LLM Integration into in silico Phenotypic and Transcriptomic Frameworks for Novel Polyploid Giant Cancer Cell Inhibitor Discovery

Samar Josyula

Abstract

Polyploid giant cancer cells (PGCCs) represent a highly resilient subtype of cancerous cells associated with cancer relapse, metastasis and therapy resistance. The recent screening methods that have emerged over the past decade have made significant advances in the identification and quantification of PGCC inhibitors through high-throughput screenings and in silico predictions, yet they remain overall inefficient due to issues in scalability, feature diversity and interpretability. This review explores how large learning models (LLMs) can be integrated into current method frameworks to help mitigate some of these limitations. Building upon the pipeline architecture laid out by Zhou et al. (2023) and Ma et al. (2025), along with the application of LLM4SD framework guidelines provided by Zheng et al. (2025), a potential hybridized model conjoining LLM capabilities to existing empirical and in silico methodologies shows real potential to become the new gold-standard in anti-PGCC discovery. This review evaluates the advantages of this potential hybrid model via few-shot compound matching, enhanced feature representations and advanced mechanistic interpretability. We also assess potential limitations that could arise within the hybridized pipeline, such as data hallucinations or data biases. Overall, the methods mentioned in this review provide a forward-looking strategy for the application of LLMs in rare cancer cell research and provides a conceptual foundation for improved scalability, interpretability, and practicality of PGCC-targeted drug discovery pipelines.

1. Introduction

1.1 Polyploid Giant Cancer Cells

Polyploid giant cancer cells (PGCCs) are subtypes of cancerous cells with additional copies of chromosomes, often resulting in significantly larger cell sizes by way of multiple nuclei or a singular, abnormally large nucleus. These result in the cancerous cells bypassing routine mitotic checkpoints, leading to endoreduplication or cell fusion events that result in polyploidy and multiple nuclei or an enlarged nucleus respectively. The formation of these cancerous cells can be attributed to several mechanisms, including aberrant cell cycle regulation, mitotic failure and response to cellular stresses, such as radiotherapy and chemotherapy (Ma et al., 2025).

1.2 PGCC Significance and Role in Cancer Relapse and Therapy Resistance

PGCC adoption of a dormant phenotype during times of cellular stress such as cancer therapies provide them with a superior resistance against them (Zhou et al., 2023). During their period of dormancy, they often initiate endoreplication to expand genomic content and variability in parallel with dedifferentiation to initiate tumor regeneration. These adaptations enable PGCCs not only to survive applied cellular stress via cancer therapies, but also regenerate tumor populations after therapy concludes. As a result, PGCCs prove to be the most resistant to therapeutic attempts combatting the tumor expansion when compared to other subtypes of cancerous cells. However, due to a lack of high-throughput methods available to quantify them, effective anti-PGCC therapies do not exist (Niu et al., 2016).

2. Current Methods and Their Limitations

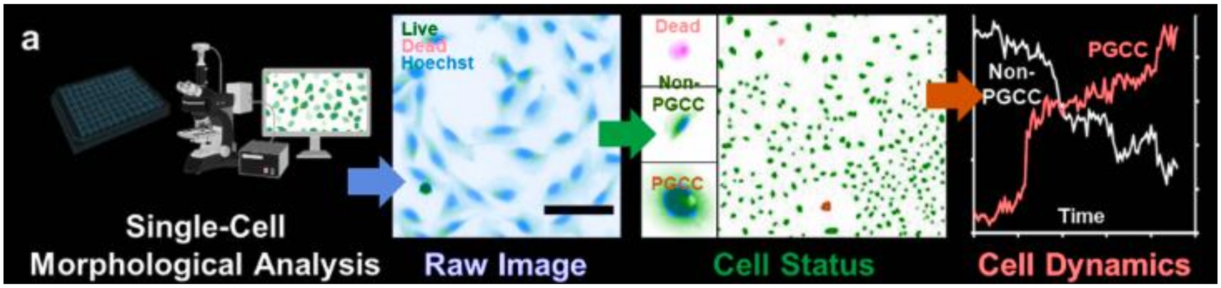
2.1 Single-Cell Morphological and Transcriptome Analysis Unveil Inhibitors of Polyploid Giant Breast Cancer Cells in Vitro (Zhou et al., 2023)

