representation detailing information about the calculated hit ratio and miss ratio values for SynGlue and the benchmark tools (LigAdvisor, SwissTargetPrediction, SuperPred) based on target predictions for the selected subset of 100 representative molecules, allowing direct comparison.

Supplementary Table 5: Tabular representation detailing information about key characteristics of Androgen Receptor (AR)-XIAP PROTACs designed *de novo* using both the data-driven and structure-guided workflows of SynGlue, potentially including compound identifiers, generating workflow, predicted properties, scores, and comparison data for experimentally validated reference AR PROTACs.

Supplementary Table 6: Tabular representation detailing information about key characteristics of Multi-Target Molecules (MTMs) designed *de novo* using SynGlue to target specific pairs (EGFR/CDK1 and PRKX/DRD1), potentially including compound identifiers, targeted protein pair, generating workflow, and predicted properties.

Supplementary Table 7: Tabular representation detailing information about key characteristics of BRD4-VHL PROTACs designed *de novo* using both the data-driven and structure-guided workflows of SynGlue, potentially including compound identifiers, generating workflow, predicted properties, scores from predictive models, and comparison data for experimentally validated reference BRD4 PROTACs.

Supplementary Table 8: Tabular representation detailing information about the *in silico* characterization of the final prioritized BRD4-VHL PROTAC candidates selected for experimental validation, including compound identifiers, key computed physicochemical properties, predicted degradation potency (DC_{50}, D_{max} values), and predicted potency class.

Supplementary Table 9: Presents a comprehensive pharmacokinetic and pharmacodynamic profile of Compound 1, including dose-dependent plasma and tumor concentrations, key PK parameters (t□/□, C□, AUC, CL, Vdss, MRT) across 2.5–10 mg/kg doses, tumor drug retention over 24 hours, formulation details, single-dose exposure at multiple time points, and target engagement data showing BRD2/3/4 and c-Myc degradation in tumor lysates**.**

References

1. Brennan, R. J. *et al.* The state of the art in secondary pharmacology and its impact on the safety of new medicines. *Nat. Rev. Drug Discov.* **23**, 525–545 (2024).
2. Brodie, J. S., Di Marzo, V. & Guy, G. W. Polypharmacology shakes hands with complex aetiopathology. *Trends Pharmacol. Sci.* **36**, 802–821 (2015).
3. Xiong, Z. *et al.* Crowdsourced identification of multi-target kinase inhibitors for RET- and TAU- based disease: The Multi-Targeting Drug DREAM Challenge. *PLoS Comput. Biol.* **17**, e1009302 (2021).
4. Liu, Z. *et al.* An overview of PROTACs: a promising drug discovery paradigm. *Mol. Biomed.* **3**, 46 (2022).
5. Chen, Z. *et al.* Flexible scaffold-based cheminformatics approach for polypharmacological drug design. *Cell* **187**, 2194–2208.e22 (2024).
6. Dar, A. C., Das, T. K., Shokat, K. M. & Cagan, R. L. Chemical genetic discovery of targets and anti-targets for cancer polypharmacology. *Nature* **486**, 80–84 (2012).
7. Ciceri, P. *et al.* Dual kinase-bromodomain inhibitors for rationally designed polypharmacology. *Nat. Chem. Biol.* **10**, 305–312 (2014).
8. Rudrapal, M., Paudel, K. R. & Pangeni, R. Editorial: Drug repurposing and polypharmacology: A synergistic approach in multi-target based drug discovery. *Front. Pharmacol.* **13**, 1101007 (2022).
9. van Veldhuisen, T. W. *et al.* Modulation of protein-protein interactions with molecular glues in a synthetic condensate platform. *J. Am. Chem. Soc.* **147**, 5386–5397 (2025).
10. Hinterndorfer, M., Spiteri, V. A., Ciulli, A. & Winter, G. E. Targeted protein degradation for cancer therapy. *Nat. Rev. Cancer* (2025) doi:10.1038/s41568-025-00817-8.
11. Nalawansha, D. A. & Crews, C. M. PROTACs: An emerging therapeutic modality in precision medicine. *Cell Chem. Biol.* **27**, 998–1014 (2020).

