Research Article

PathoAnalyzer-I: An Integrative Bioinformatics Platform for Chronic Disease Analysis

Bioinformatics and Biology Insights
Volume 20: 1–17

DOI: 10.1177/11779322261433663
journals.sagepub.com/home/bbi
Sage

Ali Aguerd[1], Faiza Bennis[1] and Fatima Chegdani[1]

Abstract

Chronic diseases impose a global health burden, contributing to high mortality and economic costs. Even with the recent surge in molecular data, these conditions remain largely incurable due to their biological complexity, data fragmentation, and analysis challenges, hindering early diagnosis, mechanistic understanding, and therapy. To address this, we developed PathoAnalyzer-I, an *in silico* platform that combines bioinformatics and machine learning to decipher chronic diseases. The tool requires no programming skills and provides a user-friendly interface for pathological analysis within a single framework. PathoAnalyzer-I uses a dataset of molecular data for 531 chronic diseases, from databases including the GWAS Catalog, PubChem, and STRING-db, enriched with machine learning–based predictions. Its dual-prediction system enhances molecular insights: one model imputes missing risk alleles with 77.6% accuracy, while a second predicts novel SNP–disease associations with 89.3% accuracy, providing avenues for future research. Applied to Alzheimer's disease, the platform identified diagnostic biomarkers (e.g. rs6733839T), core genes in disease mechanisms (e.g. *BIN1*, *APOE*, *SLC24A4*, *ABCA7*, *PTK2B*), major pathological mechanisms like amyloid processing, synaptic dysfunction, and cellular vulnerability, as well as therapeutic molecules including Beta-Lapachone and preventive compounds such as curcumin. With its features, PathoAnalyzer-I enables scientists, regardless of their resources, to conduct in-depth *in silico* studies of chronic diseases.

Keywords

Bioinformatics platform, machine learning prediction, chronic diseases, multi-omics integration, therapeutic discovery

Received: 30 October 2025; accepted: 2 March 2026

Introduction

Despite significant scientific progress in the health sector over recent decades, many chronic diseases remain incurable even with substantial investments aimed at combating them.[1] These conditions impose a considerable burden on multiple levels. They are responsible for millions of deaths worldwide each year and generate staggering economic costs.[2,3]

The persistence of these diseases is largely attributable to their biological complexity,[4] which hinders an in-depth understanding of their underlying mechanisms. As a consequence, effective prevention and curative strategies remain elusive, and early diagnosis is often challenging.

Alternatively, computational *in silico* analyses of chronic diseases represent a promising strategy to overcome current limitations. This raises a critical question: despite the availability of numerous high-throughput bioinformatics platforms and multi-omics datasets such as GWAS Catalog, Ensembl, and STRING-DB, why have these resources not been systematically used to characterize the molecular mechanisms that govern chronic pathologies, and to guide their prevention and therapeutic intervention? Several factors may contribute to this gap. First, although molecular datasets including genomics, transcriptomics, proteomics, and metabolomics are increasingly available, they remain highly fragmented and heterogeneous.[5] Currently, no integrative framework offers comprehensive interoperability across these data layers, and the functional links between different molecular entities are rarely captured, which hinders thorough multi-omics integration and analysis.[6] Second, many computational tools are technically demanding, often requiring proficiency in programming, as well as expertise in data normalization, network analysis, and predictive modeling.[7] These barriers limit broader adoption, resulting in the underutilization of resources that could

[1]Laboratory of Integrative Biology, Faculty of Science Ain Chock, University Hassan II, Casablanca, Morocco

Corresponding author:
Ali Aguerd, Laboratory of Integrative Biology, Faculty of Science Ain Chock, University Hassan II, Casablanca 20100, Morocco.
Email: aliagu97ninety@gmail.com

otherwise accelerate mechanistic understanding, biomarker discovery, and the identification of therapeutic and preventive molecules in chronic diseases.

To address these challenges, our team developed PathoAnalyzer-I, a Python-based tool designed to overcome the major obstacles in *in silico* studies of chronic diseases, including fragmented and disconnected datasets, limited interpretability, and the complexity of existing tools. This all-in-one platform enables scientists, regardless of their resources, to perform comprehensive *in silico* analyses in a fast and user-friendly manner while generating predictions that open new research directions. By integrating bioinformatics with machine learning, the tool allows researchers to:

- Elucidate the molecular mechanisms underlying chronic diseases;
- Identify biomarkers for early and confirmed diagnosis
- Discover potential therapeutic and preventive molecules;
- Explore novel research avenues through machine learning-driven predictions.

