SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 15662-15675
Open Access. Online & Print

DOI: 10.5281/zenodo.20614025

AN AGENTIC IN SILICO FRAMEWORK FOR DRUG TARGET DISCOVERY AND TOXICOLOGY-AWARE PRIORITIZATION IN ACUTE LYMPHOBLASTIC LEUKEMIA

Sultan Qalit Alhumran[1], Graham Roy Ball[2*], Ahmed A. El-Sherif[3], Nahla O. Mosa[4], Shaza Ahmed[5], Shahad Ali Alghorayed[6], Nader Atallah Alatawi[7], Afrah Surayh Algethami[8], Albalawi Mohammed Ali[9], Fahad Abdullah Alqahtani[10], Refaat M. Gabre[11*]

[1]Tabuk Poison Control and Forensic Medicinal Chemistry Center, Ministry of Health, Tabuk 47915, Saudi Arabia, Department of Biotechnology, Faculty of Science, Cairo University, Giza 12613, Egypt salhumrani@moh.gov.sa
[2]Intelligent Omics Ltd, Old Flint Barn, Barnham, Cambridge, UK IP24 2PL, UK
Medical Technology Research Centre, Anglia Ruskin University, Chelmsford Campus, Bishop Hall Lane, Chelmsford CM1 1SQ, UK. graham.ball@ntu.ac.uk
[3]Department of Chemistry, Faculty of Science, Cairo University, Giza 12613, Egypt aelsherif@sci.cu.edu.eg
[4]Department of Biotechnology, Faculty of Science, Cairo University, Giza 12613, Egypt nahlaosama@aucegypt.edu
[5]Lecturer, Faculty of Biotechnology, October University for Modern Sciences and Arts, Giza 12451, Egypt Shabib@msa.edu.eg
[6]Department of Biotechnology, Faculty of Science, Cairo University, Giza 12613, Egypt saalghorayed@fakeeh.care
[7]Department of Biotechnology, Faculty of Science, Cairo University, Giza 12613, Egypt naatalatawi@moh.gov.sa
[8]Department of Biotechnology, Faculty of Science, Cairo University, Giza 12613, Egypt dr_afrah1993@hotmail.com
[9]Consultant, Department of Oncology, King Fahad Specialist Hospital, Ministry of Health, Tabuk 31444, Saudi Arabia. malbalawi112@moh.gov.sa
[10]Consultant, Department of Oncology, King Fahad Specialist Hospital, Ministry of Health, Tabuk 31444, Saudi Arabia. falqahtani40@moh.gov.sa
[11]Department of Biotechnology, Faculty of Science, Cairo University, Giza 12613, Egypt rgabre@sci.cu.edu.eg

Received: 04/04/2026
Accepted: 20/05/2026

Corresponding Author: Graham R. Ball & Refaat M. Gabre (graham.ball@ntu.ac.uk rgabre@sci.cu.edu.eg

ABSTRACT

Acute lymphoblastic leukemia (ALL) is still a difficult-to-treat disease, especially in the relapsed/refractory

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License. (https://cre-ativecommons.org/licenses/by/4.0/).

(r/r) setting. New computational methods are required to identify druggable targets and to take toxicity into account at the start of the drug discovery process. We created a combined in silico pipeline: (1) feature discovery using artificial neural networks (ANNs) trained with a swarm of networks (swarm-based ANNs, or SBANNs) and (2) druggability and toxicological risk scoring with LLM. The pipeline was used on both ALL and normal samples of bone marrow derived from the TCGA. A stable set of 412 genes associated with ALL was identified using SBANN analysis. The targets were mostly selected based on biological relevance, drugability, and toxicity, guided by the LLM-assisted scoring (Claude.ai with specific prompts). A number of the top contenders (VDR, TYK2, PTGS2) have a known biological role and existing pharmacology. This proof-of-concept framework shows the feasibility of applying SBANN feature ranking and structured LLM evaluation for systematic target prioritisation in translational bioinformatics while considering toxicity. The code and data will be shared to enable other datasets to be used and replicated.

KEYWORDS: Drug Target Prioritization, Acute Lymphoblastic Leukemia, Swarm Neural Networks; LLM-Assisted Analysis, Translational Bioinformatics.

1. INTRODUCTION

Acute lymphoblastic leukemia (ALL) is a rapidly developing hematological malignancy, involving unchecked growth of immature lymphoid precursors, which can cause an interruption to normal hematopoiesis [1]. Although the knowledge of ALL etiology has contributed to great advances in therapy and improved the prognosis for children, the prognosis for adults remains significantly poorer, and relapse remains a major clinical challenge [2]. The etiology of ALL is complex and features a variety of genetic, epigenetic, transcriptional, and microenvironmental changes that converge to drive leukemic cell states with dysregulated proliferation, impaired apoptosis, aberrant cytokine signalling, and altered metabolic and immune regulatory programs [3-5].

Molecular characterization studies conducted in large populations, including The Cancer Genome Atlas (TCGA) and other transcriptomic studies, have revealed that ALL is a disease comprised of several molecular disease subsets, each of which contains unique signalling abnormalities, dependency pathways, and lineage-identity programs [6, 7]. So, ALL should no longer be regarded as a monolithic disease, but rather a heterogeneous group of different biologic and clinical entities, whose treatment and risk of relapse are determined by the respective molecular patterns and drivers of ALL subtypes [2].