12. Békés, M., Langley, D. R. & Crews, C. M. PROTAC targeted protein degraders: the past is prologue. *Nat. Rev. Drug Discov.* **21**, 181–200 (2022).
13. Samarasinghe, K. T. G. & Crews, C. M. Targeted protein degradation: A promise for undruggable proteins. *Cell Chem. Biol.* **28**, 934–951 (2021).
14. Gao, H., Sun, X. & Rao, Y. PROTAC technology: Opportunities and challenges. *ACS Med. Chem. Lett.* **11**, 237–240 (2020).
15. Bondeson, D. P. *et al.* Lessons in PROTAC design from selective degradation with a promiscuous warhead. *Cell Chem. Biol.* **25**, 78–87.e5 (2018).
16. Malarvannan, M., Unnikrishnan, S., Monohar, S., Ravichandiran, V. & Paul, D. Design and optimization strategies of PROTACs and its Application, Comparisons to other targeted protein degradation for multiple oncology therapies. *Bioorg. Chem.* **154**, 107984 (2025).
17. Scott, J. S., Michaelides, I. N. & Schade, M. Property-based optimisation of PROTACs. *RSC Med. Chem.* (2024) doi:10.1039/d4md00769g.
18. A S, B. G., Agrawal, D., Kulkarni, N. M., Vetrivel, R. & Gurram, K. PROTAC-design-Evaluator (PRODE): An advanced method for in-silico PROTAC design. *ACS Omega* **9**, 12611–12621 (2024).
19. Ugurlu, S. Y., McDonald, D., Enisoglu, R., Zhu, Z. & He, S. MEGA PROTAC, MEGA DOCK-based PROTAC mediated ternary complex formation pipeline with sequential filtering and rank aggregation. *Sci. Rep.* **15**, 5545 (2025).
20. Zaidman, D., Prilusky, J. & London, N. PRosettaC: Rosetta based modeling of PROTAC mediated ternary complexes. *J. Chem. Inf. Model.* **60**, 4894–4903 (2020).
21. Li, F., Hu, Q., Zhou, Y., Yang, H. & Bai, F. DiffPROTACs is a deep learning-based generator for proteolysis targeting chimeras. *Brief. Bioinform.* **25**, bbae358 (2024).
22. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–7 (2012).
23. *BindingDB: Binding Database*. (Center for Advanced Research in Biotechnology, University

of Maryland, 2007).

24. Knox, C. *et al.* DrugBank 6.0: The DrugBank knowledgebase for 2024. *Nucleic Acids Res.* **52**, D1265–D1275 (2024).
25. Szklarczyk, D. *et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **44**, D380–4 (2016).
26. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).
27. Moret, N. *et al.* Cheminformatics tools for analyzing and designing optimized small-molecule collections and libraries. *Cell Chem. Biol.* **26**, 765–777.e3 (2019).
28. Mohanty, S. K. *et al.* Deep learning reveals endogenous sterols as allosteric modulators of the GPCR-Gα interface. *bioRxiv* 2023.02. 14.528587 (2023) doi:10.1101/2023.02.14.528587.
29. Green, H. & Durrant, J. D. DeepFrag: An open-source browser app for deep-learning lead optimization. *J. Chem. Inf. Model.* **61**, 2523–2529 (2021).
30. Bai, Q. *et al.* MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Brief. Bioinform.* **22**, (2021).
31. Lewell, X. Q., Judd, D. B., Watson, S. P. & Hann, M. M. RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **38**, 511–522 (1998).
32. Axen, S. D. *et al.* A simple representation of three-dimensional molecular structure. *J. Med. Chem.* **60**, 7393–7409 (2017).
33. Dalke, A. & Hastings, J. FMCS: a novel algorithm for the multiple MCS problem. *Journal of Cheminformatics* **5**, O6 (2013).
34. Ge, J. *et al.* PROTAC-DB 3.0: an updated database of PROTACs with extended pharmacokinetic parameters. *Nucleic Acids Res.* **53**, D1510–D1515 (2025).

