

Introduction to Data Mining in Biomedicine

Piyush More
02.12.2025

Understanding biological information

What's the *first thing* you do when you start any research project?

Literature Review

- There are ~37M articles indexed in PubMed¹
- In case of NGS, there are ~7M gene expression profiles deposited in GEO²

Need for a systematic method to access and process the information

Data Mining

¹ Publications Output: U.S. Trends and International Comparisons (<https://ncses.nsf.gov/pubs/nsb202333/publication-output-by-region-country-or-economy-and-by-scientific-field>)

² GEO overview (<https://www.ncbi.nlm.nih.gov/geo/>)

What is Data Mining?



Mining: Extracting **key information** from large datasets
Followed by finding interesting relationships → meaningful associations

Frequently used techniques

- Correlation and regression
- Clustering
- Classification
- Anomaly detection
- Natural Language Processing
 - AI and Large Language Models

Types of biomedical data

- Genomic Data
 - DNA sequences (mutations and variants), gene expression, chromatin accessibility
- Proteomic and Metabolomic Data
 - Protein levels, metabolic pathways
- Imaging Data
 - Medical imaging (MRI, CT scans, etc.) processed using machine learning techniques
- Clinical Data
 - Patient demographics, medical history, diagnoses, treatment plans, outcomes

Databases and Tools useful for Data Mining

- General
 - NCBI
 - PubMed (<https://pubmed.ncbi.nlm.nih.gov/>)
 - PubChem (<https://pubchem.ncbi.nlm.nih.gov/>)
 - MeSH (<https://meshb-prev.nlm.nih.gov/>)
 - more...
- Genomics
 - GEO (<https://www.ncbi.nlm.nih.gov/geo/>)
 - CCLE (<https://sites.broadinstitute.org/ccle/>)
- Pharmacology
 - DrugComb (<https://drugcomb.org/>)
 - Genomics of Drug Sensitivity in Cancer (<https://www.cancerrxgene.org/>)
 - Therapeutic Target Database (<https://idrblab.net/ttd/>)
- R, Python, SQL, and other programming environment

Medical Subject Heading (MeSH)

- Method for categorizing PubMed articles by assigning medical terms
- Uses standardized keywords to describe main topic and subtopics of the article

Hierarchical structure

Anatomy [A] [+](#)
Organisms [B] [+](#)
Diseases [C] [+](#)
 Infections [C01] [+](#)
 Neoplasms [C04] [+](#)
 Musculoskeletal Diseases [C05] [+](#)
 Digestive System Diseases [C06] [+](#)
 Stomatognathic Diseases [C07] [+](#)
 Respiratory Tract Diseases [C08] [+](#)
 Otorhinolaryngologic Diseases [C09] [+](#)
 Nervous System Diseases [C10] [+](#)
 Eye Diseases [C11] [+](#)
 Urogenital Diseases [C12] [+](#)
 Cardiovascular Diseases [C14] [+](#)
 Hemic and Lymphatic Diseases [C15] [+](#)
 Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16] [+](#)
 Skin and Connective Tissue Diseases [C17] [+](#)
 Nutritional and Metabolic Diseases [C18] [+](#)
 Endocrine System Diseases [C19] [+](#)
 Immune System Diseases [C20] [+](#)
 Disorders of Environmental Origin [C21] [+](#)
 Animal Diseases [C22] [+](#)
 Pathological Conditions, Signs and Symptoms [C23] [+](#)
 Occupational Diseases [C24] [+](#)
 Chemically-Induced Disorders [C25] [+](#)
 Wounds and Injuries [C26] [+](#)
Chemicals and Drugs [D] [+](#)
Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E] [+](#)
Psychiatry and Psychology [F] [+](#)
Phenomena and Processes [G] [+](#)
Disciplines and Occupations [H] [+](#)
Anthropology, Education, Sociology, and Social Phenomena [I] [+](#)
Technology, Industry, and Agriculture [J] [+](#)
Humanities [K] [+](#)
Information Science [L] [+](#)
Named Groups [M] [+](#)
Health Care [N] [+](#)
Publication Characteristics [V] [+](#)
Geographicals [Z] [+](#)

Drug-regulated CD33-targeted CAR T cells control AML using clinically optimized rapamycin dosing

