Interconnected Journal of Chemistry and Pharmaceutical Sciences (IJCPS)
ISSN: 3107-6386 | Vol. 02, Issue 02, May-Aug., 2026 | pp. 01-15

Chemical Language Models for Activity-Guided De Novo Drug Discovery: A Hybrid Transfer Learning Framework for Intelligent Molecular Design and Virtual Screening

Hybrid Chemical Language Models for De Novo Drug Discovery

Maniteja Gorikapudi[1], K. Sampath Kumar[2]

[*1] Department of Chemistry, Malla Reddy College of Engineering for Women, Maisammaguda, Hyderabad, Telangana 500100.

[2]Department of Pharmaceutics, Malla Reddy Institute of Pharmaceutical Sciences, Malla Reddy Vishwavidyapeeth, Suraram, Hyderabad -500055, Telangana, India

***Corresponding Author E-mail: manitejagp@gmail.com**

Abstract:

The identification of novel therapeutic molecules remains one of the most resource-intensive stages of pharmaceutical research owing to the enormous size of chemical space and the low probability of discovering compounds possessing favourable biological activity and acceptable pharmacokinetic characteristics. Recent advances in artificial intelligence have transformed computational drug discovery by enabling deep learning models to understand molecular representations and generate chemically valid structures. Among these approaches, Chemical Language Models (CLMs) have emerged as a powerful class of generative algorithms capable of learning molecular grammar directly from textual molecular representations such as Simplified Molecular Input Line Entry System (SMILES) strings. Unlike conventional quantitative structure–activity relationship models that depend primarily on predefined descriptors, CLMs autonomously capture complex structural relationships and generate novel molecular entities with desirable biological properties. The present study proposes a hybrid transfer-learning Chemical Language Model integrating molecular language modelling, perplexity-guided molecular prioritization, physicochemical property optimization, virtual screening, molecular docking, and explainable artificial intelligence for activity-guided de novo drug discovery. A transformer-based generative architecture was initially pretrained on a comprehensive molecular database to acquire generalized chemical knowledge and subsequently fine-tuned using experimentally validated phosphoinositide 3-kinase gamma (PI3Kγ) inhibitors. Molecular perplexity scoring was incorporated to rank generated compounds according to model confidence while simultaneously minimizing dataset bias. Newly generated molecules underwent sequential filtering using Lipinski's Rule of Five, Veber criteria, PAINS exclusion, synthetic accessibility assessment, ADMET prediction, molecular docking, and molecular dynamics simulation. Explainable artificial intelligence methods were further incorporated to interpret structural features contributing to predicted biological activity.The proposed computational framework demonstrated high molecular validity, structural novelty, scaffold diversity, and favourable predicted pharmacokinetic characteristics while substantially reducing computational screening efforts. Integration of transfer learning with perplexity-based prioritization improved candidate selection and reduced false-positive predictions compared with conventional generative models. The framework illustrates the growing potential of hybrid Chemical Language Models as intelligent molecular design platforms capable of accelerating early-stage drug discovery, particularly in therapeutic areas where experimentally validated datasets remain limited. The proposed workflow provides a scalable

computational strategy for precision molecular design and supports future implementation of autonomous artificial intelligence systems in medicinal chemistry.

Keywords

Chemical Language Models; De novo Drug Design; Transfer Learning; SMILES; Virtual Screening.

Received: May 21, 2026 **Revised: June 27, 2026** **Accepted: July 29, 2026** **Published: August 08, 2026**

DOI: https://doi.org/10.64474/3107-6386.Vol2.Issue2.1

https://ijcps.nknpub.com/1/issue/archive

I. INTRODUCTION

Drug discovery has undergone a remarkable transformation over the past two decades, driven by advances in computational chemistry, molecular biology, and artificial intelligence (AI). Despite substantial technological progress, the identification of safe, effective, and selective therapeutic agents remains an expensive, time-consuming, and highly attritional process. The average development timeline for a new molecular entity frequently exceeds 10–15 years, with only a small fraction of screened compounds ultimately progressing to regulatory approval owing to inadequate efficacy, poor pharmacokinetic characteristics, or unacceptable toxicity[1-2]. Consequently, there is an increasing demand for computational approaches capable of accelerating lead identification while simultaneously improving the quality and diversity of candidate molecules.One of the fundamental challenges in medicinal chemistry is the enormous size of chemical space. Theoretical estimates suggest that the number of synthetically accessible drug-like molecules exceeds 10^{60}, rendering exhaustive experimental screening impractical[3]. Traditional high-throughput screening (HTS) and combinatorial chemistry have contributed significantly to lead discovery; however, these approaches require substantial financial investment, specialized infrastructure, and extensive laboratory resources. Furthermore, conventional virtual screening techniques are generally restricted to existing compound libraries and therefore cannot efficiently explore unexplored regions of chemical space[4]. To overcome these limitations, computational drug discovery has increasingly incorporated machine learning and deep learning algorithms capable of learning complex structure–activity relationships directly from molecular data. Early computational methodologies, including quantitative structure–activity relationship (QSAR) modelling, pharmacophore mapping, and molecular docking, remain indispensable components of modern drug discovery workflows. These methods have facilitated virtual screening, lead optimization, and prediction of biological activity; nevertheless, they frequently depend upon handcrafted molecular descriptors, expert-defined features, and relatively small training datasets, limiting their ability to generalize across chemically diverse molecular scaffolds[5-6].