35. Pinzi, L., Tinivella, A., Gagliardelli, L., Beneventano, D. & Rastelli, G. LigAdvisor: a versatile and user-friendly web-platform for drug design. *Nucleic Acids Res.* **49**, W326–W335 (2021).
36. Daina, A., Michielin, O. & Zoete, V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Res.* **47**, W357–W364 (2019).
37. Gallo, K., Goede, A., Preissner, R. & Gohlke, B.-O. SuperPred 3.0: drug classification and target prediction-a machine learning approach. *Nucleic Acids Res.* **50**, W726–W731 (2022).
38. Bertoni, M. *et al.* Bioactivity descriptors for uncharacterized chemical compounds. *Nat. Commun.* **12**, 3932 (2021).
39. Rong, Y. *et al.* Self-supervised graph Transformer on large-scale molecular data. *Adv. Neural Inf. Process. Syst.* (2020).
40. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).
41. Zeng, X. *et al.* Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence* **4**, 1004–1016 (2022).
42. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv [cs.LG]* (2020).
43. Karpińska, K. *et al.* Selective degradation of MLK3 by a novel CEP1347-VHL-02 PROTAC compound limits the oncogenic potential of TNBC. *J. Med. Chem.* **67**, 15012–15028 (2024).

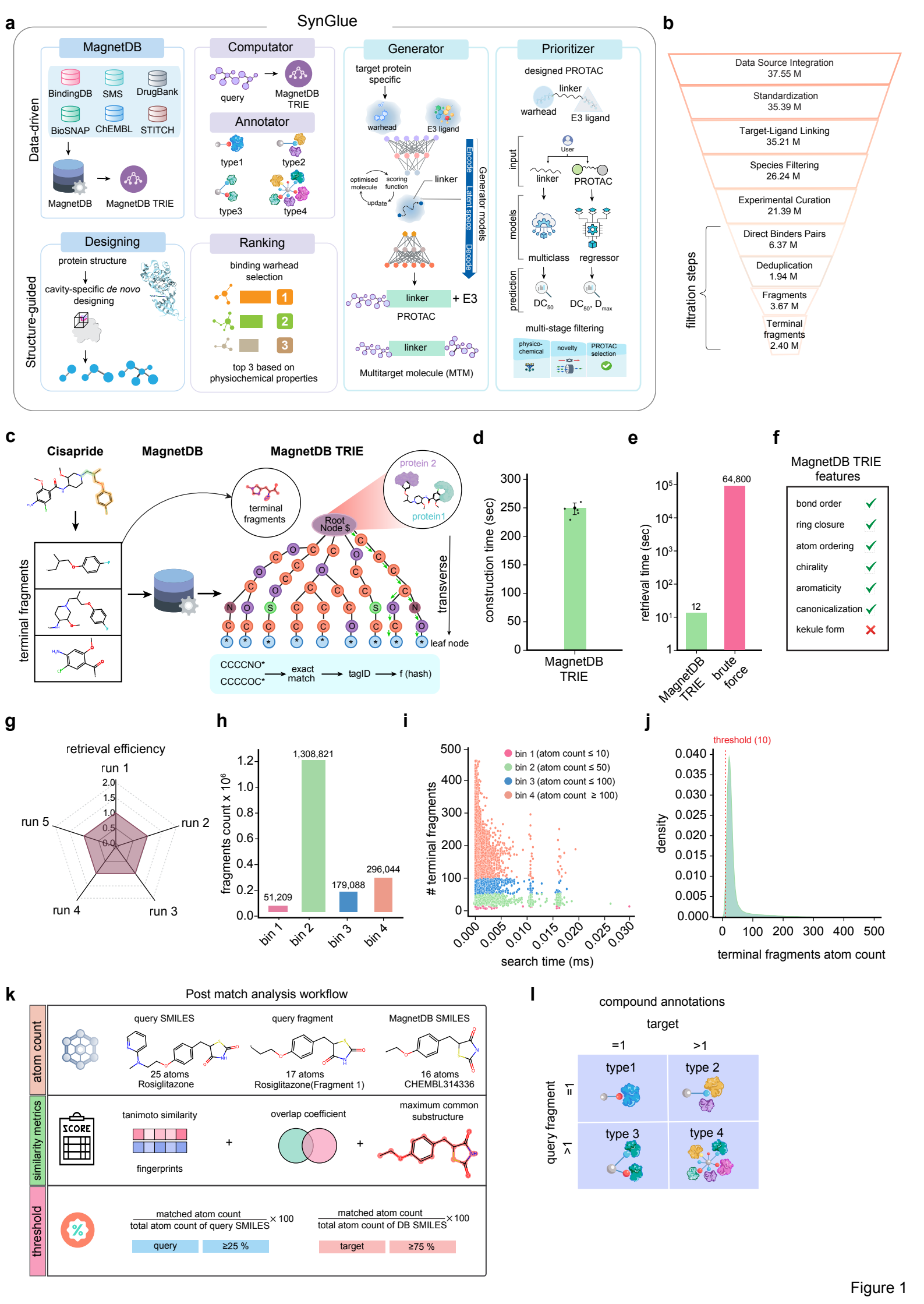

a
SynGlue
MagnetDB
Data-driven
BindingDB
SMS
DrugBank
BioSNAP
ChEMBL
STITCH
MagnetDB
MagnetDB TRIE
Computator
query
MagnetDB TRIE
Annotator
type1
type2
type3
type4
Generator
target protein specific
warhead
E3 ligand
optimised molecule
scoring function
update
linker
Encode
Latent space
Decode
Generator models
linker
+ E3
PROTAC
linker
Multitarget molecule (MTM)
Prioritizer
designed PROTAC
linker
warhead
E3 ligand
input
User
linker
PROTAC
models
multiclass
regressor
prediction
DC50
DC50, Dmax
multi-stage filtering
physico-chemical
novelty
PROTAC selection
Structure-guided
Designing
protein structure