In a landmark effort to develop a high-throughput method aimed at identifying and quantifying PGCC and non-PGCC cells, Zhou et al. (2023) developed a morphological analysis pipeline, capable of identifying compounds that act inhibitory towards PGCCs, or a variety of other cancerous cells. The analysis pipeline was performed using a custom MATLAB program that involved cell nuclei identification through Hoechst staining, filtration of dead cells through Live/Dead staining and the distinction of PGCCs from non-PGCCs based on cell nuclei size after treatment. Time lapse experiments were also utilized to track cell population dynamics, using Brightfield and fluorescence images, captured with a Nikon Ti2E inverted microscope.

To support findings from the visual screenings and analyses, the study integrated single-cell transcriptome sequencing (figure a below) which characterized cell compositional discrepancies when treated with the library of 172 compounds. Among these tested compounds, 10 were found to significantly inhibit PGCCs but not non-PGCCs, and 13 were found to inhibit both.

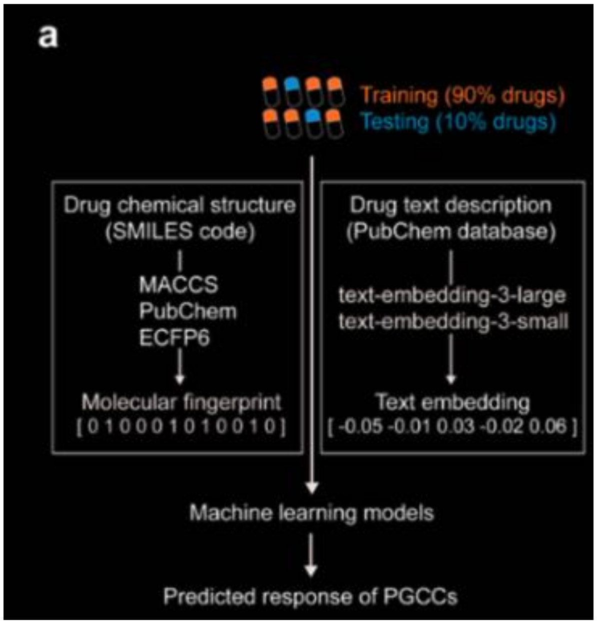
Based on visual screening and single-cell transcriptome analyses, the study found 3 classes of compounds that effectively killed off PGCCs. HDAC inhibitors were shown to inhibit Docetaxel-induced PGCCs and flow-sorted Vari068 PGCCs, when combined with other chemo and radiotherapies. Proteasome inhibitors and ferroptosis inducers not only acted as inhibitors of TNBCs, but also for HR-positive and HER2-positive breast cancer cells.

Despite the precision of this approach, it still relies heavily on the labor-intensive methods of image acquisition and segmentation. Repetition of this method may prove to be impractical and inefficient (Zhou et al., 2023).



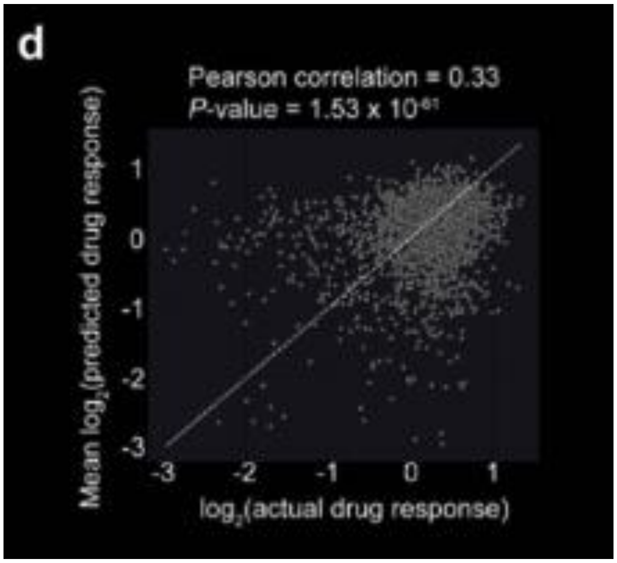
2.2 High-Throughput Empirical and Virtual Screening to Discover Novel Inhibitors of Polyploid Giant Cancer Cells in Breast Cancer (Ma et al., 2025)

To address the concerns of inefficiency and impracticality mentioned by the previous study, Ma et al. (2025) aimed to develop a machine learning model framework to predict anti-PGCCs in silico. A library consisting of 2726 phase-I approved compounds was selected and ran through the similar screening and imaging protocol of flow cytometry and visual inspection. After library filtration, the number of compounds decreased to 2430.



