OPEN ACCESS

Computer Science PERSPECTIVE

The Innovation Drug Discovery

Generative AI reshapes the PROTAC discovery paradigm: From empirical screening to rational design

Pengyun Li,[1,*] Shiyang Sun,[1] Ting Wei,[1] Zhanhong Tang,[2] Zhenze Qi,[1] Xu Cai,[1] Ziyun Zhang,[1] Zhihui Mu,[1] Xinna Wei,[1] Jiatao Qi,[1] Sihan Cui,[1] Junhai Xiao,[1,*] Yuquan Li,[2,*] Zhibing Zheng,[1,*] and Song Li[1,2]

[1]National Engineering Research Center for the Emergency Drug, Beijing 100039, China

[2]State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guizhou 550025, China

*Correspondence: bnuhuaxuelpy@163.com (P.L.); xiaojunhai@139.com (J.X.); yvquan.li@gzu.edu.cn (Y.L.); zbzheng@bmi.ac.cn (Z.Z.)

Received: April 7, 2026; Accepted: June 15, 2026; Published Online: June 25, 2026; https://doi.org/10.59717/j.xinn-drugdisc.2026.100027

Citation: Li P., Sun S., Wei T., et al. (2026). Generative AI reshapes the PROTAC discovery paradigm: From empirical screening to rational design. *The Innovation Drug Discovery* **1**:100027.

PROTACs are revolutionary therapeutics that eliminate pathogenic proteins via catalytic event-driven degradation. Conventional development suffers from unpredictable ternary complexes, empirical linker modification and blind ligand screening, yielding clinical success below 10% and requiring 3-4 years to obtain preclinical candidates. Generative artificial intelligence (AI) offers a promising avenue for data-driven rational design. This perspective explores how multimodal deep learning enhances ternary complex prediction accuracy and boosts molecular design hit rates from 5-8% to over 35%. Closed-loop research and development (R&D) ecosystems supported by databases accelerate PROTAC development. The landmark case of Insilico Medicine's PKMYT1-PROTAC, developed via its AI platform Chemistry 42, achieves a DC_{50} of 0.5 nM and exhibits a dual degradation-inhibition mechanism, and has advanced to the preclinical candidate (PCC) validation stage within approximately 12 months. We further discuss how to integrate AI design with clinical translation. Future efforts focus on advanced computational strategies, establishing standardized evaluation metrics and expanding the E3 ligase toolbox. This AI-powered strategy will expedite PROTAC research and expand the application of targeted protein degradation therapy.

INTRODUCTION

Targeted protein degradation (TPD) represented by proteolysis-targeting chimeras (PROTACs) has revolutionized modern medicinal chemistry beyond traditional small-molecule inhibition. As heterobifunctional molecules, PROTACs hijack the endogenous ubiquitin-proteasome system to achieve catalytic and selective degradation of target proteins, offering unique merits including sub-stoichiometric efficacy, accessibility to undruggable targets lacking active enzymatic pockets, and inherent potential to overcome drug resistance. FDA approved PROTAC ARV-471 for ESR1-mutant ER^+/HER^{2-} advanced breast cancer and numerous PROTAC candidates are currently under preclinical and clinical evaluation. Nevertheless, the clinical progression of PROTACs is severely hampered by intricate rational design challenges that require simultaneous optimization of dual POI/E3 binding affinity, favorable linker conformation for stable ternary complex formation, and balanced physicochemical and pharmacokinetic profiles for oral delivery. Conventional PROTAC development relies on empirical ligand screening, trial-and-error linker modification and low-throughput validation, yielding a hit rate below 8% and requiring 2.5-4 years for preclinical candidate acquisition. Furthermore, conventional computational tools fail to accurately predict dynamic ternary complex geometry, rendering linker optimization largely empirical. Collectively, these limitations restrict the clinical success rate of PROTACs to less than 10%, highlighting an urgent demand for advanced technical innovations.[1]

Generative AI, particularly deep generative models, has emerged as a promising approach to addressing these bottlenecks. Building upon foundational machine learning efforts in PROTAC property prediction and activity modeling, generative AI extends beyond retrospective analysis by enabling the de novo design of molecules with desired property profiles, trained on large-scale, high-quality chemical and biological datasets.[2] In this perspective, we discuss how generative AI is reshaping the PROTAC R&D landscape. We highlight core technological advances, the transition from linear workflows to closed-loop R&D ecosystems, key case studies, and outline critical challenges and future directions for AI-driven PROTAC discovery (Figure 1).