cavity-specific de novo designing
Ranking
binding warhead selection
1
2
3
top 3 based on physiochemical properties
b
Data Source Integration
37.55 M
Standardization
35.39 M
Target-Ligand Linking
35.21 M
Species Filtering
26.24 M
Experimental Curation
21.39 M
Direct Binders Pairs
6.37 M
Deduplication
1.94 M
Fragments
3.67 M
Terminal fragments
2.40 M
filtration steps
c
Cisapride
MagnetDB
MagnetDB TRIE
terminal fragments
terminal fragments
protein 2
protein1
Root Node $
transverse
leaf node
CCCCNO*
CCCCOC*
exact match
tagID
f (hash)
d
construction time (sec)
300
250
200
150
100
50
0
MagnetDB TRIE
e
retrieval time (sec)
10^5
10^4
10^3
10^2
10^1
1
64,800
12
MagnetDB TRIE
brute force
f
MagnetDB TRIE features
bond order
ring closure
atom ordering
chirality
aromaticity
canonicalization
kekule form
g
retrieval efficiency
run 1
run 2
run 3
run 4
run 5
2.0
1.5
1.0
0.5
0.0
h
fragments count x 10^6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
51,209
1,308,821
179,088
296,044
bin 1
bin 2
bin 3
bin 4
i
terminal fragments
500
400
300
200
100
0
bin 1 (atom count ≤ 10)
bin 2 (atom count ≤ 50)
bin 3 (atom count ≤ 100)
bin 4 (atom count ≥ 100)
0.000
0.005
0.010
0.015
0.020
0.025
0.030
search time (ms)
j
threshold (10)
density
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
500
terminal fragments atom count
k
Post match analysis workflow
atom count
query SMILES
query fragment
MagnetDB SMILES
25 atoms
Rosiglitazone
17 atoms
Rosiglitazone(Fragment 1)
16 atoms
CHEMBL314336
similarity metrics
SCORE
tanimoto similarity
fingerprints
+
overlap coefficient
+
maximum common substructure
threshold
matched atom count / total atom count of query SMILES × 100
matched atom count / total atom count of DB SMILES × 100
query
≥25 %
target
≥75 %
l
compound annotations
target
=1
>1
query fragment
=1
>1
type1
type 2
type 3
type 4