As for the model training, each compound in the library was represented as vectors of biochemical and structural features through MACCS, PubChem, and ECFP6 fingerprints, or vectors of text embeddings to encode pharmacological, biochemical and structural information. Regression models were trained and tested to predict discrepancies in PGCC counts based on quantitative representations of textual descriptors and fingerprints through 10 rounds of 10-fold cross validations, with a train/test ratio of 9:1. In the end, just under half (49.2%) of the models achieved a Pearson correlation coefficient above 0.2 across all 10 rounds.

Despite modest correlation scores across models, the integration of cheminformatics and text-based embeddings yielded meaningful predictive capacity. The best-performing models combined structural fingerprints (e.g., MACCS and PubChem) with semantic drug descriptions encoded using large text embeddings (figure a to the right). Among model architectures, histogram-based gradient boosting (HGB) and support vector machines (SVM) demonstrated higher Pearson correlations, particularly when used in an ensemble configuration. The top ensemble model achieved a Pearson correlation of 0.33 (Figure d below). These models surfaced previously untested compounds with anti-PGCC efficacy, such as Pyronaridine. This confirms that methods for PGCC inhibitor identification



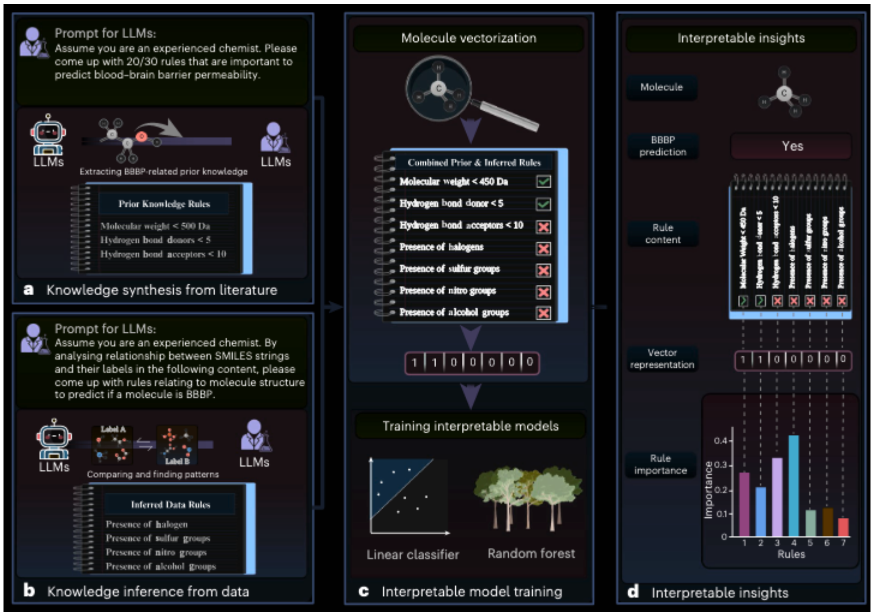
involving both empirical screening and in silico ML-based inference play a role in increasing the efficacy and practicality of anti-PGCC discovery and breast cancer research.

While the integration of these models improved the screening efficiency, the study also revealed crucial limitations in the process. A model's performance remained completely dependent on the variance of the training data set, which was restricted only to compounds with known activity. This reiterates the same problem found in the previous method where high-throughput methods to quantify PGCC inhibitors are not currently viable without a plethora of PGCC-targeted compound information, which is scarce due to a lack of high-throughput methods to quantify them. Additionally, interpretation of model results was found to be a challenge as these models cannot provide mechanistic explanations of their findings. The scarcity of biochemical mechanistic information increases the chances of drawing inaccurate conclusions based on model outputs, which are already producing suboptimal results themselves (Ma et al., 2025).