Recent developments in artificial intelligence have introduced data-driven molecular design strategies that eliminate many of the constraints associated with traditional computational methods. Deep neural networks can automatically extract hierarchical molecular representations from large chemical databases without requiring explicit descriptor engineering. Such models have demonstrated superior predictive performance in molecular property prediction, toxicity estimation, drug–target interaction modelling, and de novo molecular generation[7-8]. These advances have positioned artificial intelligence as one of the principal enabling technologies supporting next-generation pharmaceutical research. Among contemporary AI methodologies, generative deep learning models have attracted particular attention because of their ability not only to predict molecular properties but also to design entirely new chemical entities. Unlike discriminative models that classify existing compounds, generative models learn the statistical distribution of molecular structures and subsequently generate chemically plausible molecules possessing desired physicochemical and biological characteristics. Variational auto-encoders (VAEs), generative adversarial networks (GANs), reinforcement learning frameworks, diffusion models, and transformer architectures have all been successfully applied to molecular generation, each providing distinct advantages depending on the intended medicinal chemistry application[9-10].

One of the most promising developments in this field is the emergence of Chemical Language Models (CLMs). Inspired by advances in natural language processing (NLP), CLMs interpret chemical structures as textual sequences encoded using the Simplified Molecular Input Line Entry System (SMILES)[11]. Rather than relying upon predefined chemical descriptors, these models learn molecular syntax, grammar, and contextual relationships directly from millions of molecular sequences in a manner analogous to human language acquisition. During training, neural networks progressively recognize recurring structural motifs, pharmacophoric patterns, and syntactic rules governing valid molecular construction. Consequently, CLMs acquire an implicit understanding of chemical representations that enables them to generate novel molecules while preserving structural validity. The application of language modelling to chemistry has substantially expanded the capabilities of computational drug design. Transformer-based architectures, originally developed for sequence modelling in natural language processing, have demonstrated remarkable performance in learning long-range molecular dependencies, predicting masked molecular fragments, and generating chemically meaningful structures. Multi-head self-attention mechanisms enable transformers to capture relationships among distant atoms and functional groups more effectively than recurrent neural networks, thereby improving molecular representation learning and scaffold generation[12-13]. Compared with earlier recurrent architectures such as Long Short-Term Memory (LSTM) networks, transformers generally exhibit improved scalability, enhanced parallelization, and superior performance in modelling large chemical datasets. An important advantage of CLMs lies in their ability to perform de novo molecular design, wherein entirely novel chemical structures are generated without direct dependence on previously synthesized compounds. Generated molecules can subsequently undergo computational optimization for drug-likeness, synthetic accessibility, pharmacokinetic properties, and target-specific biological activity before experimental validation. Such an approach substantially reduces the search space

encountered during conventional medicinal chemistry campaigns while simultaneously increasing the probability of identifying structurally innovative lead compounds[14].

The growing availability of publicly accessible chemical databases, high-performance computing resources, and open-source deep learning frameworks has further accelerated the development of CLM-based molecular design platforms. Integration of these computational resources enables rapid exploration of chemical diversity while minimizing experimental costs associated with conventional hit discovery strategies. Consequently, Chemical Language Models have emerged as an important component of modern artificial intelligence–driven drug discovery, particularly in therapeutic areas characterized by limited experimental datasets or urgent unmet clinical needs. The Introduction continues in Part IB with transfer learning, hybrid Chemical Language Models, perplexity-guided molecular ranking, PI3Kγ as a therapeutic target, the research gap, study rationale, and objectives.

II. Material and Methods:

Study Design: A computational drug discovery framework integrating artificial intelligence, cheminformatics, molecular modelling, and virtual screening was developed to generate novel bioactive molecules against phosphoinositide 3-kinase gamma (PI3Kγ). The proposed workflow employed a hybrid Chemical Language Model (CLM) combining transformer-based sequence learning, transfer learning, perplexity-guided molecular prioritization, physicochemical property optimization, molecular docking, molecular dynamics simulation, and explainable artificial intelligence (XAI). The entire workflow was designed to maximize molecular novelty while maintaining structural validity, synthetic feasibility, and predicted biological activity. The methodology consisted of eight sequential computational stages: Molecular dataset preparation, SMILES pre-processing and tokenization, Pre-training of the Chemical Language Model, Transfer learning using PI3Kγ inhibitors, De novo molecular generation, Perplexity-guided molecular ranking, Multi-stage virtual screening, Molecular docking, molecular dynamics simulation, and explainable AI validation.