Figure 1

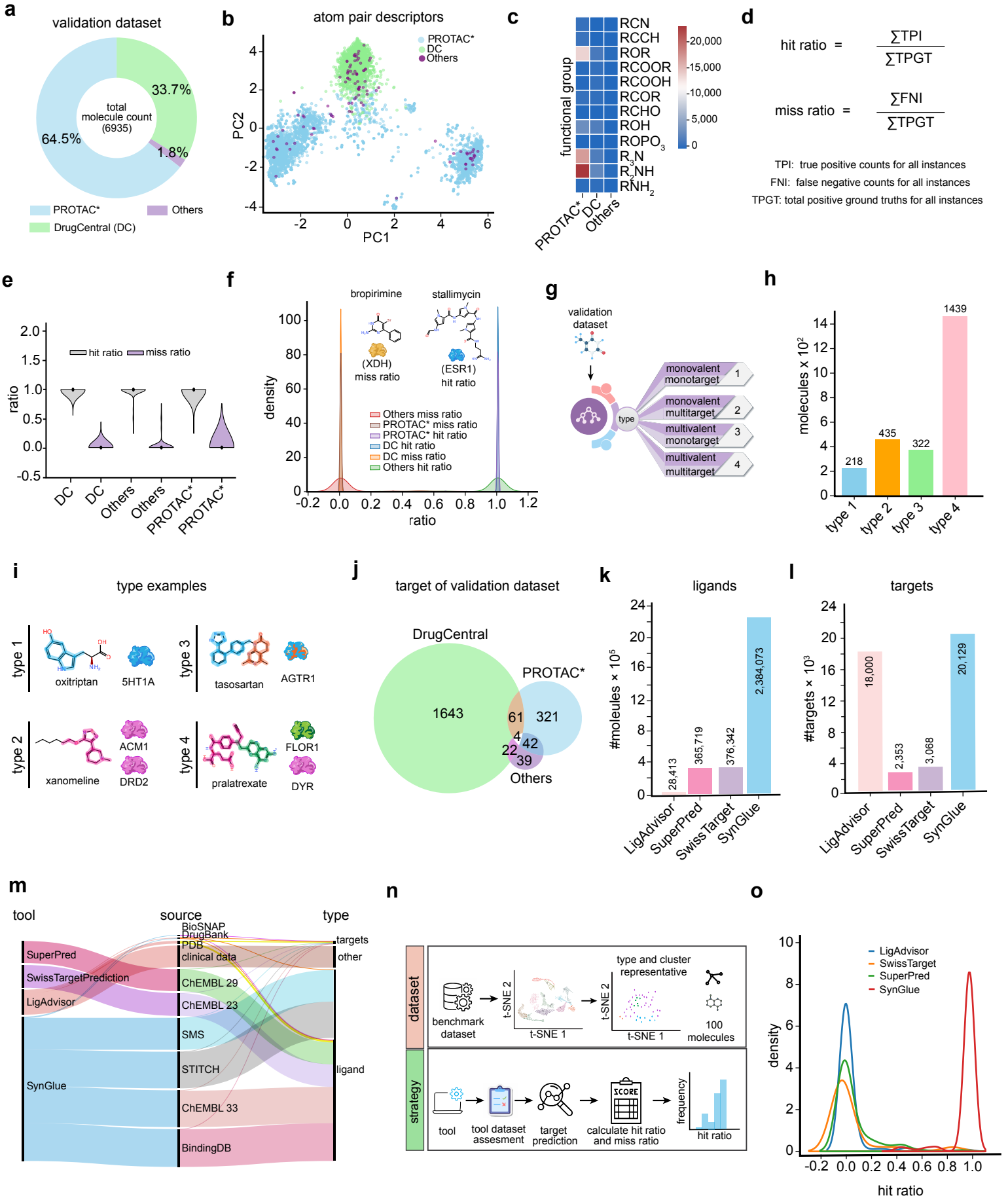