MULTIMODAL GENERATIVE AI FOR PROTAC RATIONAL DESIGN

AI-driven PROTAC design has rapidly evolved from basic predictive algorithms to sophisticated multimodal generative frameworks that address the multiscale structural and computational challenges of ternary complex modeling. The synergistic combination of Graph Neural Networks (GNNs) and Transformers constitutes a robust hybrid architecture for precise PROTAC modeling: GNNs capture the spatial characteristics, 3D interfacial geometry and multiscale interaction features of proteins and molecules to predict ternary complex stability and conformation, while Transformer models leverage self-attention mechanisms to decode long-range sequential and allosteric regulatory effects of protein and ligand sequences.[3] This multimodal integration improves ternary complex prediction accuracy from below 50% to over 77%, transforming computational modeling from retrospective analysis to prospective design guidance.[4] Furthermore, state-of-the-art diffusion models enable end-to-end de novo generation of intact PROTACs, breaking the limitations of conventional ligand modification strategies. By exploring untapped chemical space, these models generate synthetically feasible warheads and linkers for undruggable targets while optimizing binding affinity, linker configuration and lipophilicity. Emerging large vision-language models further complement this workflow by enabling multimodal structural interpretation and heterogeneous data integration to decipher complex structure-activity relationships. Collectively, these advanced AI architectures raise the PROTAC design hit rate from 5%-8% to over 35% and effectively overcoming the high-cost, time-consuming bottlenecks of traditional PROTAC research and development.[4]

THE CLOSED LOOP ECOSYSTEM FOR ACCELERATED PROTAC DEVELOPMENT

High-quality curated PROTAC databases serve as the fundamental foundation for robust generative AI model training. Representative platforms including PROTAC-PatentDB,[5] and PROTAC-DB 3.0 systematically integrate massive PROTAC-related resources,[6] ranging from structural information, degradation efficacy indicators (DC_{50} and D_{max}), selectivity profiles, physicochemical properties and pharmacokinetic parameters to computationally modeled ternary complex structures, embedding rich structure-activity relationship knowledge and critical spatial features to support rational PROTAC design. Leveraging these specialized datasets, generative AI establishes a sustainable closed-loop iterative R&D workflow for PROTAC development. Continuously updated datasets containing both active and inactive compounds provide standardized training resources for deep learning architectures, which generate novel PROTAC libraries compliant with predefined target product profiles. Subsequent in silico validations covering molecular docking, ternary complex prediction, ADMET evaluation and synthetic accessibility assessment screen high-quality lead candidates, which are further synthesized automatically and evaluated via high-throughput biological assays. New experimental data are ultimately supplemented into databases to iteratively retrain and optimize AI models, forming a self-upgrading design system. This innovative closed-loop paradigm effectively resolves the low-efficiency drawbacks of traditional linear PROTAC development pipelines and significantly accelerates the clinical translation of targeted protein degradation therapeutics. Nevertheless, large-scale application still faces notable practical constraints. Major hurdles include inconsistent cross-platform data