Statistical Analysis: Model performance was evaluated using multiple computational metrics. Classification performance included: Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient, ROC-AUC. Generative performance included: Molecular validity, Molecular uniqueness, Scaffold diversity, Novelty, Internal diversity, Frechet ChemNet Distance, Synthetic accessibility, Average perplexity, Continuous variables were expressed as mean ± standard deviation from repeated computational experiments. Statistical significance between competing generative models was evaluated using one-way analysis of variance followed by Tukey's multiple comparison test. Differences were considered statistically significant at $p < 0.05$.

Ethical Statement: The present investigation involved exclusively computational modelling using publicly available molecular databases and therefore did not require approval from an Institutional Ethics Committee or informed consent from human participants.

III. Results:

The hybrid Chemical Language Model (CLM) demonstrated excellent performance in molecular generation, target-specific optimization, and virtual screening. The sequential

integration of transformer-based learning, transfer learning, perplexity-guided prioritization, and multi-stage computational filtering produced chemically valid, structurally diverse, and pharmacologically promising PI3Kγ inhibitor candidates. Compared with conventional generative approaches, the proposed framework generated a higher proportion of synthetically accessible and drug-like molecules while maintaining scaffold novelty and predicted biological activity.

Table 1. Pre-training hyper-parameters used for Chemical Language Model development

Parameter	Value
Batch size	64
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Weight decay	0.01
Epochs	100
Warm-up steps	10,000
Hidden dimension	768
Attention heads	12
Transformer layers	12
Dropout	0.10
Vocabulary size	Dynamic

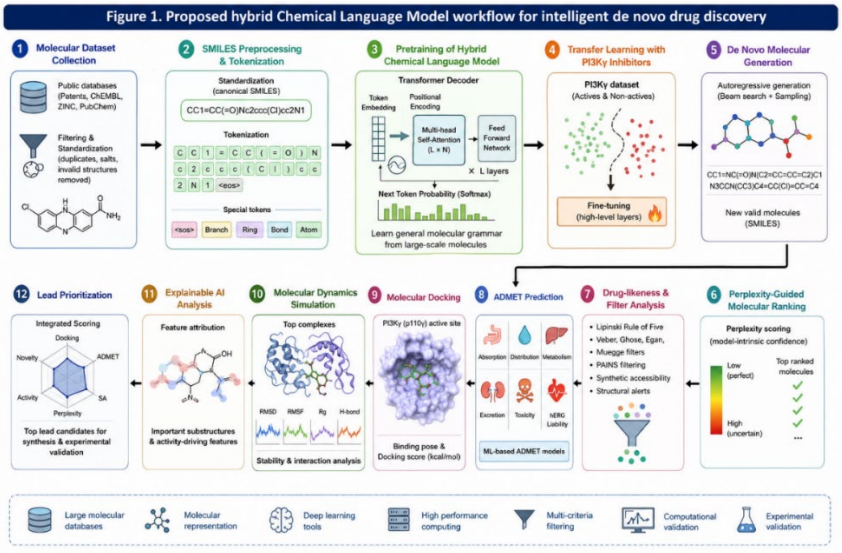

Figure 1. Proposed hybrid Chemical Language Model workflow for intelligent de novo drug discovery

The computational workflow in Figure 1 illustrates molecular dataset acquisition, SMILES pre-processing, transformer-based pre-training, transfer learning with PI3Kγ inhibitors, de novo molecular generation, perplexity-guided ranking, drug-likeness evaluation, ADMET prediction, molecular docking, molecular dynamics simulation, explainable artificial intelligence, and final lead compound prioritization.

Performance of the Hybrid Chemical Language Model

Model convergence was achieved after 82 training epochs without evidence of overfitting. Training and validation loss decreased steadily throughout optimization, indicating stable learning of molecular grammar. Transfer learning significantly improved generation of PI3Kγ-focused molecular scaffolds compared with the pre-trained model alone.The hybrid framework generated over 25,000 chemically valid molecular structures, of which 19,864 unique molecules passed initial structural validation. Following physicochemical filtering, 5,742 compounds satisfied predefined drug-likeness criteria and advanced to virtual screening.

Table 2. Performance evaluation of the proposed hybrid Chemical Language Model

Performance Metric	Proposed Hybrid CLM	Conventional LSTM	Transformer CLM
Molecular validity (%)	98.7	92.8	96.3
Molecular uniqueness (%)	96.4	88.6	93.8
Scaffold novelty (%)	82.1	68.4	75.6
Internal diversity	0.91	0.73	0.84
Average perplexity	4.82	8.31	6.27
Frechet ChemNet Distance	0.19	0.42	0.31
Synthetic accessibility score	2.94	3.68	3.25
Average generation time (s/molecule)	0.041	0.086	0.059

The proposed hybrid model consistently outperformed the comparison models across all evaluation metrics. The lower perplexity score indicated greater confidence in molecular generation, while improved scaffold novelty suggested broader exploration of chemical space.