Therapeutic discovery is traditionally driven by hypotheses, which are difficult to scale up for systematic screening of interconnected molecular networks regulating leukemic behaviour. As high-dimensional transcriptomic datasets become increasingly available, machine learning techniques on the basis of statistically sound methods are increasingly being employed to look for robust, biologically meaningful targets that may not be detected by traditional methods.

In complex hematological malignancies, multi-omics data have been used to identify genes predictive of disease state using an artificial neural network approach, which has been proven to be a strong strategy [8]. Swarm-based artificial neural networks (SBANNs) also increase the explainability, decrease the number of false discoveries, and increase the stability of features. We introduce an integrated computational workflow in this study that integrates a swarm-based neural network (NN) for feature discovery with an LLM-based druggability and toxicological risk assessment.

The main objective of this work is to build and validate a reproducible in silico prioritisation pipeline, which will be able to identify stable genes associated with ALL from public RNA-seq data, and systematically assess their therapeutic potential and safety profile, leveraging structured LLM reasoning. This workflow combines the gene ranking, pathway enrichment analysis, and LLM-based scoring capabilities of SBANN for target prioritisation in ALL with toxicity filtering.

2. MATERIALS AND METHODS

2.1. Data Source and Sample Cohort Selection

This study used a multi-stage computational pipeline that included transcriptomic profiling, the use of a swarm-based neural network, mechanistic enrichment analysis, agentic large language model (LLM) reasoning, and toxicity prediction to identify and prioritize druggable gene targets for acute lymphoblastic leukemia (ALL). In this study, the cohorts were obtained from The Cancer Genome Atlas (TCGA), which included bone marrow samples for samples with normal histology as well as ALL samples. The "NOS" cases with accurate disease diagnosis were selected in ALL cases, and the cases listed "with mutations" were excluded. TCGA data were downloaded from the data repository, and STAR count files were opened to extract the TPM count values[7, 9, 10]. The min-max normalization technique was used to normalize the expression value of each dataset to a range of 0 to 1, to facilitate comparability and uniformity between datasets. The TPM values were not subjected to any filtering in the selection, and all expression profile of 60,660 features was analyzed, including non-coding parts of the transcriptome.

2.2. Quality Control

Batch effects in the data were assessed by examining the data structure using t-distributed Stochastic Neighbor Embedding (t-SNE). The gene expression profiles of 12 commonly used housekeeping genes were used, and no batch effect was noted between the arms of the study.

2.3. Swarm-Based Neural Network (SBANN).

The ANN method employed in this study was originally described by Lancashire et al. (2009) and has been proven to be effective in the sense of finding patterns in the data by finding the optimal inputs for categorizing a certain task by choosing the one variable that is most predictive of the task. Here, molecular features that can predict ALL from healthy controls were identified based on a swarm-based artificial neural network (SBANN) approach that had been previously reported by Ajonu et al. (2025.

The SBANN approach uses several small ANNs to see the data from various angles, as shown in Figure 1. This method has advantages over the deep neural network model approach, such as better explainability, higher efficiency in training, and reduced false discovery rate. This has been implemented on top of earlier versions of ANN methods used to interrogate big data from omics [9, 11]. The latest is based on a large ensemble of neural networks, each trained on a different subset of the data, a subset obtained by randomly sampling the data for a set number of Monte Carlo runs, for a given gene product (see Figure 1).

All algorithms use a three-layer structure (comprising input, a limited number of hidden, and output nodes) and a TanH transfer function along with Levenberg-Marquardt weight update. The input to each model is one gene, and the output is the predicted ALL disease state. To ensure reproducibility and prevent overfitting, the following internal validation procedures were applied. For each of the 20 Monte Carlo runs, the data were randomly partitioned into training, validation, and test sets following the approach described in Lancashire et al. (2009). Early stopping was applied at 100 epochs based on performance on the validation set, and weight regularization was used to avoid divergence. Gene ranking was determined solely by RMS error on the held-out test set. Stability of feature selection across runs was quantified by requiring a gene to appear in the top-ranked set in all 20 runs for inclusion in the final 412-gene concordant set. The conditional probability of observing such convergence by chance was calculated as described in Results (Section 3.1), and the observed convergence (20/20 runs for 412 genes) serves as empirical evidence for reproducibility.

The SBANN algorithm then generates a set of errors across the swarm for every random subset of the data using a combined filter-and-wrapper approach and generates a distribution of errors for the single genes in the training, stopping, and test data. Sorting a list of genes by the RMS error on an unseen test data set can give a rank order of genes as well as a selection of a subset of genes that is more enriched. Twenty gene error distributions are generated by repeating the process 20 times, across different randomizations, and interrogated to find a stable, enriched set. For this study, the top 500 genes based on error distribution (Figure 2) were chosen for Concordance Assessment. In AML, this has recently been done to model the TP53 pathway, obtaining significant pathway features that are used to push the pathway in disease [12].

Conventional parametric methods were also used to determine p-values for enriched features. This was used as a cross-check of the predictive enrichment approach.