Jacob Appelbaum ^{1 2 3 4}, April E Price ⁵, Kaori Oda ¹, Joy Zhang ⁵, Wai-Hang Leung ⁵, Giacomo Tampella ¹, Dong Xia ⁵, Pauline Pl So ⁵, Sarah K Hilton ⁵, Claudya Evandy ¹, Semanti Sarkar ¹, Unja Martin ⁵, Anne-Rachel Krostag ⁵, Marissa Leonardi ¹, Daniel E Zak ⁵, Rachael Logan ¹, Paula Lewis ⁵, Secil Franke-Welch ⁵, Njabulo Ngwenyama ⁵, Michael Fitzgerald ¹, Niklas Tulberg ¹, Stephanie Rawlings-Rhea ¹, Rebecca A Gardner ¹, Kyle Jones ⁶, Angelica Sanabria ⁶, William Crago ⁶, John Timmer ⁶, Andrew Hollands ⁶, Brendan Eckelman ⁶, Sanela Bilic ⁷, Jim Woodworth ⁷, Adam Lamble ^{1 4}, Philip D Gregory ⁵, Jordan Jarjour ⁵, Mark Pogson ⁵, Joshua A Gustafson ¹, Alexander Astrakhan ⁵, Michael C Jensen ¹

Affiliations + expand

PMID: 38502193 PMCID: [PMC11060733](#) DOI: [10.1172/JCI162593](#)

Abstract

Chimeric antigen receptor (CAR) designs that incorporate pharmacologic control are desirable; however, designs suitable for clinical translation are needed. We designed a fully human, rapamycin-regulated drug product for targeting CD33+ tumors called dimerizing agent-regulated immunoreceptor complex (DARIC33). T cell products demonstrated target-specific and rapamycin-dependent cytokine release, transcriptional responses, cytotoxicity, and in vivo antileukemic activity in the presence of as little as 1 nM rapamycin. Rapamycin withdrawal paused DARIC33-stimulated T cell effector functions, which were restored following reexposure to rapamycin, demonstrating reversible effector function control. While rapamycin-regulated DARIC33 T cells were highly sensitive to target antigen, CD34+ stem cell colony-forming capacity was not impacted. We benchmarked DARIC33 potency relative to CD19 CAR T cells to estimate a T cell dose for clinical testing. In addition, we integrated in vitro and preclinical in vivo drug concentration thresholds for off-on state transitions, as well as murine and human rapamycin pharmacokinetics, to estimate a clinically applicable rapamycin dosing schedule. A phase I DARIC33 trial has been initiated (PLAT-08, [NCT05105152](#)), with initial evidence of rapamycin-regulated T cell activation and antitumor impact. Our findings provide evidence that the DARIC platform exhibits sensitive regulation and potency needed for clinical application to other important immunotherapy targets.

Keywords: Cancer immunotherapy; Hematology; Leukemias; T cells; Therapeutics.

MeSH terms

[> Animals](#)
[> Female](#)
[> Humans](#)
[> Immunotherapy, Adoptive](#)
[> Leukemia, Myeloid, Acute* / drug therapy](#)
[> Leukemia, Myeloid, Acute* / immunology](#)
[> Leukemia, Myeloid, Acute* / pathology](#)
[> Leukemia, Myeloid, Acute* / therapy](#)
[> Male](#)
[> Mice](#)
[> Receptors, Chimeric Antigen / immunology](#)
[> Sialic Acid Binding Ig-like Lectin 3* / immunology](#)
[> Sialic Acid Binding Ig-like Lectin 3* / metabolism](#)
[> Sirolimus* / administration & dosage](#)
[> Sirolimus* / pharmacology](#)
[> T-Lymphocytes* / drug effects](#)
[> T-Lymphocytes* / immunology](#)
[> Xenograft Model Antitumor Assays](#)



Substances

[> CD33 protein, human](#)
[> Receptors, Chimeric Antigen](#)
[> Sialic Acid Binding Ig-like Lectin 3](#)
[> Sirolimus](#)

A curated database regarding chemical information including physical and chemical properties, biological activities, and literature evidence