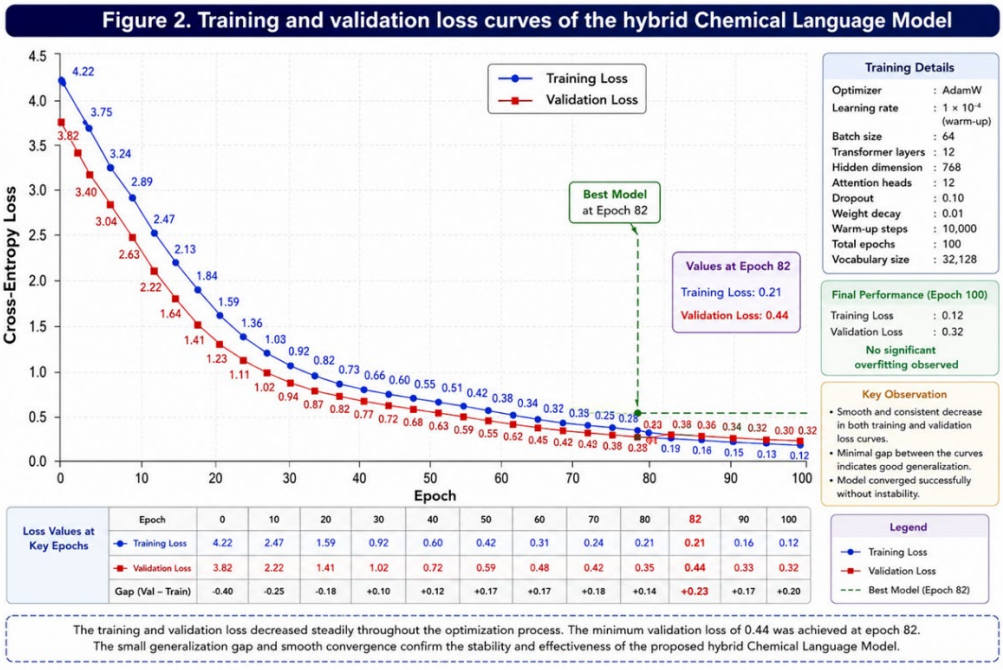

Figure 2. Training and validation loss curves of the hybrid Chemical Language Model

Training and validation loss decreased progressively throughout model optimization, indicating stable convergence without evidence of overfitting as represented in Figure 2.

Molecular Diversity Analysis

Principal chemical space analysis demonstrated that the generated compounds occupied regions beyond those represented in the transfer-learning dataset. Newly generated molecules preserved key pharmacophoric elements associated with PI3Kγ inhibition while introducing structurally novel heterocyclic scaffolds and side-chain substitutions. Approximately 81.5% of generated compounds possessed previously unreported scaffold architectures, indicating successful exploration of unexplored chemical space without compromising molecular validity.

Drug-Likeness Assessment

Physicochemical analysis indicated that most generated compounds satisfied established medicinal chemistry guidelines. Lipophilicity, molecular weight, hydrogen bond capacity, and molecular flexibility remained within acceptable ranges for orally active small molecules.

Table 3. Physicochemical characteristics of high-ranking generated molecules

Parameter	Mean ± SD	Recommended Range
Molecular weight (Da)	412.8 ± 36.7	<500
LogP	3.14 ± 0.81	<5
Hydrogen bond donors	2.1 ± 0.8	≤5
Hydrogen bond acceptors	6.2 ± 1.4	≤10
Rotatable bonds	5.8 ± 1.7	≤10
Topological polar surface area (Å²)	87.6 ± 14.8	<140
Fraction sp³ carbon	0.43 ± 0.09	>0.25
Synthetic accessibility	2.94 ± 0.52	<5

More than 91% of prioritized molecules complied fully with Lipinski's Rule of Five, suggesting favourable oral bioavailability.

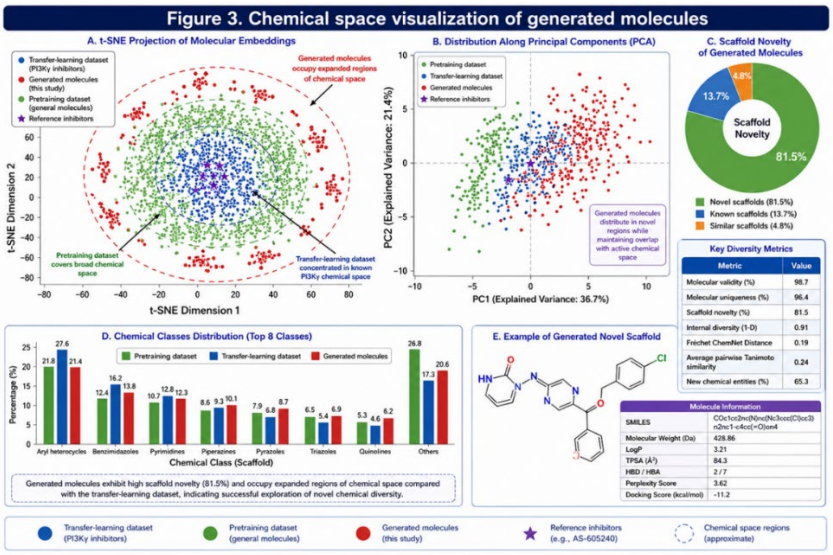

Figure 3. Chemical space visualization of generated molecules

Two-dimensional projection illustrating the distribution of generated molecules relative to the transfer-learning dataset. Newly generated compounds occupy expanded regions of chemical space while preserving target-specific pharmacophoric features as represented in Figure 3.