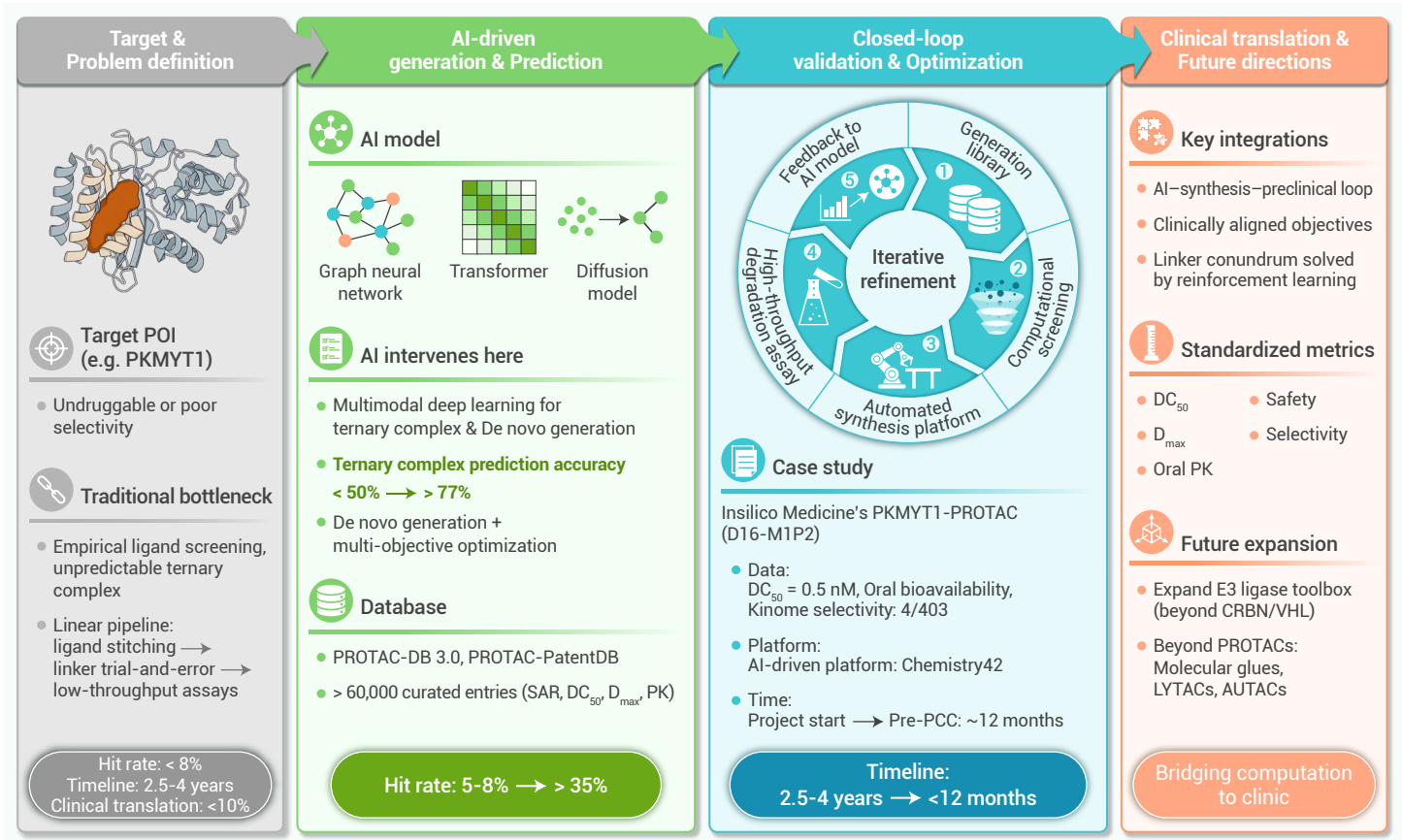

Figure 1. Schematic diagram of generative AI expediting the PROTAC research and translational pipeline Generative AI tools reshape the four-step PROTAC research workflow, resolving traditional bottlenecks in ternary complex prediction and low-throughput empirical screening via multi-objective de novo molecular generation and closed-loop automated validation and optimization. The PKMYT1-PROTAC program developed on the Chemistry42 platform demonstrates substantial improvements in hit rate, predictive accuracy as well as project timelines, with integrated preclinical translational benchmarks. Future directions for targeted protein degrader modalities beyond canonical PROTACs are outlined.

standards, poor compatibility between AI-generated molecules and automated synthetic workflows, and high experimental verification costs restrict cross-institutional data sharing.