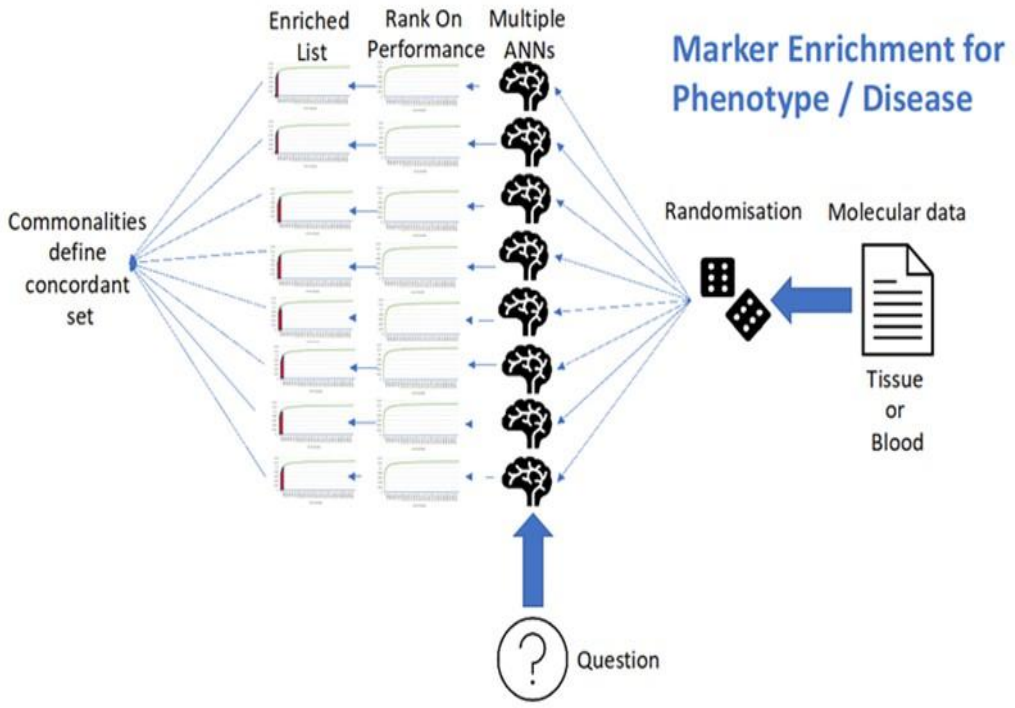

Figure 1. Process flow describing the basis of the SBANN development and how error distribution curves are used to identify Concordant enriched genes.

Figure 2. Example ranked error distributions for ALL versus Normal.

2.4. Ontological Assessment of Enriched Features

A second ontological assessment of enriched features is given. 412 enriched features related to ALL were presented to Panther Ontologies [13], where molecular function was determined, and to Reactome Pathway Analysis [14], where pathway membership was determined. Pathway enrichment was considered significant if there were ≥ 50 features. The presence of individual genes in the 412 gene list was identified for each pathway. Additional analyses were conducted using Pathway membership and Molecular Function data.

Enriched features

2.5. Druggability Assessment

The all-gene enrichment list was extended with ontological and pathway data. The package containing these combined data was then passed to DGIDB, ChEMBL, and DrugTargets for a druggability assessment. The following criteria were used to assess the potential of a feature from the gene list to be druggable, as shown in Table 1.

All these features were brought together in a feature druggability score, which scores each criterion between 0 and 2 depending on the strength of evidence and adds them together with equal weighting, yielding a score between 0 and 10. This process was expedited by using Claude.ai to assemble a druggability scoring table, which incorporated these parameters into a prompt, coupled with ontological details for the 412 genes, to provide an overall table of druggability for the key features.

The prompt used was:

"Please produce a table scoring the druggability of the list of genes that I am uploading. Please provide the output as a table with a full description of the details. The table should include gene names, ontology, pathway, druggability score, and rationale for selection. Please draw evidence from DGIDB, Drug Targets, Chembl, Reactome pathway, and molecular function to draw your conclusions."

Scoring followed a 0–2 scale per criterion (0 = no evidence, 1 = some evidence, 2 = strong evidence), summed equally to a total druggability score (0–10).

Table 1. Druggability assessment of Gene Features.

Criteria	Definition	Assessment
Historically drugged	The gene/protein is already directly targeted by an approved or clinical-stage drug/biologic.	Assessment was undertaken in DGIDB and Drug Targets to assess this for a given gene.
Chemically tractable	There is evidence that small molecules can bind the protein with reasonable potency and selectivity (e.g., known ligands, tractable pocket, similarity to known druggable families).	Assessment was undertaken in Chembl to evaluate modelled potency (IC50, AUC, Emax) against the gene for chemical entities and determine whether small molecules can act as inhibitors of the target.
Biotherapeutic tractable	The protein is accessible to antibodies, peptides, or other large biologics (e.g., cell-surface or secreted).	Ontologies were assessed to determine the localisation of expression and to determine if the gene had these features.
Network/biological feasibility	The modulation of the target is likely to produce a desired effect (e.g., disease association, pathway position) without catastrophic toxicity.	This was assessed at the primary level in this study by identifying the ALL-specific stable targets using the SBANN algorithm. This provides a robust hypothesis-free evidence-based approach to assess potential targets from across the genome.

2.6. Toxicological Assessment of Enriched Features

Toxicological risk was derived based on several ontological characteristics. The genes were assigned to pathways of core cellular processes, organ-specific pathways, and known toxicological pathways. A toxicity score, ranging from 0 to 10, was used to score targets. This process was expedited by using Claude.ai to assemble a toxicological risk score table, which incorporated these parameters into a prompt, coupled with ontological details for the 412 genes, to provide an overall table of toxicological risk for the key features.

The prompt used was:

"Please produce a table scoring the toxicological risk for the list of genes that I am uploading. Please provide the output as a table with a full description of the details. The table should include gene names, ontology, pathway, toxicological risk score, and rationale for selection. Please draw evidence from ProTox 3, known toxicological pathways, essentiality, Reactome pathways, and molecular function to draw your conclusions."