3. LLM implementation in Morphological Prediction

3.1 Large Language Models and their Potential Application in Molecular Property Prediction (Zheng et al., 2025)

Zheng et al. (2025) introduced a novel framework, LLM4SD (Large Language Models for Scientific Discovery). The framework functions by performing knowledge synthesis and inferences from existing scientific literature. First, by retrieving molecular features from established data libraries. Second, by using it's understanding of morphological conformations (through the SMILES notation) and chemistry knowledge to identify patterns from experimental data. These connections create interpretable vectors of features. By training previously mentioned models with these feature vectors, we can draw molecular property predictions with a far greater confidence than previously possible and with concrete mechanistic conclusions provided by the LLM.


Prompt for LLMs:
Assume you are an experienced chemist. Please come up with 20/30 rules that are important to predict blood-brain barrier permeability.
LLMs
Extracting BBBP-related prior knowledge
LLMs
Prior Knowledge Rules
Molecular weight < 500 Da
Hydrogen bond donors < 5
Hydrogen bond acceptors < 10
a Knowledge synthesis from literature
Prompt for LLMs:
Assume you are an experienced chemist. By analysing relationship between SMILES strings and their labels in the following content, please come up with rules relating to molecule structure to predict if a molecule is BBBP.
LLMs
Label A
Label B
Comparing and finding patterns
LLMs
Inferred Data Rules
Presence of halogen
Presence of sulfur groups
Presence of nitro groups
Presence of alcohol groups
b Knowledge inference from data
Molecule vectorization
C
H
Combined Prior & Inferred Rules
Molecular weight < 450 Da ✓
Hydrogen bond donor < 5 ✓
Hydrogen bond acceptors < 10 ☒
Presence of halogens ☒
Presence of sulfur groups ☒
Presence of nitro groups ☒
Presence of alcohol groups ☒
1 1 0 0 0 0 0
Training interpretable models
Linear classifier
Random forest
c Interpretable model training
Interpretable insights
Molecule
BBBP prediction
Yes
Rule content
Molecular Weight < 450 Da
Hydrogen bond donor < 5
Hydrogen bond acceptors < 10
Presence of halogens
Presence of sulfur groups
Presence of nitro groups
Presence of alcohol groups
Vector representation
1 1 0 0 0 0 0
Rule importance
Importance
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
Rules
d Interpretable insights

The LLM4SD pipeline consists of knowledge synthesis from existing literature, knowledge inferences from the respective data, model training and interpretation of results. Although LLM4SD does not aim to replace the standard ML methods used today, it does augment them to act more interactively with data inputs by providing embeddings that feed into downstream models. The ability of LLM4SD to bridge gaps between structural, textual and mechanistic knowledge offer much insight when compared to the explicit black-box predictions that have been curated before. LLMs may also offer additional explanations on why certain compounds act as PGCC inhibitors, by leveraging its vast knowledge of scientific literature towards the scope (Zheng et al., 2025).

3.2 Requirements for LLM Integration in PGCC Inhibitor Prediction

Building on the empirical and ML-based pipelines developed by Zhou et al. (2023) and Ma et al. (2025), we can further improve a hybrid framework that incorporates large language models, in attempts to enhance the practicality and efficacy of high-throughput anti-PGCC molecular discovery. This can all be done with enhanced confidence in results brought by various mechanistic conclusions of any correlational data found.

The first requirement is a foundation of empirical data. Phenotypic and molecular features extracted from high-throughput cell imaging from Zhou et al. (2023) and compound-response profiles via Ma et al. (2025). These can be used later to distinguish PGCCs from other cancerous cell states and to assign response scores to candidate compounds based on observed impact on PGCC formation/elimination in future testing datasets.

To fully leverage the interpretive strengths of LLM4SD, single-cell transcriptomic data and the morphological imaging can be transformed into structural or semi-structured text for embedding. Snapshots previously used only as visual verification, will now serve the additional role of semantic features through cluster annotation based on compound density and fed into the same modal embedding space.