INSILICO MEDICINE'S PKMYT1 PROTAC

As a promising synthetic lethal target in genomically aberrant cancers, PKMYT1 remains poorly targeted by conventional small-molecule inhibitors due to inadequate selectivity, dose-limiting toxicity, and suboptimal efficacy. Insilico Medicine's end-to-end generative AI platform Chemistry42 enabled the streamlined development of the PKMYT1-targeted PROTAC via a multi-step rational workflow.[7] The platform generated over 2000 novel de novo PKMYT1 ligands to bypass the constraints of traditional ligand-based optimization, followed by AI-guided hit refinement, rational linker design, and conjugation of PKMYT1 warheads with CRBN-recruiting moieties. Synchronous multi-objective optimization of ternary complex assembly, degradation potency, target selectivity, metabolic stability, and oral bioavailability yielded well-balanced, druggable candidates. The optimized lead D16-M1P2 achieves potent PKMYT1 degradation (DC_{50} = 0.5 nM) and direct enzymatic inhibition (IC_{50} = 7.6 nM). It also displays exceptional kinome selectivity, favorable physicochemical and *in vivo* pharmacokinetic properties. Notably, the entire pipeline from project initiation to pre-PCC screening was completed within 12 months, and D16-M1P2 has advanced to formal PCC validation, positioning it as one of the most mature AI-designed PROTAC candidates.[8] This landmark case substantiates the capacity of generative AI to accelerate R&D cycles and enable the rational discovery of high-quality, clinically viable PROTAC therapeutics. Direct comparative studies with conventionally designed PROTACs and systematic tracking of AI-designed molecules through clinical development will be important to establish robust evidence as the field matures.

INTEGRATING AI DESIGN WITH CLINICAL TRANSLATION

The advanced AI-driven PROTAC optimization paradigm relies on three core technical advancements to improve design rationality and clinical translatability. First, a closed-loop workflow integrating AI design, automated synthesis, and preclinical validation enables iterative model refinement, wherein synthetic accessibility is prioritized during molecular optimization to exclude unfeasible structures, and real-time data from high-throughput biological assays, and *in vivo* PK/PD profiling is digitized and fed back to update design algorithms. Second, current AI design pipelines are recalibrated to align computational optimization with clinical target product profiles, whereby rational modulation of molecular physicochemical properties improves oral bioavailability, tumor penetration, metabolic stability, and target selectivity, bridging the gap between in silico prediction and practical bench-to-bedside translation. Third, reinforcement learning-enhanced generative AI addresses the longstanding linker conundrum; it efficiently explores broad linker chemical space and decodes the undefined nonlinear structure-activity relationships between linker features and ternary complex formation for specific POI-E3 ligase pairs.

DISCUSSION AND FUTURE PERSPECTIVES

Despite advances in this field, generative AI still encounters core technical bottlenecks in PROTAC design. First, training datasets exhibit pronounced bias, which are largely composed of CRBN- and VHL-recruiting degraders. Such imbalanced data substantially impairs model generalization capacity toward less-studied E3 ligases and novel target variants. Second, poor model interpretability hinders chemists from deciphering AI-derived molecular structures and distilling actionable structure-activity relationships. Third, these models show limited cross-target transferability, with dramatic performance degradation when applied to unreported targets beyond training datasets. Fourth, evident discrepancies persist between computational

predictions and experimental outcomes. Numerous theoretically optimal candidates fail wet-lab validation owing to neglected druggability properties. Furthermore, existing regulatory and intellectual property frameworks have not been fully adapted to AI-originated therapeutics. Unresolved concerns cover patent eligibility of AI-designed molecules, data traceability and privacy protection of cross-institutional training data, as well as unified evaluation criteria for clinical translation of AI-generated drug leads. Nevertheless, addressing several core challenges and prioritizing key future directions is essential to mature the field and fully unlock its translational and clinical potential.

Advanced computational strategies

To tackle persistent challenges in PROTAC linker design and multi-objective optimization, next-generation AI frameworks adopt advanced computational strategies. Modular feature extraction disassembles PROTAC molecules into warhead, linker, and E3 ligand modules, allowing the model to capture transferable component-specific design rules, improve generalizability, and support independent optimization and combinatorial refinement of novel PROTAC constructs. Furthermore, the integration of generative models with reinforcement learning enables systematic exploration of the expansive linker chemical space. Guided by wet-lab experimental feedback, such frameworks autonomously identify the complex nonlinear correlations between structural features and degradation performance, enabling iterative optimization of PROTAC design schemes.

Establishing standardized evaluation metrics for clinical translation

A major obstacle limiting the advancement of AI-designed PROTACs is the lack of standardized computational and experimental evaluation criteria. Standardized computational metrics includes validated cutoffs for ternary complex prediction confidence, synthetic accessibility, and ADMET properties correlated with *in vivo* druggability; systematic retrospective analysis of existing PROTAC datasets can establish such criteria to enable rigorous pre-experimental compound filtering and reduce unnecessary synthetic and biological validation costs. Additionally, unified preclinical evaluation frameworks including degradation efficiency, druggability, safety and selectivity should be standardized. Furthermore, standardized minimum reporting criteria covering SMILES strings, predicted/experimental ternary complex conformations, training dataset composition, model architectures, and full experimental protocols should be established for AI-designed PROTACs, thereby ensuring cross-laboratory reproducibility and independent validation.