COMPOUND SUMMARY

Etoposide

PubChem CID	36462
Structure	 2D 3D
Chemical Safety	 Laboratory Chemical Safety Summary (LCSS) Datasheet
Molecular Formula	$C_{29}H_{32}O_{13}$
Synonyms	etoposide Velipid 33419-42-0 Toposar trans-Etoposide View More...
Molecular Weight	588.6 g/mol Computed by PubChem 2.2 (PubChem release 2021.10.14)
Dates	Create: 2004-09-16 Modify: 2024-11-30
Description	Etoposide can cause cancer according to California Labor Code. It can cause developmental toxicity according to state or federal government labeling requirements. <ul style="list-style-type: none">California Office of Environmental Health Hazard Assessment (OEHHA) Etoposide is a beta-D-glucoside, a furonaphthohydroisole and an organic heterotetracyclic compound. It has a role as an antineoplastic agent and a DNA synthesis inhibitor. It is functionally related to a podophylotoxin and a 4'-demethyllepidopodophylotoxin. <ul style="list-style-type: none">CHEBI A semisynthetic derivative of podophylotoxin that exhibits antitumor activity. Etoposide inhibits DNA synthesis by forming a complex with topoisomerase II and DNA. This complex induces breaks in double stranded DNA and prevents repair by topoisomerase II binding. Accumulated breaks in DNA prevent entry into the mitotic phase of cell division, and lead to cell death. Etoposide acts primarily in the G2 and S phases of the cell cycle. <ul style="list-style-type: none">DrugBank: Toxin and Toxin Target Database (T3DB) View More...

Cite Download

CONTENTS

- Title and Summary
- 1 Structures
- 2 Names and Identifiers
- 3 Chemical and Physical Properties
- 4 Spectral Information
- 5 Related Records
- 6 Chemical Vendors
- 7 Drug and Medication Information
- 8 Pharmacology and Biochemistry
- 9 Use and Manufacturing
- 10 Identification
- 11 Safety and Hazards
- 12 Toxicity
- 13 Associated Disorders and Diseases
- 14 Literature
- 15 Patents
- 16 Interactions and Pathways
- 17 Biological Test Results
- 18 Taxonomy
- 19 Classification
- 20 Information Sources

14 Literature

14.1 Consolidated References

51,154 items

Search



SORT BY Publication Date - Most Recent

Treatment of High-Risk Gestational Trophoblastic Neoplasia

Publication Name: Hematology/Oncology Clinics of North America
Publication Date: 2024-12
PMID: 39322460 DOI: 10.1016/j.hoc.2024.08.014

Correction to: Colitis associated with persistent drug-induced immune dysregulation

Publication Name: Virchows Archiv : an international journal of pathology
Publication Date: 2024-11-09
PMID: 39120656 DOI: 10.1007/s00428-024-03878-6

Efficacy and safety of combined anlotinib-oral etoposide treatment for patients with platinum-resistant ovarian cancer

Publication Name: Journal of Gynecologic Oncology
Publication Date: 2024-11
PMCID: PMC11543247 PMID: 38670563 DOI: 10.3802/jgo.2024.35.e100

A Phase I Open-Label Study of Cediranib Plus Etoposide and Cisplatin as First-Line Therapy for Patients With Extensive-Stage Small-Cell Lung Cancer or Metastatic Neuroendocrine Non-Small-Cell Lung Cancer

Publication Name: Clinical Lung Cancer
Publication Date: 2024-11
PMID: 39307607 DOI: 10.1016/j.clc.2024.08.015

O-GlcNAcylation inhibition redirects the response of colon cancer cells to chemotherapy from senescence to apoptosis

Publication Name: Cell Death & Disease
Publication Date: 2024-10-19
PMCID: PMC11490504 PMID: 39426963 DOI: 10.1038/s41419-024-07131-5

<< First < Previous Page 1 of 10,231 Next > Last >>

PubChem

Therapeutic Target Database (TTB)

Home Advance Search Target Group Drug Group Patient Data Download

Search Whole Database

Search for Targets: Search Reset

Examples: EGFR; Vascular endothelial growth factor; Peramivir; Renal cell carcinoma; EGFR_HUMAN ...

Therapeutic Target & Drug Data for Coronavirus (COVID-19, MERS-CoV, SARS-CoV)

A comprehensive collection of anti-coronavirus drugs (small molecular drugs, monoclonal antibodies, protein/peptide drugs, combination drugs, vaccines, etc.) together with their corresponding therapeutic targets data from previous and recent coronavirus researches ([Click to Explore All Data](#))

Jump to the Star Target & Star Drug for COVID-19 of This Week

Search for Drugs: Search Reset

Examples: Cabozantinib; Ebola virus infection; Oseltamivir ...