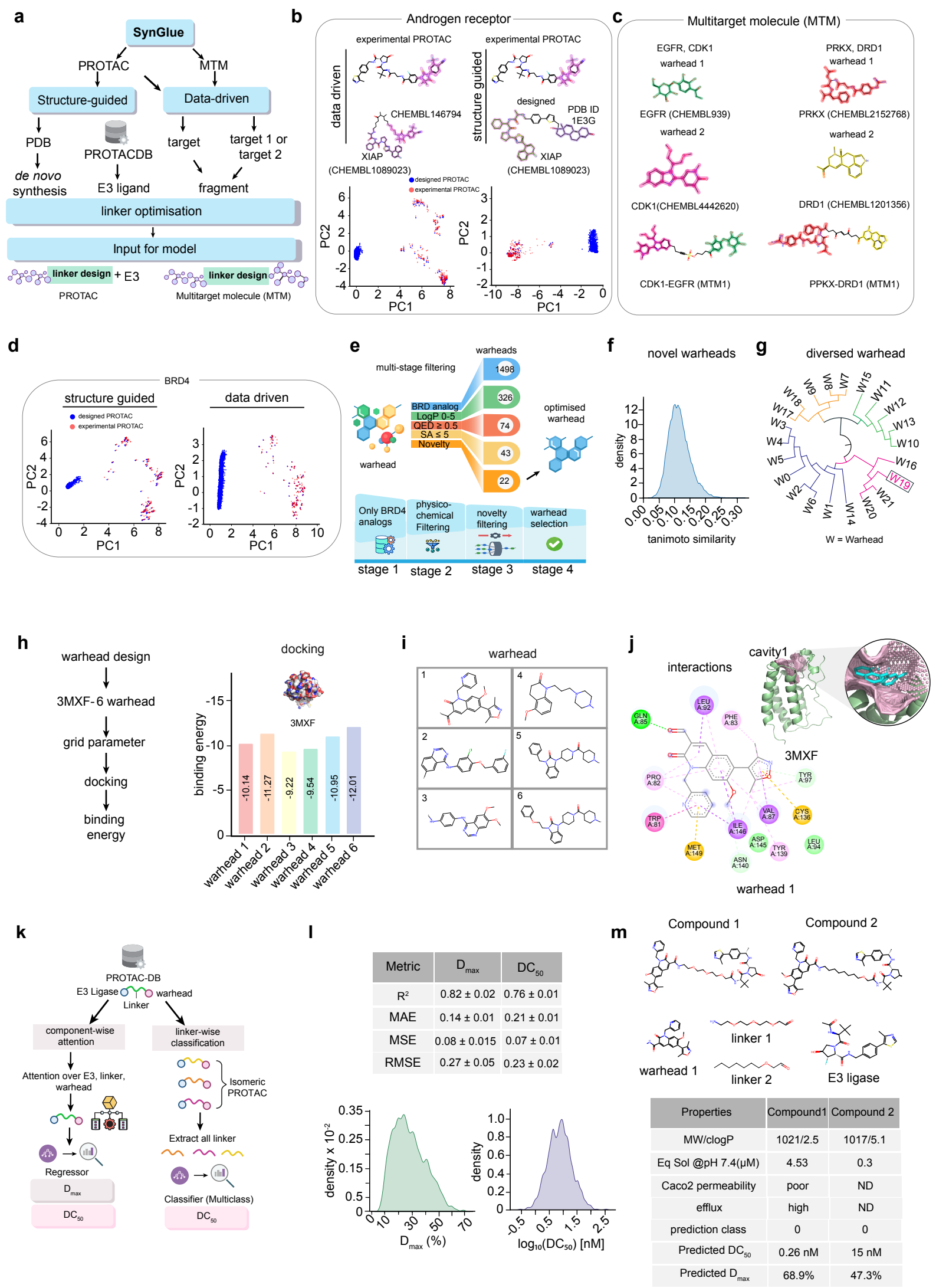