By encoding these signatures into a natural language format, they can be embedded using LLMs and fused with chemical and pharmacological data. Multi-modal reasoning, or the ability of the model to reason across various modalities, will further streamline the pathway while results develop greater interpretability and transparency.

4. Practicality of LLM4SD Integration into Previous PGCC Inhibitor Identification Methods

4.1 Supplementing Drug Feature Representations

Where Ma et al. (2025) implemented fixed morphological fingerprints and descriptive text embeddings from their own empirical conclusions, LLMs may be able to generate richer, contextual embedding of drug mechanisms, target pathways and the functionalities of both. It was found that 91.3% of inferred rules were statistically significant, higher than the calculation of synthesized rules, and an average of 74% of these were already documented in existing scientific literature

(Zheng et al., 2025). The incorporation of an LLM's knowledge database on these PGCC inhibitors can find connections between implicit biochemical and pharmacological pathways that they share to provide further biologically informed and relevant features for prediction modelling.

For example, the pathways of HDAC/Proteasome inhibitors and Ferroptosis inducers found to be anti-PGCCs in Zhou et al. (2023) could be used as seed inputs to draw various initial mechanistic descriptors using an LLM. These current descriptors from input literature or LLM knowledge databases, would then enrich the feature space of the initial (N+1) train/test dataset. As the compounds are screened and validated, the LLM links molecular activity to additional, mechanistically similar compounds found within its knowledge databases. This would generate an even broader, more contextually informed set of features for the training of the next (N+2) iteration, thus exponentially improving compound prioritization.

4.2 Enhanced Interpretability and Mechanistic Insights

A key limitation in the study performed by Ma et al. (2025) was the potential corruption of conclusions drawn due to the necessity of human interpretation of the statistical results. LLMs trained on the LLM4SD pipeline can offer interpretable outputs, such as a compounds relation to an established PGCC or hypothesis on a drug's mode of action. By employing interpretable models, the LLM enables quantification of each rule's importance in prediction, which elucidates the contribution to the model's final decision (Zheng et al., 2025). This transparency allows for intuitive comprehension of the LLM's 'thought process' and allows for seamless interpretation of hypotheses.

4.3 More Accurate Results from Few-Shot or Zero-Shot Predictions

The lack of PGCC molecular empirical data or genetic information acted as a crucial handicap in previous methods. Regardless of the lack of high throughput application in the study conducted by Ma et al. (2025), A deficit of viable compounds prior to filtration may have led to a model that was suboptimal. Conversely, LLMs can infer 'few-shot' relationships between novel compounds and PGCC phenotypes based on semantic similarity, due to the natural language ability. The previously mentioned exponential improvement of compound prioritization shown in feature representations also contributes to these accurate predicational outcomes. Few-shot or zero-shot learning capabilities, allow LLMs to generalize task knowledge from very few examples. In the context of PGCC inhibitor identification, a large language model would be able to fill the gaps of limited data sources by synthesizing them through its hypothesis generation.