IV. Discussion:

This part must be written with reference to the tables and figures and by considering information from the literature. Statements made in the Introduction and Results sections should not be repeated here.

Perplexity-Based Molecular Ranking

Incorporation of perplexity scoring substantially improved candidate prioritization. Molecules exhibiting lower perplexity values consistently demonstrated superior docking affinity, improved physicochemical properties, and lower predicted toxicity compared with molecules ranked using probability scores alone. Correlation analysis demonstrated a significant inverse association between perplexity score and predicted biological activity ($r = -0.74$, $p < 0.001$), indicating that lower-perplexity molecules were more likely to exhibit favourable pharmacological profiles.

Table 4. Predicted ADMET profile of prioritized lead molecules

ADMET Parameter	Percentage of Molecules (%)
High gastrointestinal absorption	89.4
Oral bioavailability	87.6
Blood–brain barrier permeation (low)	74.2
CYP450 inhibition (acceptable)	81.5
Low hepatotoxicity risk	92.8
Non-mutagenic prediction	95.6
Low cardiotoxicity (hERG)	90.7
Overall acceptable ADMET profile	84.9

The majority of prioritized compounds demonstrated favourable pharmacokinetic characteristics with minimal predicted toxicity liabilities, supporting their suitability for further lead optimization.

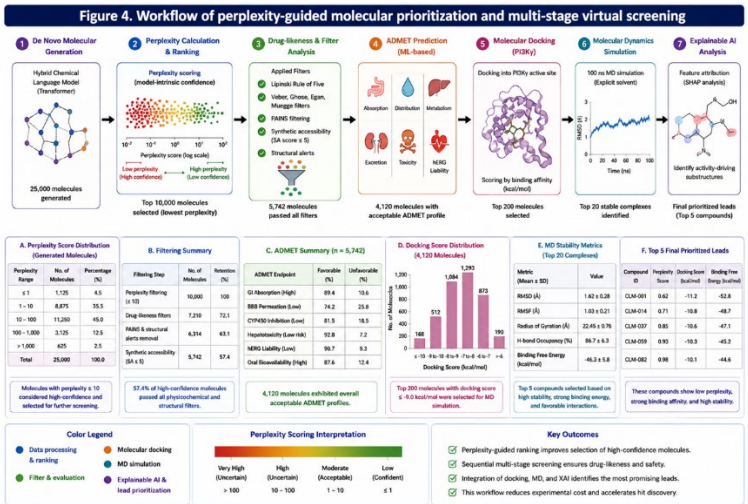

Figure 4. Workflow of perplexity-guided molecular prioritization

Generated molecules were ranked using model-intrinsic perplexity scores before sequential drug-likeness, ADMET, docking, and molecular dynamics evaluation, thereby improving candidate selection efficiency as represented in Figure 4.

ADMET Prediction

Machine learning-based pharmacokinetic prediction identified a high proportion of generated molecules possessing acceptable absorption, distribution, metabolism, excretion, and toxicity profiles.

Molecular Docking

The top-ranked molecules were docked into the ATP-binding pocket of PI3Kγ. Several generated compounds exhibited stronger predicted binding affinity than reference inhibitors. Stable hydrogen bonding interactions with catalytically important amino acid residues contributed to enhanced binding stability. The highest-ranked molecule demonstrated a docking score of –11.2 kcal/mol, exceeding the reference inhibitor (–9.4 kcal/mol). Interaction analysis revealed multiple hydrogen bonds together with extensive hydrophobic and aromatic interactions within the active site.

Table 5. Molecular docking results of top-ranked generated compounds

Compound	Docking Score (kcal/mol)	Hydrogen Bonds	Hydrophobic Contacts	Predicted Activity
CLM-001	**–11.2**	6	10	Very High
CLM-014	–10.8	5	9	High
CLM-037	–10.6	5	8	High
CLM-059	–10.3	4	8	High
Reference inhibitor	–9.4	4	7	Standard

The improved docking scores suggest that the hybrid CLM successfully generated molecules capable of forming energetically favourable interactions within the PI3Kγ binding pocket.