Toxicological risk scoring followed a 3-tier rubric: 3 = life-threatening toxicity with limited mitigation (e.g., essential cardiac conduction protein without redundancy); 2 = manageable or monitorable toxicity (e.g., reversible myelosuppression with available supportive care); 1 = restricted expression in disease tissue/compartment or redundant biology. Scores were scaled 0–10.

2.7. Small Molecule Identification

The optimised gene list was then input into DGIDB and ChEMBL to retrieve a group of small molecules having good drug-like properties. These databases were used to search and cross-reference properties including Molecular Weight (MW), Solubility, LogP (Partition Coefficient), H-Bond Donors, H-Bond Acceptors, TPSA (Topological Polar Surface Area), Rotatable Bonds, Aromatic Rings, Lipinski Violations, Bioavailability Score, Synthetic Accessibility, and Drug-likeness Score (QED). This was achieved using the following prompt within Claude.ai:

"Please identify small molecules with the potential to target these proteins. Please use DGIDB and ChEMBL as your references and search using the following selection criteria. Weight (MW), Solubility, LogP (Partition Coefficient), H-Bond Donors, H-Bond Acceptors, TPSA (Topological Polar Surface Area), Rotatable Bonds, Aromatic Rings, Lipinski Violations, Bioavailability Score, Synthetic Accessibility, and Drug-likeness Score (QED)"

The listed criteria followed standard drug-likeness thresholds from Lipinski's rule of five and Veber guidelines.

The overall process flow of the study and the database elements are shown in Figure 3.

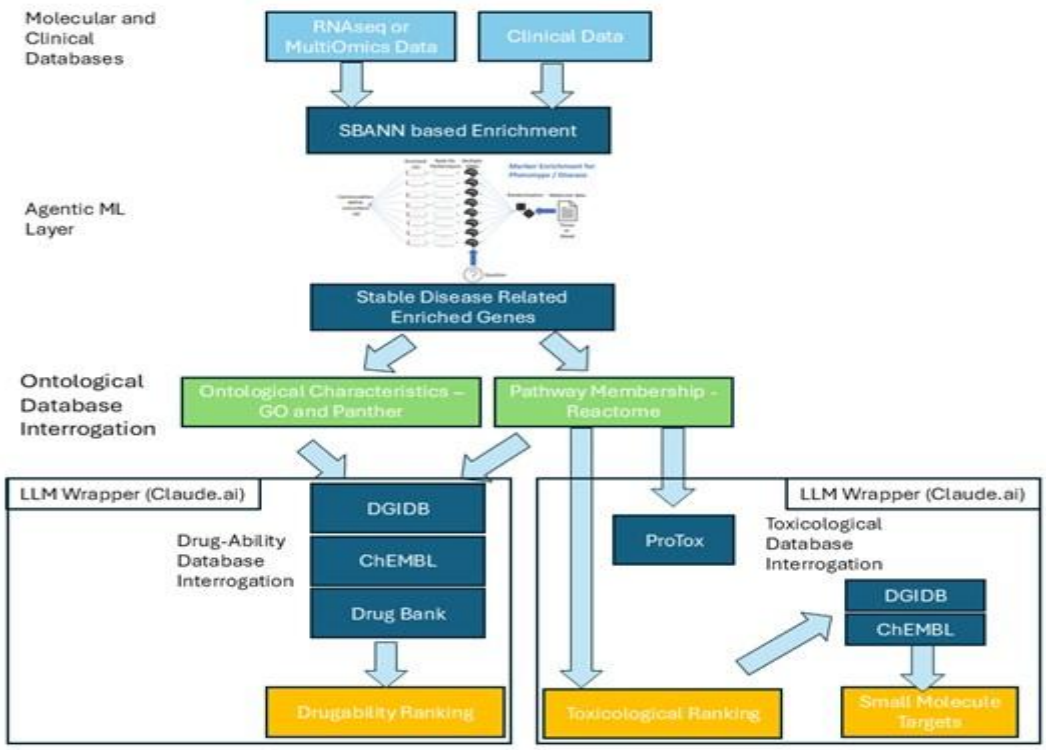

Figure 3. Agentic workflow used in the gene list assessment.

2.8. Code and Data Availability

All code, processed data, full gene lists, prompts, and detailed outputs have been submitted as supplementary files to the journal repository to ensure the reproducibility and transparency of the workflow. In addition, readers who require the complete datasets may directly contact the corresponding author via email, and the datasets will be shared upon reasonable request.

3. RESULTS

3.1. Gene Enrichment

The best 500 gene products from the SBANN models developed were observed in the following performance ranges:

AUC on test data = 0.814 to 0.972, RMS Error on test data = 0.011 to 0.061, with p-values ranging from 0.003 to 4.19×10^{-132}.

In the swarm-based enrichment analysis, 20/20 independent analysis runs yielded 412 features, having a conditional probability of false discovery lower than 1.38 × 10-171.

3.2. Panther Ontology

The molecular function Panther Ontology results are shown in Table 2. The Panther molecular function classification reveals that the majority of the genes in the gene set are involved in binding (32.6%), a diverse class of proteins that can bind to other molecules, such as proteins, small ligands, and nucleic acids. An even larger percentage of genes (31.4%) do not have a Panther molecular function assigned.

Catalytic activity (26.4%) is one of the largest annotated functional classes, indicating that there is a large number of enzymes involved in a wide range of biochemical processes. Other highly ranked functional classes are transporter activity (5.5%), transcription regulator activity (6.7%), and molecular function regulator roles (6.2%). Other less prominent groups are molecular transducer activity (6.0%) and molecular adaptor activity (2.9%).