5. Potential Limitations and Risks

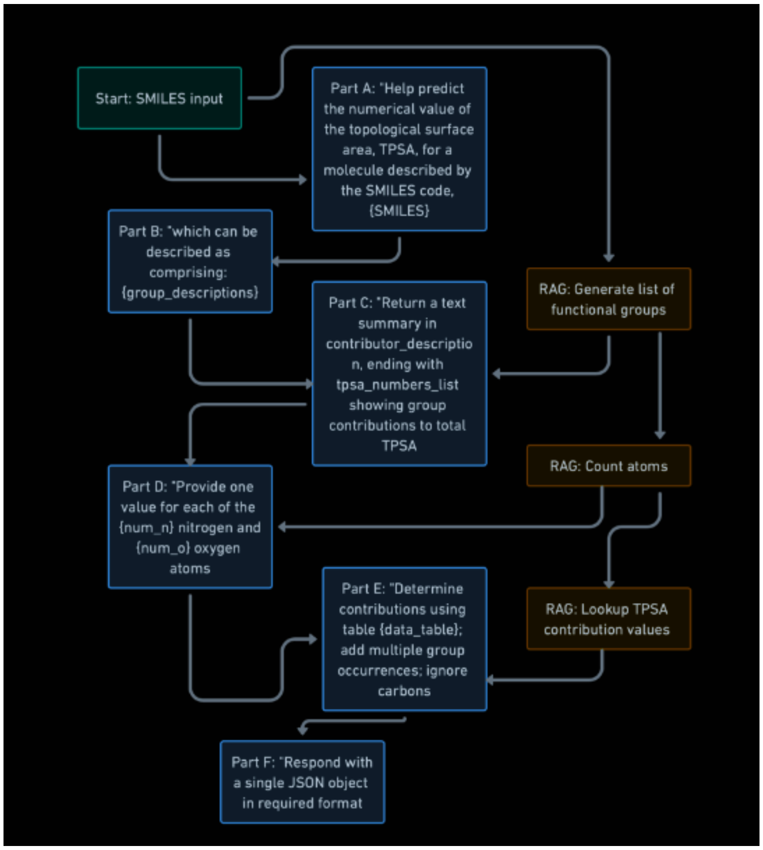
5.1 Data Hallucination

LLM Implementation within the ongoing search for more efficient methods of anti-PGCC identification does not come without risks. Hallucinations can happen, where the model produces outputs that are vastly different from the expected output. This is often due to gaps or inconsistencies in the training data sets. Even minor discrepancies can lead to significant misinterpretations by the model. There are some approaches that could decrease the risk of these chemical hallucinations.

Retrieval-Augmented-Generation (RAG) combines the data retrieval system within the generative model. With RAG implemented, our LLM is able to dynamically fetch and calculate relevant information from external databases, outside of the scope of the input. RAG would be a particularly crucial implementation for our specific framework, as it is shown to be valuable for molecular property prediction using group contribution methods.

Multiprompt Instruction Proposal Optimizer (MIPRO) is a prompt optimization framework that synthesizes and refines LLM prompts for enhanced precision and reliability. This is done through the addition of supplementary instructions to add to the prompt. Using a PyTorch powered machine learning framework, MIPRO is then able to select few-shot examples that illustrate successful executions of the given task.

When implemented together, RAG can address issues that arise with outdated or incomplete by grounding the LLM’s response in current data sources. MIPRO is then able to optimize prompt structure, allowing our LLM to interpret and utilize the retrieved data more effectively through better specified instructions (Reed, 2025).


Start: SMILES input
Part A: "Help predict the numerical value of the topological surface area, TPSA, for a molecule described by the SMILES code, {SMILES}
Part B: "which can be described as comprising: {group_descriptions}
RAG: Generate list of functional groups
Part C: "Return a text summary in contributor_description, ending with tpsa_numbers_list showing group contributions to total TPSA
RAG: Count atoms
Part D: "Provide one value for each of the {num_n} nitrogen and {num_o} oxygen atoms
RAG: Lookup TPSA contribution values
Part E: "Determine contributions using table {data_table}; add multiple group occurrences; ignore carbons
Part F: "Respond with a single JSON object in required format