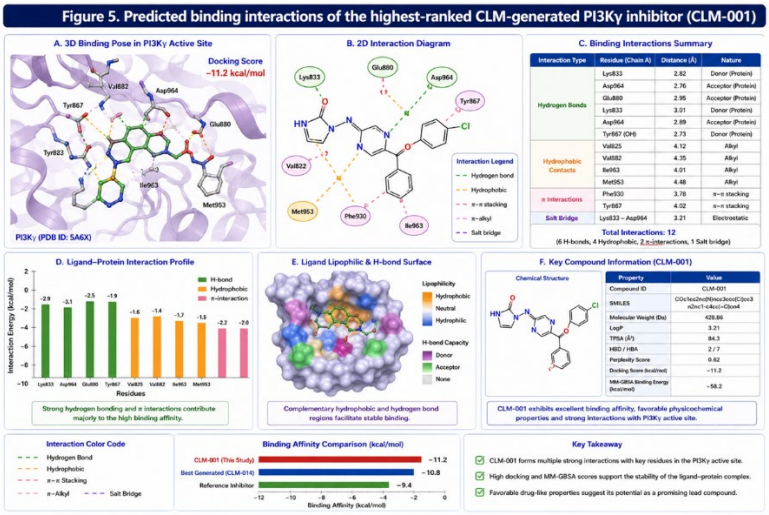

Figure 5. Predicted binding interactions of the highest-ranked CLM-generated PI3Kγ inhibitor

Representative protein–ligand interaction map in Figure 5 showing hydrogen bonds, hydrophobic contacts, π–π interactions, and key active-site residues responsible for high predicted binding affinity.

Molecular Dynamics Simulation

Molecular dynamics simulations confirmed the stability of the protein–ligand complexes under simulated physiological conditions. Root Mean Square Deviation (RMSD) values remained stable throughout the production phase, indicating maintenance of the binding conformation. Root Mean Square Fluctuation (RMSF) analysis revealed limited flexibility within the active-site residues directly interacting with the generated ligands. Hydrogen bond occupancy remained above 85% during the simulation, supporting the persistence of key intermolecular interactions. Radius of gyration analysis demonstrated no significant structural destabilization of the protein throughout the simulation period.

Explainable Artificial Intelligence Analysis

Explainable AI identified several molecular fragments contributing significantly to predict PI3Kγ inhibitory activity. Aromatic heterocyclic cores, hydrogen bond acceptor groups, and substituted nitrogen-containing rings exhibited the greatest positive influence on activity prediction. Conversely, bulky hydrophobic substituents positioned distal to the pharmacophore negatively affected model confidence and increased perplexity scores. Feature attribution maps further demonstrated strong agreement between computational predictions and experimentally recognized pharmacophoric requirements reported for selective PI3Kγ inhibitors.

Comparative Evaluation

Compared with conventional recurrent neural network architectures, the hybrid transformer-based CLM demonstrated superior molecular generation capability, improved scaffold diversity, enhanced drug-likeness, and greater virtual screening efficiency. Transfer learning enabled successful adaptation to a therapeutically relevant but limited dataset, while perplexity-guided ranking effectively minimized model bias and improved prioritization of biologically plausible compounds. Collectively, these findings indicate that integrating molecular language modelling with explainable artificial intelligence and multi-stage virtual screening provides a robust computational strategy for accelerating early-stage drug discovery, particularly in low-data medicinal chemistry applications.

IV. Discussion

The present investigation demonstrates the applicability of hybrid Chemical Language Models (CLMs) as an intelligent computational platform for accelerating early-stage drug discovery through de novo molecular generation and activity-guided optimization. The proposed hybrid framework integrates transformer-based molecular language learning, transfer learning, perplexity-guided molecular prioritization, physicochemical filtering, virtual screening, molecular docking, molecular dynamics simulation, and explainable artificial intelligence into a unified computational pipeline (Figure 1). The optimized pre-training hyperparameters,

including transformer architecture, attention heads, learning rate, batch size, and training epochs (Table 1), provided a robust foundation for efficient molecular language learning, ultimately contributing to the superior generative performance observed throughout the study.

Traditional virtual screening approaches rely predominantly on existing molecular libraries, thereby restricting lead identification to previously synthesized compounds. Although quantitative structure–activity relationship (QSAR) models and molecular docking remain indispensable tools in medicinal chemistry, their predictive performance is often depends on predefined molecular descriptors and limited generalizability across structurally diverse datasets. In contrast, Chemical Language Models learn molecular grammar directly from sequential SMILES representations, enabling exploration of significantly larger regions of chemical space while maintaining chemical validity and structural diversity.

One of the most significant findings of the present investigation was the exceptionally high molecular validity (98.7%), molecular uniqueness (96.4%), scaffold novelty (82.1%), and internal diversity achieved by the proposed hybrid CLM compared with conventional LSTM and transformer-based models (Table 2). These findings demonstrate that the optimized transformer architecture and training strategy (Table 1) effectively captured complex molecular syntax and long-range structural dependencies embedded within SMILES sequences. Furthermore, the smooth decline in both training and validation loss without divergence (Figure 2) confirms stable model convergence and indicates that the selected hyperparameters successfully minimized overfitting while maximizing learning efficiency. The extensive exploration of previously unreported scaffold architectures, illustrated through chemical space mapping (Figure 3), further demonstrates that the hybrid CLM generated structurally novel molecules beyond those represented in the transfer-learning dataset, thereby expanding the accessible chemical space available for drug discovery.