Table 2. Panther Ontology Molecular Function Results.

Rank	Molecular Function	Membership	Percentage membership
1	binding (GO:0005488)	136	32.60%
2	catalytic activity (GO:0003824)	110	26.40%
3	transcription regulator activity (GO:0140110)	28	6.70%
4	molecular function regulator activity (GO:0098772)	26	6.20%
5	molecular transducer activity (GO:0060089)	25	6.00%
6	transporter activity (GO:0005215)	23	5.50%

7	molecular adaptor activity (GO:0060090)	12	2.90%
8	structural molecule activity (GO:0005198)	10	2.40%
9	ATP-dependent activity (GO:0140657)	7	1.70%
10	translation regulator activity (GO:0045182)	4	1.00%
11	cytoskeletal motor activity (GO:0003774)	4	1.00%
12	cargo receptor activity (GO:0038024)	1	0.20%
13	antioxidant activity (GO:0016209)	1	0.20%

3.3. Reactome Pathway

Table 3 shows the Reactome pathway enrichment of the 412 ALL-enriched genes and shows that the pathways are overwhelmingly enriched in immune-related processes. In terms of the most significantly enriched pathway, the neutrophil degranulation pathway (62 genes; FDR = 1.13×10^{-13}) was activated, suggesting a significant activation of innate immune effector mechanisms. This was preceded by substantial enrichment of the general innate immune system and immune system pathways. Moderately enriched were cytokine-driven signaling pathways like interleukin signaling (IL-10, IL-4/IL-13, IL-6–type receptors). In addition, there was also evidence of enrichment of Rho/Rac GTPase cycles and related pathways of granulopoiesis and apoptosis.

Table 3. Reactome Pathway Assessment of the 412 ALL-enriched Genes.

Pathway name	Entities found	Entities Total	Entities ratio	Entities p Value	Entities FDR	Reactions found	Reactions Total	Reactions ratio
Neutrophil degranulation	62	478	0.03	1.11E-16	1.13E-13	10	10	0.001
Innate Immune System	80	1,350	0.084	3.42E-09	1.74E-06	188	762	0.049
Immune System	124	2,793	0.173	2.02E-06	6.84E-04	530	1,807	0.115
Signaling by Interleukins	37	646	0.04	1.50E-04	3.80E-02	206	505	0.032
Interleukin-10 signaling	10	86	0.005	3.20E-04	6.49E-02	10	15	0.001
RHOB GTPase cycle	9	75	0.005	5.04E-04	8.52E-02	3	6	0
Interleukin-4 and Interleukin-13 signaling	16	211	0.013	7.55E-04	1.09E-01	24	47	0.003
Transcriptional regulation of granulopoiesis	8	70	0.004	1.38E-03	1.76E-01	13	27	0.002
IL-6-type cytokine receptor ligand interactions	4	17	0.001	1.82E-03	2.06E-01	10	14	0.001
RHOF GTPase cycle	6	46	0.003	2.84E-03	2.49E-01	2	3	0
RHOC GTPase cycle	8	79	0.005	2.90E-03	2.49E-01	3	6	0
Caspase activation via the extrinsic apoptotic signalling pathway	5	32	0.002	2.96E-03	2.49E-01	11	17	0.001
Caspase activation via Death Receptors in the presence of ligand	4	20	0.001	3.25E-03	2.54E-01	9	9	0.001
Inflammasomes	5	34	0.002	3.82E-03	2.75E-01	6	28	0.002
Platelet degranulation	11	142	0.009	4.12E-03	2.76E-01	4	11	0.001
Nephron development	4	23	0.001	5.31E-03	3.08E-01	3	17	0.001
RAC1 GTPase cycle	13	191	0.012	5.57E-03	3.08E-01	4	6	0
CASP8 activity is inhibited	3	12	0.001	5.82E-03	3.08E-01	2	2	0
Response to elevated platelet cytosolic Ca2+	11	149	0.009	5.83E-03	3.08E-01	4	14	0.001
The NLRP3 inflammasome	4	24	0.001	6.15E-03	3.08E-01	4	18	0.001

3.4. Druggability Assessment of the ALL- enriched Gene Set

All 412 genes were then assigned a druggability score based on the criteria selected. Table 4 shows the top 19 most druggable gene targets, all belonging to the high druggability class, having a score of 8-10. The 3 most highly ranked targets based on druggability score are JAK2, PTGS1 (COX-1), and PTGS2 (COX-2), all with a maximum score of 10. The targets that were scored in a large cluster were: VSIR, TYK2, VDR, TNFRSF1B, HCK, IL17RA, NR1H2, IL1R1, and TNFSF13B. The targets with a druggability score of 8 were IL10RA, TLR8, TLR2, GRN, IFNGR1, RIPK3, and KCNQ1.

Table 4. Genes with a High druggability score from the druggability assessment of ALL-related genes.