Overall, PathoAnalyzer-I enables comprehensive *in silico* investigations of chronic diseases, facilitating a deeper understanding of their persistence and ultimately contributing to advancements in public health.

Methodology

Theoretical background

Despite the human genome being 99% identical, the crucial 1% of genetic variation holds the key to understanding disease susceptibility.[8] These differences form the core of genome-wide association studies (GWAS), which link specific single-nucleotide polymorphisms (SNPs) to particular diseases and aid in deciphering incurable diseases, including their understanding, treatment, prevention, and early diagnosis. To achieve this goal, several *in silico* analyses must be undertaken:

- **Extract SNPs associated with chronic diseases:** These SNPs are considered biomarkers that enable diagnosis and their analysis provides a deeper understanding of chronic diseases.
- **Identify the features of SNPs associated with chronic diseases:**
  - **SNP localization**: Identifying whether a SNP located within or outside a gene helps predict its biological impact whether it affects gene regulation or protein function.
  - **Gene and Network Analysis**: The interaction of the genes carrying these SNPs allows us to construct interaction networks that can reveal the signaling pathways involved in the disease. Studying the proteins encoded by these genes and their biological functions is essential for understanding the signaling pathways and the mechanisms underlying the pathology.
  - **Therapeutic and Preventive Target Discovery:** After identifying key genes and their characteristics (such as encoded proteins and functions), we can pinpoint regulatory molecules that could serve as potential treatments or preventive agents.

The analysis of these elements necessitates processing tremendous volumes of biological data, which makes automation essential. Bioinformatics supported by tools such as Python, provides efficient solutions through libraries like Biopython[9] and Pandas,[10] enabling accurate and rapid data manipulation.

However, traditional bioinformatics has limitations, especially when addressing chronic diseases that may entail subtle or "silent" biomarkers. In this context, artificial intelligence represents a particularly powerful approach. Machine learning models, such as Random Forest,[11] can generate predictions on SNP-disease associations by leveraging linkage disequilibrium patterns, enabling the identification of potential biomarkers, which help deciphering chronic diseases.

To make this decoding process faster, more accessible, and fully automated, it is essential to bring all these stages together within a unified platform. Such a program would enable users to deeply understand any incurable disease and provide access to potential therapeutic strategies, all through a single, integrated tool. The core principle of our tool is summarized in Figure 1 below.

Dataset

Our methodology was designed to develop a Python program that provides the following functionalities:

- Comprehensive disease descriptions for chronic diseases
- Analysis of SNPs associated with chronic diseases
- Identification and visualization of signaling pathways associated with chronic diseases, with interactive data enrichment
- Discovery of potential therapeutic and preventive molecules for chronic diseases
- Prediction of novel SNP-disease associations

Raw data. To decode chronic diseases, it is essential to rely on reliable, comprehensive, and accurate data. To meet this need, our program's dataset was built by integrating high-quality information into a robust system, further enriched with predictions generated by a powerful machine learning model.

First and foremost, this dataset was built based on the GWAS Catalog[12] via web scraping, from which we extracted genomic data associated with 531 chronic pathologies in TSV format. Each file encompasses valuable columns on SNPs, including *riskAllele*, which contains the SNP ID and risk allele (e.g. in the format rs123456-A);