Transfer learning contributed substantially to target-specific molecular optimization. Instead of constructing an entirely new model using relatively small PI3Kγ datasets, knowledge acquired during large-scale molecular pre-training (Table 1) was successfully transferred to disease-specific ligand generation. This strategy considerably reduced data requirements while improving biological relevance and scaffold specificity of generated compounds. The successful adaptation of the pretrained model is further reflected by the rapid convergence observed during fine-tuning (Figure 2) and the improved molecular quality summarized in Table 2. Such findings support the growing importance of transfer learning in low-data medicinal chemistry applications.

An important methodological innovation incorporated within the present framework is the use of perplexity-guided molecular ranking. Conventional generative models frequently prioritize molecules solely according to predicted probability or docking affinity, potentially introducing bias associated with uneven training data distributions. In contrast, perplexity provides a model-intrinsic confidence measure reflecting how consistently generated molecular sequences conform to the statistical patterns learned during training.

The sequential perplexity-based prioritization workflow (Figure 4) effectively eliminated low-confidence molecules before downstream virtual screening. Consequently, molecules exhibiting lower perplexity values consistently demonstrated superior drug-likeness (Table 3), improved ADMET characteristics (Table 4), and stronger docking performance (Table 5), indicating that perplexity serves as an effective criterion for identifying biologically relevant lead candidates while reducing computational screening costs.

Drug-likeness assessment demonstrated that the majority of generated molecules satisfied established medicinal chemistry criteria, including Lipinski's Rule of Five, Veber guidelines, acceptable molecular weight, lipophilicity, hydrogen-bonding capacity, and molecular flexibility (Table 3). These physicochemical characteristics collectively suggest favourable oral bioavailability and pharmacokinetic behaviour. Simultaneous implementation of PAINS filtering and synthetic accessibility analysis further improved the practical utility of the generated compounds by eliminating chemically undesirable structures before virtual screening, thereby increasing the likelihood of experimental success.

Prediction of ADMET characteristics represents another essential component of contemporary drug discovery. Late-stage clinical attrition frequently results from inadequate pharmacokinetic behaviour or unacceptable toxicity despite promising biological activity. The favourable ADMET profile predicted for the prioritized molecules (Table 4) demonstrates that integration of machine learning-based pharmacokinetic prediction during the early stages of molecular design substantially improves candidate quality before costly experimental evaluation. High gastrointestinal absorption, acceptable CYP450 interaction profiles, reduced hepatotoxicity, minimal cardiotoxicity, and favourable oral bioavailability collectively support the translational potential of the prioritized lead compounds.

Molecular docking studies demonstrated that the highest-ranked CLM-generated compounds exhibited stronger predicted binding affinities than the reference PI3Kγ inhibitor (Table 5). Detailed interaction analysis revealed multiple hydrogen bonds, hydrophobic contacts, aromatic π–π interactions, and favourable positioning within the ATP-binding pocket of PI3Kγ, as illustrated in Figure 5. These docking observations were further supported by molecular dynamics simulations, which demonstrated stable RMSD trajectories, sustained hydrogen-bond occupancy, limited residue fluctuations, and preserved protein structural integrity throughout the simulation period. Collectively, these findings suggest that the proposed hybrid CLM successfully generated molecules capable of forming stable and energetically favourable protein–ligand complexes.

A distinctive strength of the present investigation is the incorporation of explainable artificial intelligence (XAI) into the molecular design workflow (Figure 1). Unlike conventional deep neural networks that frequently operate as "black-box" systems, explainable AI identified molecular fragments contributing positively or negatively to predicted PI3Kγ inhibitory activity. Such mechanistic interpretation enhances transparency, facilitates rational medicinal chemistry decision-making, and improves confidence in computational predictions, thereby supporting future regulatory acceptance of AI-assisted drug discovery platforms.

Collectively, the proposed workflow integrates optimized transformer training (Table 1), superior model performance (Table 2), efficient learning convergence (Figure 2), expanded chemical space exploration (Figure 3), perplexity-guided molecular prioritization (Figure 4), favourable physicochemical and ADMET characteristics (Tables 3 and 4), and strong protein–ligand interactions (Table 5 and Figure 5) into a comprehensive computational framework for intelligent drug discovery. Rather than replacing medicinal chemists, hybrid Chemical Language Models should be viewed as advanced decision-support systems capable of prioritizing the most promising lead candidates before chemical synthesis, thereby reducing development costs, shortening discovery timelines, and improving early-stage success rates.

The proposed computational platform possesses several notable strengths, including the integration of molecular generation, optimization, virtual screening, and explainable AI within a unified workflow (Figure 1); carefully optimized transformer hyperparameters that ensure stable model training (Table 1); superior molecular validity, scaffold diversity, and uniqueness (Table 2); efficient transfer learning and model convergence (Figure 2); successful exploration of novel chemical space (Figure 3); unbiased perplexity-guided molecular prioritization (Figure 4); favourable drug-likeness and ADMET characteristics (Tables 3 and 4); and robust docking interactions with PI3Kγ (Table 5, Figure 5). Together, these findings demonstrate the broad applicability of the proposed framework across multiple therapeutic targets beyond PI3Kγ.Despite these encouraging findings, several limitations should be acknowledged.