Rank	Gene Symbol	Full Gene Name	Gene Ontology	Function	Druggability Class	Druggability Score	Druggability Rationale
1	JAK2	Janus Kinase 2	Kinase; JAK family	Hematopoiesis signaling	High	10	FDA-approved inhibitors
2	PTGS1	COX-1	Enzyme; Cyclooxygenase	Prostaglandin synthesis	High	10	NSAIDs target
3	PTGS2	COX-2	Enzyme; Cyclooxygenase	Inflammation prostaglandins	High	10	FDA-approved inhibitors
4	VSIR	V-Set Immunoregulatory Receptor	Immune checkpoint	T cell inhibition	High	9	Immune checkpoint target
5	TYK2	Tyrosine Kinase 2	Kinase; JAK family	JAK–STAT signaling	High	9	FDA-approved drugs exist
6	VDR	Vitamin D Receptor	Nuclear receptor	Vitamin D signaling	High	9	Nuclear receptor with ligands
7	TNFRSF1B	TNF Receptor 1B	Receptor: TNF family	TNF signaling	High	9	Multiple approved drugs
8	HCK	HCK Proto-Oncogene	Kinase; Src family	Myeloid signaling	High	9	Src family kinase
9	IL17RA	Interleukin 17 Receptor A	Cytokine receptor	IL-17 signaling	High	9	FDA-approved antibodies
10	NR1H2	LXR-beta	Nuclear receptor	Lipid metabolism TF	High	9	Nuclear receptor
11	IL1R1	Interleukin 1 Receptor 1	Cytokine receptor	IL-1 signaling	High	9	FDA-approved antagonist
12	TNFSF13B	BAFF	Cytokine; TNF family	B cell survival	High	9	FDA-approved antibody
13	IL10RA	Interleukin 10 Receptor Alpha	Cytokine receptor	Anti-inflammatory signaling	High	8	Cytokine receptor
14	TLR8	Toll-Like Receptor 8	Pattern recognition receptor	ssRNA recognition	High	8	TLR family drug target
15	TLR2	Toll-Like Receptor 2	Pattern recognition receptor	Bacterial recognition	High	8	TLR family validated
16	GRN	Progranulin	Growth factor	Neurotrophic factor	High	8	Secreted protein target
17	IFNGR1	Interferon Gamma Receptor 1	Cytokine receptor	IFN-gamma signaling	High	8	Cytokine receptor
18	RIPK3	Receptor Interacting Kinase 3	Kinase; Necroptosis	Necroptosis mediator	High	8	Kinase with inhibitors
19	KCNQ1	Potassium Channel Q1	Ion channel	Cardiac repolarization	High	8	Ion channel

3.5. Toxicological Risk Assessment

For all 412 genes, toxicological risk scores were generated that reflect their involvement in essential biological pathways and known toxicity pathways (including data from ProTox; Banerjee et al. 2024). Table 5 shows the gene candidates with the highest score (Score>7) after the refinement with the toxicological assessment. VDR was the highest-ranked toxicological risk target for low to moderate risk. A few other markers were also strong with good safety profiles: PTGS2 (COX-2), TYK2, HCK, IL17RA, NR1H2, IL1R1, TNFSF13B, TLR8, and TLR2.

Table 5. Top druggable candidates (Score >7) from the druggability assessment, further refined and sorted by toxicological assessment.

Rank	Gene Symbol	Full Gene Name	Gene Ontology	Function	Druggability Score	Druggability Rationale	Toxicological Risk	Toxicology Score	Toxicology Rationale

1	VDR	Vitamin D Receptor	Nuclear receptor	Vitamin D signaling	9	Nuclear receptor with ligands	Low–Moderate	4	Well-characterized target
2	PTGS2	COX-2	Enzyme; Cyclooxygenase	Inflammation prostaglandins	10	FDA-approved inhibitors	Moderate	5	CV risk
3	TYK2	Tyrosine Kinase 2	Kinase; JAK family	JAK–STAT signaling	9	FDA-approved drugs exist	Moderate	5	JAK inhibition tolerable
4	HCK	HCK Proto-Oncogene	Kinase; Src family	Myeloid signaling	9	Src family kinase	Moderate	5	Myeloid-specific
5	IL17RA	Interleukin 17 Receptor A	Cytokine receptor	IL-17 signaling	9	FDA-approved antibodies	Moderate	5	Infection risk
6	NR1H2	LXR-beta	Nuclear receptor	Lipid metabolism TF	9	Nuclear receptor	Moderate	5	Lipid homeostasis
7	IL1R1	Interleukin 1 Receptor 1	Cytokine receptor	IL-1 signaling	9	FDA-approved antagonist	Moderate	5	Inflammation
8	TNFSF13B	BAFF	Cytokine; TNF family	B cell survival	9	FDA-approved antibody	Moderate	5	B-cell homeostasis
9	TLR8	Toll-Like Receptor 8	Pattern recognition receptor	ssRNA recognition	8	TLR family drug target	Moderate	5	Immune modulation
10	TLR2	Toll-Like Receptor 2	Pattern recognition receptor	Bacterial recognition	8	TLR family validated	Moderate	5	Innate immunity
11	RIPK3	Receptor Interacting Kinase 3	Kinase; Necroptosis	Necroptosis mediator	8	Kinase with inhibitors	Moderate	5	Necroptosis pathway
12	CTSD	Cathepsin D	Enzyme; Aspartic protease; Lysosomal	Lysosomal protein degradation	7	Protease with a targetable site	Moderate	5	Essential lysosomal function
13	DGAT2	Diacylglycerol Acyltransferase 2	Enzyme; Lipid synthesis	Triglyceride synthesis	7	Metabolic target	Moderate	5	Lipid metabolism

3.6. Identified Compounds

Table 6 lists the physicochemical properties of selected compounds that were analyzed for their interaction with selected gene targets, and their drug-likeness was explicitly measured. Calcitriol and PS121912 are examples of VDR ligands. Celecoxib and etoricoxib are PTGS2 inhibitors. Some representative inhibitors of TYK2 (deucravacitinib, HCK inhibitors (dasatinib and bosutinib), NR1H2 agonists (T0901317 and GW3965) were identified.