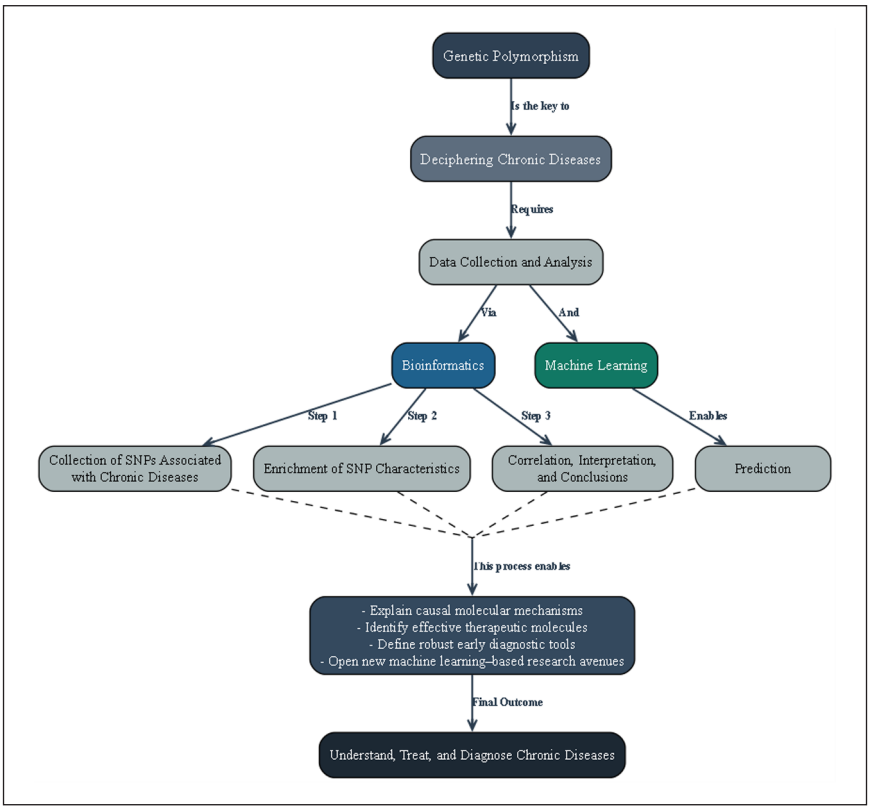

Figure 1. Core principle of PathoAnalyzer-I.

locations, which provides the chromosome harboring the SNP and its precise genomic position (e.g. formatted as 14:45678); *pValue*, which indicates the statistical significance of the SNPs; *mappedGenes*, which indicates the gene harboring the SNP; as well as *traitName*, *accessionId*, and *authors*.

Preparing predictions. A central methodological objective of this work is to generate predictive biomarkers of chronic diseases in the form of single-nucleotide polymorphisms (SNPs), by leveraging SNPs already associated with chronic conditions to predict novel candidates under the rationale that genetically proximal variants, as indicated by their genomic positions, may exert similar effects and, through linkage disequilibrium, be linked to the same disorder.

The first step of prediction was the pre-processing of the data, to this end, we followed the procedure below:

1. Relevant feature selection: *RiskAllele*, *location*, *mappedGene*, and *pValue*, as these features capture the genomic position of variants and their statistical significance.
2. Context enrichment: Added the *geneticontext* (e.g. intron/exon status) using Ensembl[13] via web scraping, serving as a functional and positional indicator for each SNP.
3. Quality control: Removed approximately 4.07% of SNPs with incomplete annotations.
4. Field normalization:
   - Split *location* into *chromosome* and *position*.
   - Separated *RiskAllele* into *variant* (RSID) and *riskAllele* (A, T, C, G).

During this process, we encountered a limitation: approximately 11.69% of the entries contained only the SNP locus ID without the corresponding risk allele (marked as "?"). This prompted a two-tiered prediction strategy: first, predicting missing risk alleles to complete the dataset; second, using the complete dataset to discover novel SNP–disease associations.

Phase 1. Predicting missing risk alleles. The first phase, aimed at predicting missing risk alleles, is detailed in Table 1, which presents the data sources, engineered features, and model specifications.

The combination of raw and engineered features provides a biologically grounded framework, linking variants, genes, and pathologies while capturing regional, density, and frequency patterns to predict missing risk alleles. CatBoostEncoder transforms categorical variables into numerical representations while preserving their statistical relationships with the target, and Random Forest leverages these representations to model complex interactions and accurately impute missing risk alleles.

Phase 2. Prediction of novel SNP–pathology associations. The second phase of the framework, which aimed to predict new SNP–pathology associations, is detailed in Table 2.

Table 1. Predicting missing risk alleles.

Objective	Predict missing risk alleles (approximately 11.69% of the total)
Data Used	Fully annotated SNPs (approximately 88.31% of the total)
Raw Features	*RSID, riskAllele, chromosome, position, geneticontext, mappedGenes*, **pathology name** (Generated from file name).
Engineered Features	• **interval**: 5 kb genomic windows • **chrom_interval**: chromosome + interval combination • **gene_count**: number of associated genes. • **variant_freq**: variant frequency.
Encoding	CatBoostEncoder.
Model	Random Forest – 100 trees, max_depth = 50, min_samples_split = 3, min_samples_leaf = 2, class weighting

Table 2. Predicting new SNP–pathology associations.