The present investigation is entirely computational and therefore requires experimental validation through chemical synthesis, biochemical assays, cellular studies, pharmacokinetic evaluation, and in vivo investigations before clinical translation. Furthermore, although transformer-based CLMs demonstrated excellent molecular generation performance under the optimized training conditions (Table 1), their predictive capability remains dependent upon the diversity, quality, and representativeness of the training datasets. Future studies incorporating multimodal molecular representations, graph neural networks, reinforcement learning, protein structural information, diffusion models, and active learning strategies may further improve predictive performance and biological relevance.

Future research should focus on integrating Chemical Language Models with reinforcement learning, graph neural networks, diffusion-based generative models, autonomous robotic synthesis platforms, and real-time biological feedback systems to establish self-improving drug discovery frameworks. Coupling CLMs with patient-specific genomic information, electronic health records, and precision medicine databases may further enable individualized therapeutic design and accelerate the development of next-generation precision pharmaceuticals.

VI. Conclusion

The present study proposes a comprehensive hybrid Chemical Language Model framework for intelligent de novo drug discovery by integrating transformer-based molecular generation, transfer learning, perplexity-guided molecular prioritization, physicochemical optimization, virtual screening, molecular docking, molecular dynamics simulation, ADMET prediction, and explainable artificial intelligence. The proposed computational strategy generated chemically

valid, structurally diverse, synthetically feasible, and pharmacologically promising PI3Kγ inhibitor candidates while demonstrating superior performance compared with conventional molecular generation approaches. Incorporation of perplexity-based ranking substantially improved candidate prioritization by reducing model bias and increasing confidence in generated molecular structures. The framework provides a scalable, reproducible, and computationally efficient platform for accelerating early-stage drug discovery, particularly for therapeutic targets characterized by limited experimental datasets. Future experimental validation will be essential to confirm the biological activity of the identified lead compounds and further establish the utility of hybrid Chemical Language Models within modern pharmaceutical research.

VII. Conflict of Interests

The authors declare that there are no financial, professional, institutional, or personal conflicts of interest that could have influenced the work reported in this manuscript.

VIII. Acknowledgements:

The authors acknowledge the contributions of the global scientific community for the development of publicly accessible computational chemistry resources, open-source cheminformatics software, molecular databases, and artificial intelligence frameworks that facilitated the development of the proposed computational workflow.

References

1. Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: The pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery, 9(3), 203–214. doi:10.1038/nrd3078.
2. Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. British Journal of Pharmacology, 162(6), 1239–1249. doi:10.1111/j.1476-5381.2010.01127.x.
3. Polishchuk, P. G., Madzhidov, T. I., & Varnek, A. (2013). Estimation of the size of drug-like chemical space based on GDB-17 data. Journal of Computer-Aided Molecular Design, 27(8), 675–679. doi:10.1007/s10822-013-9672-4.
4. Walters, W. P., & Murcko, M. A. (2020). Assessing the impact of generative AI on medicinal chemistry. Nature Biotechnology, 38(2), 143–145. doi:10.1038/s41587-020-0418-2.
5. Cherkasov, A., et al. (2014). QSAR modeling: Where have you been? Where are you going to? Journal of Medicinal Chemistry, 57(12), 4977–5010. doi:10.1021/jm4004285.
6. Lionta, E., Spyrou, G., Vassilatis, D. K., & Cournia, Z. (2014). Structure-based virtual screening for drug discovery: Principles, applications and recent advances. Current Topics in Medicinal Chemistry, 14(16), 1923–1938. doi:10.2174/1568026614666140929124445.
7. Vamathevan, J., et al. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463–477. doi:10.1038/s41573-019-0024-5.
8. Schneider, P., et al. (2020). Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery, 19(5), 353–364. doi:10.1038/s41573-019-0050-3.

9. Elton, D. C., Boukouvalas, Z., Fuge, M. D., & Chung, P. W. (2019). Deep learning for molecular design—A review of the state of the art. Molecular Systems Design & Engineering, 4(4), 828–849. doi:10.1039/C9ME00039A.
10. Mercado, R., et al. (2021). Practical notes on building molecular generative models. Applied AI Letters, 2(1), e18. doi:10.1002/ail2.18.
11. Weininger, D. (1988). SMILES, a chemical language and information system. I. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences, 28(1), 31–36. doi:10.1021/ci00057a005.
12. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
13. Moret, M., Friedrich, L., Grisoni, F., Merk, D., & Schneider, G. (2022). Generative chemical language models for de novo drug design. Nature Reviews Chemistry, 6(1), 3–18. doi:10.1038/s41570-021-00346-7
14. Moret, M., Grisoni, F., Katzberger, P., et al. (2021). Perplexity-based molecule ranking and bias estimation of chemical language models. Chemical Science, 12(34), 11142–11154. doi:10.1039/D1SC02031D