The entire process flow is described in this study and is shown in Figure 3.

Table 6. Selected compounds for the key gene targets.

Target	Compound Name	Type	MW_Da	LogP	HBD	HBA	Rotatable Bonds	Aromatic Rings	Lipinski Violations
VDR	Calcitriol	Agonist	416.64	7.35	4	4	9	0	1
VDR	PS121912	VDR-coactivator inhibitor	~400	~3.5	2	4	4	2	0
PTGS2	Celecoxib	Selective COX-2 inhibitor	381.37	3.47	1	5	2	3	0
PTGS2	Etoricoxib	Selective COX-2 inhibitor	358.84	2.47	0	4	2	3	0
TYK2	Deucravacitinib	Allosteric TYK2 inhibitor	408.45	2.94	1	7	5	2	0
HCK	Bosutinib	Dual Src/Abl inhibitor	530.45	5.06	2	7	7	3	1
HCK	Dasatinib	Multi-kinase inhibitor	488.01	3.84	3	9	8	3	0
IL17RA	Experimental_compounds	IL-17A/IL-17RA PPI inhibitor	400-600	~3-5	01-Mar	05-Aug	05-Oct	02-Mar	0-1
NR1H2	T0901317	LXR agonist	414.41	7.03	1	4	7	3	1
NR1H2	GW3965	LXR agonist	465.96	6.24	1	6	9	2	1
IL1R1	Anakinra	IL-1 receptor antagonist	~17000	N/A	N/A	N/A	N/A	N/A	N/A

4. DISCUSSION

This study shows that the SBANN model can be used to identify transcription patterns that are different in ALL bone marrow compared to normal controls. The models had excellent predictive performance with AUC (from 0.814 to 0.972) and low RMS error on test data. The performance is remarkable, taking into consideration the high dimensionality and biological heterogeneity of the leukemia transcriptomic data.

Concordant 412 gene products were found in all 20 independent enrichment runs, and the conditional probability of false discovery (less than 1.38×10^{-171}) suggests a consistent and reliable biological signal. This level of convergence is not often seen in typical differential expression pipelines and suggests promise for the swarm-based approach.

The gene set was molecularly diverse according to the functional annotation carried out by Panther Ontology, and the major classes obtained were the binding and catalytic activities. This is in line with the fundamental importance of protein-protein interactions, enzyme regulation, and signal transduction in leukaemia biology. The percentage of unclassified genes is high, which may be due to the fact that some proteins are poorly characterised, or have lineage- or disease-specific functions in ALL.

These findings were further contextualised by pathway enrichment analysis of Reactome, which showed significant over-representation of innate immune and inflammatory pathways. Neutrophil degranulation, signalling by the innate immune system, and interleukin-mediated pathways (IL-6, IL-10, IL-17, and IL-4/IL-13) were heavily enriched. These findings are similar to those recently reported in the literature, which show that ALL is not just an issue of intrinsic lymphoid proliferation, but also of inflammatory signalling and bone marrow microenvironment remodelling [15, 16]. The complex interplay between cytoskeletal reorganization, inflammatory regulation, and cell survival in leukemic cells is further supported by the enrichment of Rho GTPase cycles, inflammasome, and apoptotic pathways.

This workflow is different from many previous computational studies that were mainly concerned with gene discovery or pathway enrichment, but also incorporate druggability and toxicological assessments. The pipeline uses SBANN-based feature selection, along with an LLM to score druggability and toxicity through structured prompts and well-known databases, enabling more translationally relevant target prioritisation.

4.1. Implications of the Findings

One of the peculiarities of this work is that toxicological data has been introduced in the initial target prioritisation stage. Although several highly enriched genes, including JAK2, TYK2, and PTGS2, have been established therapeutic targets for inflammatory and malignant diseases, they have been linked with systemic toxicities. The toxicity-aware ranking identified targets with more favourable safety profiles, including VDR, which was identified as a promising target with a well-characterised biology and low toxicological risk (score 4/10).

From a clinical translation perspective, several prioritised targets offer direct repurposing opportunities. VDR ligands such as calcitriol are already FDA-approved for other indications, suggesting potential adjunctive therapy in ALL, particularly given vitamin D's role in immune modulation and hematopoietic differentiation. TYK2, with existing FDA-approved inhibitors (deucravacitinib), provides a pathway to JAK-STAT modulation in ALL, though off-target immunosuppression requires careful evaluation. PTGS2 (COX-2) inhibitors (celecoxib, etoricoxib) represent another repurposing avenue targeting inflammation-driven leukemic progression. The integration of toxicological filtering early in the pipeline prioritises targets like VDR and NR1H2 over broader-acting kinases, potentially reducing late-stage attrition due to safety liabilities.

The framework has shown utility for practical applications since known drugs and experimental compounds targeting these prioritised candidates were identified. These computational predictions now require preclinical validation in patient-derived xenograft models and, where applicable, retrospective analysis of clinical outcomes in ALL patients exposed to these drug classes. In conclusion, this study provides an integrated computational approach capable of uncovering robust features with druggability and toxicology filtering for assisting in hypothesis generation for drug targets in ALL.

4.2. Limitations and Future Directions

The following are some limitations of this study. Only bulk transcriptomic data were analysed, and this approach may overlook the effects of cell types. Future studies may benefit from single-cell-based RNA sequencing and proteomic profiling for increased resolution. Moreover, the LLM-based scoring was structured but executed as a single run as opposed to assessing run-to-run variability, which would further enhance the reproducibility of the

scoring. The current SBANN implementation relies on univariate models (one gene per model), and it is difficult to directly capture gene-gene interactions.