a
validation dataset
total molecule count (6935)
64.5%
33.7%
1.8%
PROTAC*
Others
DrugCentral (DC)
b
atom pair descriptors
PROTAC*
DC
Others
PC1
PC2
c
functional group
RCN
RCCH
ROR
RCOOR
RCOOH
RCOR
RCHO
ROH
ROPO3
R3N
R2NH
RNH2
PROTAC*
DC
Others
20,000
15,000
10,000
5,000
0
d
hit ratio = ∑TPI / ∑TPGT
miss ratio = ∑FNI / ∑TPGT
TPI: true positive counts for all instances
FNI: false negative counts for all instances
TPGT: total positive ground truths for all instances
e
hit ratio
miss ratio
ratio
DC
DC
Others
Others
PROTAC*
PROTAC*
f
bropirimine
stallimycin
(XDH) miss ratio
(ESR1) hit ratio
Others miss ratio
PROTAC* miss ratio
PROTAC* hit ratio
DC hit ratio
DC miss ratio
Others hit ratio
density
ratio
g
validation dataset
type
monovalent monotarget 1
monovalent multitarget 2
multivalent monotarget 3
multivalent multitarget 4
h
molecules x 10^2
218
435
322
1439
type 1
type 2
type 3
type 4
i
type examples
type 1
oxitriptan
5HT1A
type 3
tasosartan
AGTR1
type 2
xanomeline
ACM1
DRD2
type 4
pralatrexate
FLOR1
DYR
j
target of validation dataset
DrugCentral
PROTAC*
1643
61
321
4
22
42
39
Others
k
ligands
#moleules x 10^5
28,413
365,719
376,342
2,384,073
LigAdvisor
SuperPred
SwissTarget
SynGlue
l
targets
#targets x 10^3
18,000
2,353
3,068
20,129
LigAdvisor
SuperPred
SwissTarget
SynGlue
m
tool
source
type
SuperPred
SwissTargetPrediction
LigAdvisor
SynGlue
BioSNAP
DrugBank
PDB
clinical data
ChEMBL 29
ChEMBL 23
SMS
STITCH
ChEMBL 33
BindingDB
targets
other
ligand
n
dataset
benchmark dataset
t-SNE 2
t-SNE 1
type and cluster representative
100 molecules
strategy
tool
tool dataset assesment
target prediction
calculate hit ratio and miss ratio
SCORE
frequency
hit ratio
o
LigAdvisor
SwissTarget
SuperPred
SynGlue
density
hit ratio