Search Drugs and Targets by Disease or ICD Identifier: Search Reset

Examples: Influenza; Alzheimer; ICD9: 331; ICD10: G30 ...

Search for Biomarkers: Search Reset

Examples: 5-HT 2A receptor; Rheumatic fever; Candidosis; Dengue; HIV infection ...

Search for Drug Scaffolds: Please select a drug scaffold name Search Reset

A curated database concerning therapeutic protein, pathways, diseases, and corresponding drugs

Target Information

Content Navigation (All None)

Target General Information

Target ID	T15739 (Former ID: TTDC00118)	Top
Target Name	Cellular tumor antigen p53 (TP53)	
Synonyms	Tumor suppressor p53; Phosphoprotein p53; P53; Antigen NY-CO-13	
Gene Name	TP53	
Target Type	Clinical trial target	[1]
Disease	[+] 3 Target-related Diseases	+
Function	Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling a set of genes required for this process. One of the activated genes is an inhibitor of cyclin-dependent kinases. Apoptosis induction seems to be mediated either by stimulation of BAX and FAS antigen expression, or by repression of Bcl-2. Click to Show/Hide	
UniProt ID	P53_HUMAN ↗	
Sequence	MEEFQSDPSVEPPLSQEFTSFLKLLPENNVLSPQAMDMLSPDIEQWFTEDGPPDEAPRPEAAPVAPAAPPTPAAPAPAPSPMLSSVSPQKTYQ65YGRLLGFLHSGTAKSVTCTYSPALNKHFCOLAKTQVQLWLDSTPPPGTRVRMAIYKQ5QHTEVRRCPHHE Click to Show/Hide	
3D Structure	Click to Show 3D Structure of This Target	AlphaFold ↗
ADReCS ID	BADD_A02521 ↗ ; BADD_A03004 ↗ ; BADD_A05675 ↗ ; BADD_A06440 ↗ ; BADD_A08298 ↗	
HIT2.0 ID	T82.WE ↗	

Drugs and Modes of Action

Clinical Trial Drug(s)	[+] 17 Clinical Trial Drugs	+
Discontinued Drug(s)	[+] 3 Discontinued Drugs	+
Mode of Action	[+] 5 Modes of Action	+

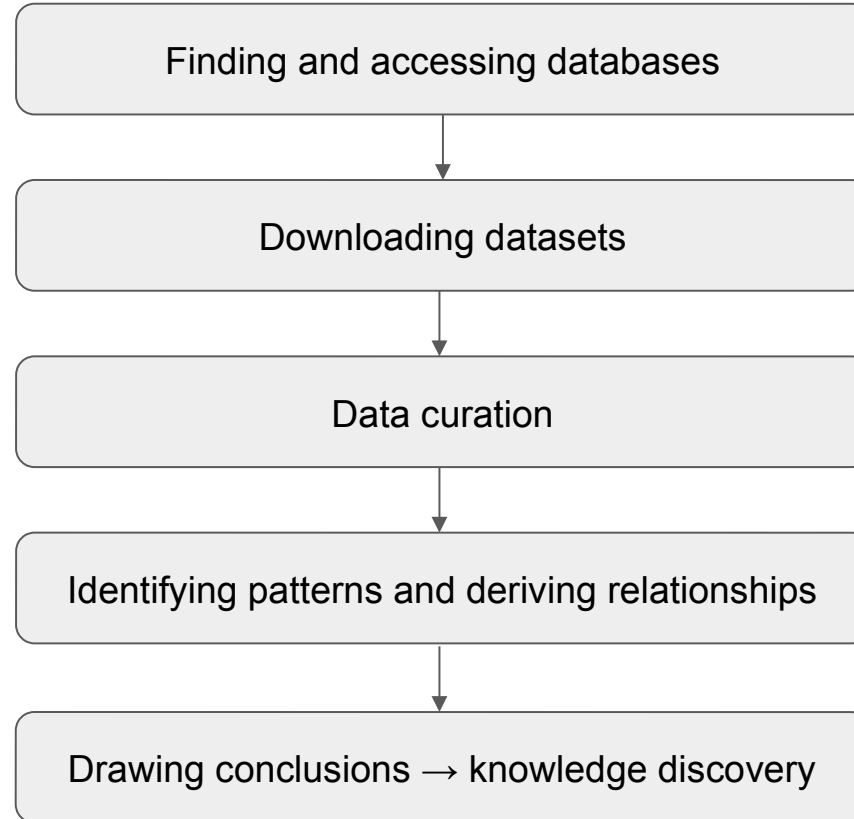
Cell-based Target Expression Variations

Cell-based Target Expression Variations [Info \[↗\]\(#\)](#)