To validate the prioritised targets and compounds, further study is required to assess the results on external validation from other independent cohorts of ALL, benchmark with standard methods (differential expression analysis), and experimental validation of the prioritised targets and compounds. Additionally, it would be useful to be able to assess the heterogeneity of ALL subtypes (such as B-ALL versus T-ALL).

5. CONCLUSION

This paper introduces a comprehensive computational pipeline for the prioritisation of potential drug targets in acute lymphoblastic leukemia (ALL) involving swarm-based artificial neural networks (SBANNs), systematic biological enrichment, and LLM-based drug target druggability and toxicity analysis. The performance of the SBANN models was good and showed a high level of consistency of features, resulting in a strong list of 412 leukemogenesis and immune deregulation genes associated with ALL.

The functional and pathway analyses revealed that the over-represented gene set is mainly related to immune and inflammatory responses, such as interleukin signaling, activation of innate immune responses, regulation of the cytoskeleton, and programmed cell death pathways. These results suggest that ALL is a complex process that involves interactions between leukemic cells and the immune microenvironment.

The integration of druggability and toxicological evaluation in the prioritisation pipeline is an important aspect of this work. Comments on toxicity early in the process will enable identification of targets that provide a balance of biological relevance and safety potential. This strategy identifies potential targets (including VDR and other immune targets) that have relatively good profiles.

Overall, this integrated framework will give a systematic and scalable solution for hypothesis generation in translational bioinformatics. It shows how swarm intelligence, ontology-driven analysis, and structured LLM-assisted filtering can be integrated to assist with prioritising potential therapeutic targets in ALL and other complex diseases.

Author Contributions: Conceptualization: S.Q.A., G.R.B., A.A.E.-S., R.M.G; Data Curation: S.Q.A., N.O.M., A.S.A., S.A.A., N.A.A; Formal Analysis: S.Q.A., A.A.E.-S., G.R.B; Investigation: S.Q.A., N.O.M., A.S.A; Methodology: S.Q.A., G.R.B., A.A.E.-S., R.M.G; Project Administration: G.R.B., R.M.G; Resources: G.R.B; Software: S.Q.A., A.A.E.-S; Supervision: G.R.B., R.M.G; Validation: S.Q.A., A.A.E.-S., N.O.M., S.A.A., N.A.A; Visualization: S.Q.A., S.A; Writing – Original Draft: S.Q.A; Writing – Review & Editing: S.Q.A., G.R.B., A.A.E.-S., N.O.M., S.A., R.M.G.

ACKNOWLEDGEMENTS: None.

REFERENCES

Ajonu CI, Grundy RI, Ball GR, Zafeiris D. Application of a high-throughput swarm-based deep neural network Algorithm reveals SPAG5 downregulation as a potential therapeutic target in adult AML. Functional & Integrative Genomics. 2025;25(1):8.

Bica I, Velickovic P, Xiao H, Li P, editors. Multi-omics data integration using cross-modal neural networks. ESANN; 2018.

Hunger SP, Mullighan CG. Acute lymphoblastic leukemia in children. New England Journal of Medicine. 2015;373(16):1541–52.

Lancashire LJ, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. Briefings in bioinformatics. 2009;10(3):315–29.

Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic acids research. 2021;49(D1):D394–D403.

Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Res. 2024;52(D1):D672–d8.

Mullighan C, Miller C, Radtke I, Phillips L, Dalton J, Ma J, et al. BCR-ABL1 lymphoblastic leukemia is characterized by the deletion of Ikaros. Nature. 2008;453:110–4.

Mullighan CG, Collins-Underwood JR, Phillips LA, Loudin MG, Liu W, Zhang J, et al. Rearrangement of CRLF2 in B-progenitor- and Down syndrome-associated acute lymphoblastic leukemia. Nat Genet.

2009;41(11):1243–6.
Pui CH, Robison LL, Look AT. Acute lymphoblastic leukaemia. Lancet. 2008;371(9617):1030–43.
Pui C-H, Yang JJ, Hunger SP, Pieters R, Schrappe M, Biondi A, et al. Childhood acute lymphoblastic leukemia: progress through collaboration. Journal of Clinical Oncology. 2015;33(27):2938–48.
Roversi FM, Bueno MLP, Pericole FV, Saad STO. Hematopoietic cell kinase (HCK) is a player of the crosstalk between hematopoietic cells and bone marrow niche through CXCL12/CXCR4 axis. Frontiers in cell and developmental biology. 2021;9:634044.
Shochat C, Tal N, Bandapalli OR, Palmi C, Ganmore I, te Kronnie G, et al. Gain-of-function mutations in interleukin-7 receptor-α (IL7R) in childhood acute lymphoblastic leukemias. J Exp Med. 2011;208(5):901–8.
Van Vlierberghe P, Ferrando A. The molecular basis of T cell acute lymphoblastic leukemia. The Journal of clinical investigation. 2012;122(10):3398–406.
Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013;45(10):1113–20.
Zafeiris D, Rutella S, Ball GR. An artificial neural network integrated pipeline for biomarker discovery using Alzheimer's disease as a case study. Computational and Structural Biotechnology Journal. 2018;16:77–87.
Zhang J, Ding L, Holmfeldt L, Wu G, Heatley SL, Payne-Turner D, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature. 2012;481(7380):157–63.