5.2 Data Bias Issues

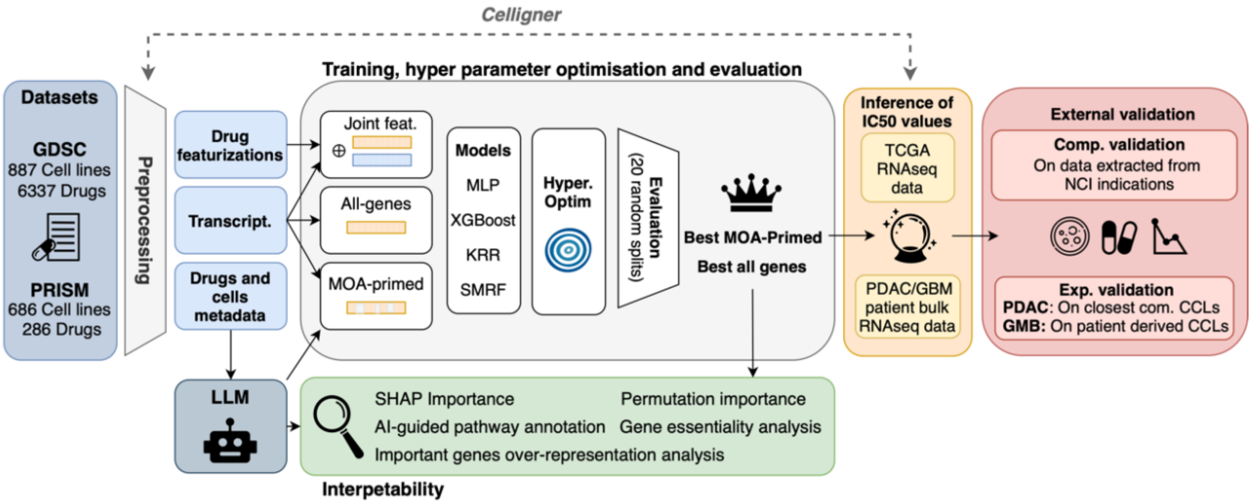
Large language trained on biased datasets may be inclined to exacerbate existing biases or hallucinate data, as previously mentioned. It is crucial that data inputs fed into an LLM-integrated pipeline are free of any data corruption. This issue can be further eliminated through diversification of the training data sets to ensure fair and unbiased outputs (Sarumi & Heider, 2024).

6. Related Literature for LLM-Based Molecular Property Prediction

6.1 Learning and actioning general principles of cancer cell drug sensitivity (Carli et al., 2025)

Carli et al. (2025) implements an LLM pipeline to explore drug sensitivity across various distinct cancerous phenotypes. The study shared a notable similarity with PGCC screening, as it focuses on how genetic heterogeneity within cancerous populations can influence selective drug responses. Empirical drug response data and leveraged transcriptomic and phenotypic clustering was used to train their model. The authors used LLM-based embeddings to infer drug-subtype relationships and validated their predictions across publicly available pharmacogenomic datasets.

Their results showcased the practicality of LLM implementation to support compound prioritization, specifically based on phenotypic viability. This corelates closely with our directive, with the need to target PGCCs as a rare and therapy resistant cell state within heterogeneous tumors.



6.2 Cancer Gene Identification through Integrating Casual Prompting Large Language Model with OMICS Data-Driven Casual Inference (Zeng et al., 2025)

Zeng et al. (2025) introduces ICGI, a novel framework for identifying cancerous genes across multi-omics domains. Multi-omics refers to the integration of multiple “omics” data types (ex. transcriptomics, genomics, etc.). ICGI’s integration of transcriptomic, CNV and methylation data, with LLM-based reasoning over biomedical literature, enables the discovery of gene-phenotype relationships based in biochemical mechanisms. Unlike other correlational models, ICGI focuses on mechanistic interpretability, providing natural language justifications and model transparency for rationale behind its findings (Zeng et al., 2025).

6.3 Procedures from Similar Literature to Assimilate into LLM-Based Anti-PGCC Identification Methods

In Carli et al. (2025) the authors implemented a phenotype-aware LLM pipeline that mapped transcriptomic and molecular cancerous subtypes to compounds with the predicted selective efficacy. These procedures provide an abstract architectural skeleton for our pipeline implementation as we work on a hybridized LLM-based framework for anti-PGCC detection. Moreso, implementation of the interface provided by Zeng et al. (2025), or similar, could allow us to enhance our framework to optimize the retrieval of mechanistic explanations of correlational results. By incorporating their prompting model into our framework, we could be able to combine structured phenotypic data from transcriptomic and morphological features, with prompt-based rationale over mechanistic knowledge. Though it does not align completely to the LLM4SD framework, architectural inspiration taken from ICGI would allow us to prioritize more compounds that act as PGCC inhibitors through further explainable biochemical and pharmacological pathways by use of multi-omics.

7. Conclusion

The recent evolution of anti-PGCC drug discovery, from high-throughput empirical screening, to in silico predicational frameworks, has revealed both the future potential and current limitations of the methodologies. While Ma et al. (2025) improved upon the framework of Zhou et al. (2023) by implementing machine learning screening, both approaches still face the constraints of scalability and interpretability. The integration of large learning models, particularly ones through frameworks like LLM4SD or similar, offer an exciting new avenue to explore in the realm of computational drug discovery. These models could enable lusher feature representation, a more rational hypothesis generation technique, and seamless cross-modal examination.