Drug Binding Sites of Target

Ligand Name: Norleucine	Ligand Info	
Structure Description	SIRT1/Activator/Substrate Complex	PDB:4ZZJ
Method	X-ray diffraction	Resolution: 2.74 Å Mutation: No [38]
PDB Sequence	RHK ↗	

Typical data mining process



Typical challenges

- Data Quality
 - Noisy or incomplete data
- Data Integration
 - Combining heterogeneous data types (e.g., genomic, clinical, and imaging data)
- Interpretability
 - Understanding the results of complex models and ensuring they are actionable for biologists/clinicians
- Privacy and Ethics
 - Protecting patient data and ensuring ethical use of medical data.

Application 1: Drug Discovery

Goal: Identifying **new drug candidates** with therapeutic effects or **repurposing existing** drug candidates for novel therapeutic effect

Target identification

High-throughput drug screening

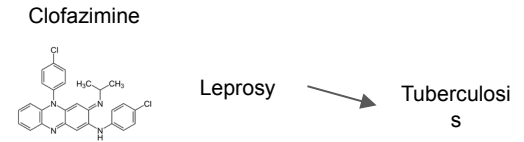
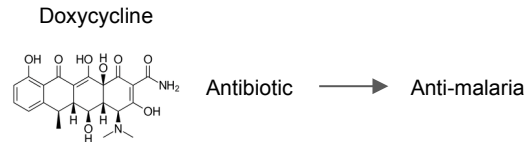
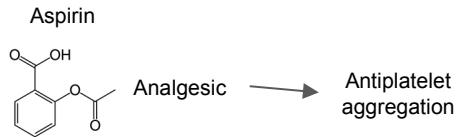
Drug-target interaction

ML models on large datasets to predict new interaction

Biomarker discovery

Mining genetic datasets to predict response

Examples



Application 2: Personalized Medicine

Goal: Tailor treatment to individual patients based on genetic, physiological, and other characteristics

Genetic sequencing

NGS and patient-specific biomarkers

Data integration

Electronic health records to modify treatment

Clinical trials

Identifying patients to be most benefited

Examples

HER2+ breast cancer treatment

Standard treatment
60-80% 5 year OS



Trastuzumab treatment
85-90% 5 year OS³

Drug dosing optimization

Statin dosing based on
SLCO1B1 genetic variant⁴

Wearable technologies

Glucose monitoring →
real-time insulin
adjustments

³ Piccart-Gebhart et al., 2005, Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer. *The New England Journal of Medicine*, 2005;353:1659-1672, DOI: 10.1056/NEJMoa052306

⁴ The SEARCH Collaborative Group, 2008, *SLCO1B1* Variants and Statin-Induced Myopathy — A Genomewide Study. *The New England Journal of Medicine*, 2008;359:789-799, DOI: 10.1056/NEJMoa0801936

To remember...

Before starting any experiment, mine the literature and databases to understand the pre-existing data

After obtaining the results, try to integrate with different types of publicly available data to get interesting insights

Exercise

The RNA-Seq data from Hs.505T (Chronic lymphocytic leukemia) cell line shows *BCL2* (Apoptosis regulator Bcl-2) overexpression. Identify drug approved for Chronic lymphocytic leukemia targeting this protein and identify which other genes are co-associated with BCL2 and the identified drug in the literature.

Steps

- Search Target BCL2 in the TTD database (<https://idrblab.net/ttd/>)
- Identify approved cancer drug and obtain the PubChem ID
- Search PubChem (<https://pubchem.ncbi.nlm.nih.gov/>) to get PubMed PMIDs of the articles
- Use provided 02_gene2pubmed-subset.csv file and use it with the 04_R-script.R script to identify co-associated genes (<https://tinyurl.com/datamining2025>)