Expanding the E3 ligase toolbox

Currently, the vast majority of PROTACs rely on a highly limited subset of E3 ubiquitin ligases—predominantly CRBN and VHL. To address these bottlenecks, the PROTAC field has expanded E3 ligase toolkits toward non-CRBN/VHL subtypes, with DCAF15, DCAF11 and KEAP1 identified as promising alternatives. However, most newly exploited E3 ligases lack potent and well-validated ligands, while conventional experimental strategies exhibit low efficiency in ligand mining and optimization, substantially restricting their broader application in PROTAC development.[9] By contrast, generative AI offers distinct strengths to overcome this bottleneck, greatly accelerating ligand discovery, high-throughput screening and functional validation of novel E3 ligases.

CONCLUSION

Generative AI has revolutionized PROTAC research by shifting traditional empirical screening toward efficient, data-driven rational design, effectively resolving longstanding bottlenecks in conventional PROTAC development. The PKMYT1-targeted PROTAC program developed by Insilico Medicine exemplifies this advancement, with the ongoing clinical progression of this candidate further validating the translational reliability of AI-designed PROTACs. Maximizing such technological advantages relies on closed-loop R&D ecosystems that tightly integrate computational modeling, experimental validation, and high-quality specialized PROTAC databases. At the interface of generative AI and TPD, considerable translational potential remains untapped. Generative AI offers a compelling pathway to address the limitations inherent to conventional empirical approaches, advancing the field toward rapid, rational, and robust PROTAC drug discovery. This transformative innovation is poised to deliver next-generation therapeutics for refractory diseases that have long remained intractable to conventional pharmacological interventions.

REFERENCES

1. Guedeney N, Cornu M, Schwalen F, et al. (2023). PROTAC technology: A new drug design for chemical biology with many challenges in drug discovery. *Drug Discov Today* **28**:103395. DOI:10.1016/j.drudis.2022.103395
2. Ge J, Hsieh C-Y, Fang M, et al. (2024). Development of PROTACs using computational approaches. *Trends Pharmacol Sci* **45**:1162–1174. DOI:10.1016/j.tips.2024.10.006
3. Dou B, Zhu Z, Merkurjev E, et al. (2023). Machine Learning Methods for Small Data Challenges in Molecular Science. *Chem Rev* **123**:8736–8780. DOI:10.1021/acs.chemrev.3c00189
4. Chen Z, Gu C, Tan S, et al. (2025). Interpretable PROTAC Degradation Prediction With Structure-Informed Deep Ternary Attention Framework. *Adv Sci* **12**:08138. DOI:10.1002/advs.202508138
5. Cai H, Yao G, Shi Y, et al. (2025). PROTAC-PatentDB: A PROTAC Patent Compound Dataset. *Sci Data* **12**:1840. DOI:10.1038/s41597-025-06136-9
6. Ge J, Li S, Weng G, et al. (2025). PROTAC-DB 3.0: an updated database of PROTACs with extended pharmacokinetic parameters. *Nucleic Acids Res* ***53***:1510-1515. DOI:10.1093/nar/gkae768.
7. Ivanenkov YA, Polykovskiy D, Bezrukov D, et al. (2023). Chemistry42: An AI-Driven Platform for Molecular Design and Optimization. *J Chem Inf Model* **63**:695–701. DOI:10.1021/acs.jcim.2c01191
8. Wang Y, Wang X, Liu T, et al. (2025). Discovery of a bifunctional PKMYT1-targeting PROTAC empowered by AI-generation. *Nat Commun* **16**:10759. DOI:10.1038/s41467-025-65796-8
9. Schapira M, Calabrese MF, Bullock AN, et al. (2019). Targeted protein degradation: expanding the toolbox. *Nat Rev Drug Discov* **18**:949–963. DOI:10.1038/s41573-019-0047-y

FUNDING AND ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 82404417) and Beijing Natural Science Foundation (No. 7264376). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.