Despite the promise of this approach, challenges surrounding unfamiliar implementation and bug-handling remain prevalent. LLM's are prone to data hallucination and bias if given suboptimal input or if grounded in training sets of low quality. Previous methods would simply show low correlational scores as result of the same inputs. Integration of LLMs in anti-PGCC detection would require a high level of attentiveness when analyzing results and careful surveillance of the mechanistic rationale provided, if any. Techniques such as RAG and MIPRO, along with interfaces like ICGI, can help mitigate these risks but one must remain cautious.

By combining the precision of previous empirical pipelines with the semantic depth of LLMs provided through natural language detection, we may be capable of accelerating PGCC inhibitor discovery while retrieving the bonus of biochemical and pharmacogenic mechanistic transparency. This potential development of the anti-PGCC discovery pipeline may not only act as an advancement within the specific scope but serve as an exemplar shift in the way rare cancerous cells like PGCCs may be targeted using evolving computational techniques.

References

Carli, F., Di Chiaro, P., Morelli, M., Arora, C., Bisceglia, L., De Oliveira Rosa, N., Cortesi, A., Franceschi, S., Lessi, F., Di Stefano, A. L., Santonocito, O. S., Pasqualetti, F., Aretini, P., Miglionico, P., Diaferia, G. R., Giannotti, F., Liò, P., Duran-Frigola, M., Mazzanti, C. M., … Raimondi, F. (2025). Learning and actioning general principles of cancer cell drug sensitivity. *Nature Communications*, *16*(1). https://doi.org/10.1038/s41467-025-56827-5

Ma, Y., Shih, C.-H., Cheng, J., Chen, H.-C., Wang, L.-J., Tan, Y., Zhang, Y., Brown, D. D., Oesterreich, S., Lee, A. V., Chiu, Y.-C., & Chen, Y.-C. (2025). High-throughput empirical and virtual screening to discover novel inhibitors of polyploid giant cancer cells in breast cancer. *Analytical Chemistry*, *97*(10), 5498–5506. https://doi.org/10.1021/acs.analchem.4c05138

Niu, N., Zhang, J., Zhang, N., Mercado-Uribe, I., Tao, F., Han, Z., Pathak, S., Multani, A. S., Kuang, J., Yao, J., Bast, R. C., Sood, A. K., Hung, M.-C., & Liu, J. (2016). Linking genomic reorganization to tumor initiation via the giant cell cycle. *Oncogenesis*, *5*(12). https://doi.org/10.1038/oncsis.2016.75

Reed, S. (2025). *Augmented and Programmatically Optimized LLM Prompts Reduce Chemical Hallucinations*. https://doi.org/10.26434/chemrxiv-2025-rwgt8

Sarumi, O. A., & Heider, D. (2024). Large language models and their applications in bioinformatics. *Computational and Structural Biotechnology Journal*, *23*, 3498–3505. https://doi.org/10.1016/j.csbj.2024.09.031

Zeng, H., Yin, C., Chai, C., Wang, Y., Dai, Q., & Sun, H. (2025). Cancer gene identification through integrating causal prompting large language model with OMICS data–driven causal inference. *Briefings in Bioinformatics*, *26*(2). https://doi.org/10.1093/bib/bbaf113

Zheng, Y., Koh, H. Y., Ju, J., Nguyen, A. T., May, L. T., Webb, G. I., & Pan, S. (2025). Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, *7*(3), 437–447. https://doi.org/10.1038/s42256-025-00994-z

Zhou, M., Ma, Y., Chiang, C.-C., Rock, E. C., Butler, S. C., Anne, R., Yatsenko, S., Gong, Y., & Chen, Y.-C. (2023). Single-cell morphological and transcriptome analysis unveil inhibitors of polyploid giant breast cancer cells in vitro. *Communications Biology*, *6*(1). https://doi.org/10.1038/s42003-023-05674-5