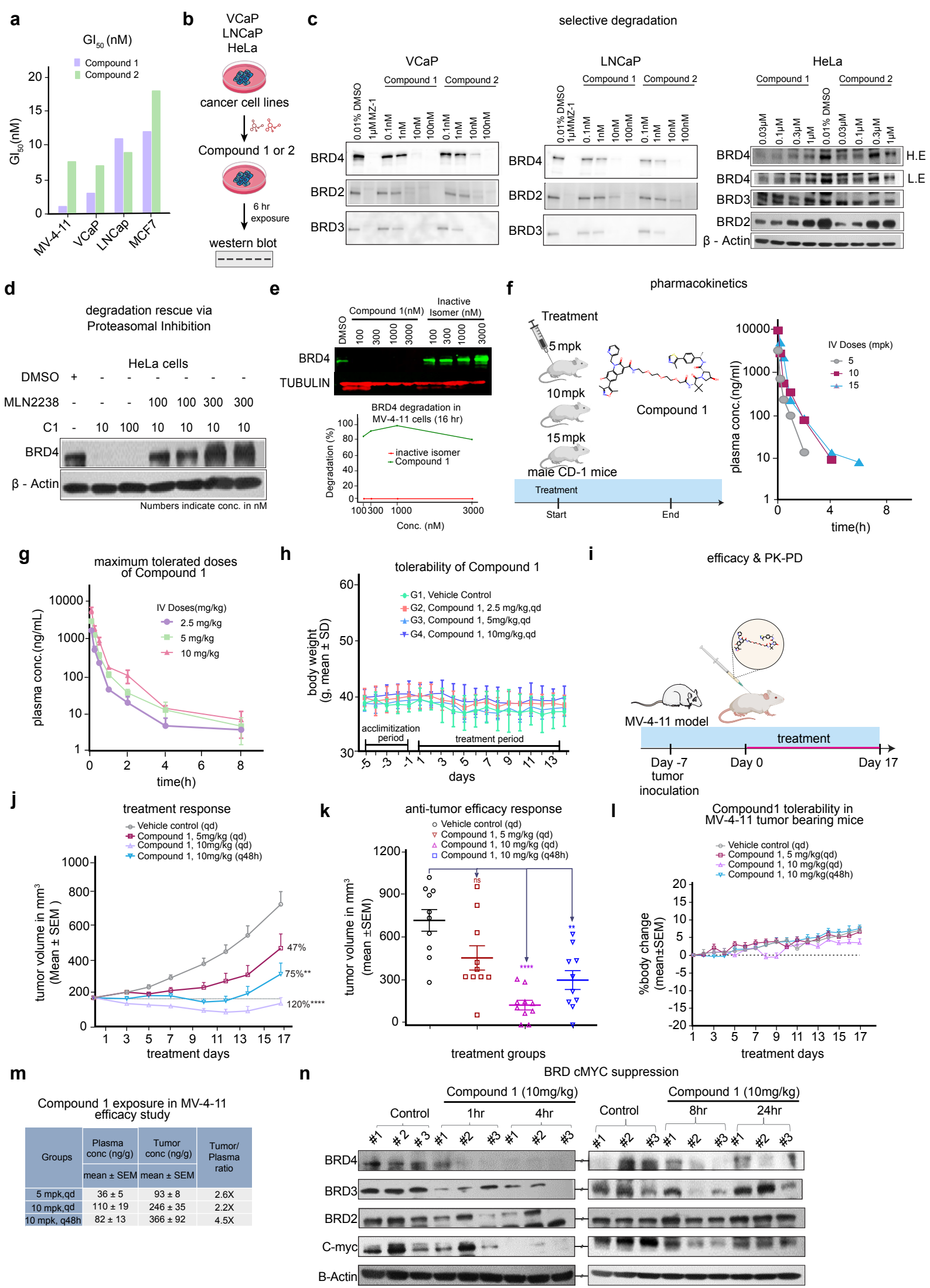
Figure 2



Metric	D_{max}	DC_{50}
R^2	0.82 ± 0.02	0.76 ± 0.01
MAE	0.14 ± 0.01	0.21 ± 0.01
MSE	0.08 ± 0.015	0.07 ± 0.01
RMSE	0.27 ± 0.05	0.23 ± 0.02

Properties	Compound1	Compound 2
MW/clogP	1021/2.5	1017/5.1
Eq Sol @pH 7.4(μM)	4.53	0.3
Caco2 permeability	poor	ND
efflux	high	ND
prediction class	0	0
Predicted DC_{50}	0.26 nM	15 nM
Predicted D_{max}	68.9%	47.3%

Figure 3



Groups	Plasma conc (ng/g)	Tumor conc (ng/g)	Tumor/ Plasma ratio
	mean ± SEM	mean ± SEM	
5 mpk,qd	36 ± 5	93 ± 8	2.6X
10 mpk,qd	110 ± 19	246 ± 35	2.2X
10 mpk, q48h	82 ± 13	366 ± 92	4.5X

Figure 4