Objective	Predict new SNP–pathology associations
Raw Features	***chromosome; position; variant; riskAllele; pValue*** - These features provide the biological and statistical foundation for prediction.
Engineered Features	• **interval**: 10 kb spatial windows (Linkage Disequilibrium block modeling) • **chrom_interval**: Unique spatial identifier (chromosome + interval) • **log_pvalue**: Logarithmic transformation ($-\log_{10}$[pValue]) • **pvalue_quantile**: Cross-study normalized p-values - These features capture the genomic topology, normalize association signals, and model linkage patterns.
Model	Random Forest – 200 trees, max depth = 30, Synthetic Minority Over-sampling Technique (SMOTE) for rare class oversampling

To mitigate overfitting, SNPs were divided into groups according to their occurrence frequency (Table 3).

As shown in Table 3, rare variants (⩽4 occurrences) are particularly prone to overfitting due to statistical underrepresentation.

Frequency-based stratification therefore allows to:

- Learn group-specific predictive rules
- SMOTE to compensate class imbalance
- Prevent dominance of common variants in training
- Enable stratified cross-validation for unbiased rare-variant evaluation

The two prediction phases were implemented in Python using standard libraries for data manipulation (Pandas, NumPy),[14] machine learning and categorical encoding (scikit-learn,[15] category_encoders[16]), class imbalance correction imbalanced-learn,[17] and visualization (Seaborn,[18] matplotlib[19]). Additional utilities were used for file management, logging, and computational support (os,[20] datetime,[21] logging,[22] sys,[23] hashlib,[24] warnings,[25] collections,[26] joblib[27]).

Table 3. Variant classification by occurrence frequency.

Group	Occurrence Threshold
COMMON_VARIANT	⩾50
MODERATE_VARIANT	20–49
RARE_VARIANT	5–19
VERY_RARE_VARIANT	2–4
ULTRA_RARE_VARIANT	1

Model validation. The robustness and generalizability of the machine learning models were assessed using validation frameworks based on simulated genomic data. These validations were conducted without altering the original preprocessing pipelines, feature engineering logic, or hyperparameter configurations, in order to closely reflect real analytical conditions. For the risk allele prediction model, simulated GWAS data comprising 20 000 variants were generated with biologically realistic associations among genomic contexts, genes, pathologies, and nucleotide distributions. Model performance was evaluated using stratified 5-fold cross-validation. A similar validation strategy was applied to the SNP–pathology association prediction model. In this case, 20 000 simulated variants were distributed across five frequency-based categories, ranging from COMMON_VARIANT to ULTRA_RARE_VARIANT, with *P*-value distributions calibrated to reflect realistic GWAS patterns. Performance was likewise assessed via 5-fold cross-validation. This simulated-data validation phase confirmed the robustness and stability of both models (detailed results are presented in Results: Option 5 – AI-based SNP prediction).

Final dataset. Following these steps, we constructed a clean and robust dataset containing the information required to elucidate the molecular mechanisms underlying chronic and incurable pathologies. Specifically, this dataset comprises 531 files, each named after a chronic incurable pathology and containing two subfolders: the "Reference" subfolder, which includes a TSV file listing GWAS study references associated with the corresponding pathology. Among the key columns in this file are:

- ***accessionId*** (GWAS study identifier)
- ***discoverySampleAncestry*** (ancestry and populations of the discovery sample)
- ***replicationSampleAncestry*** (ancestry and populations of the replication sample)

The "Pathology" subfolder contains a TSV file with the following columns:

- ***riskAllele*** (contains SNP ID and risk allele)
- ***pValue*** (indicates SNP significance)
- ***mappedGenes*** (Contains the gene carrying the SNP)
- ***locations*** (Contains the chromosome carrying the SNP and the exact chromosomal position)

- ***geneticontext*** (contains genetic context e.g. intron, exon . . .)
- ***Prediction*** (contains predicted SNPs)
- ***Context-Prediction*** (contains the genetic context of the predicted SNPs, e.g. intron, exon . . .)
- ***Location*** (Contains the chromosome carrying the predicted SNP and the exact chromosomal position)
- ***MappedGene*** (Contains the gene carrying the predicted SNP)
- ***accessionId***

Dynamism of our dataset. Alongside our extensive local dataset, the Python-based program PathoAnalyzer-I incorporates a dynamic data-retrieval system to enhance and update analyses in real time. Data is gathered from reliable and trusted databases, including:

- OLS (Ontology Lookup Service):[28] used for detailed pathology descriptions
- GWAS Catalog: used for SNP-disease associations
- PDB (Protein Data Bank):[29] used for 3D protein structures
- Universal Protein Resource (uniprot):[30] used for FASTA sequences and protein functions.
- Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) Database:[31] used for gene and protein interaction networks
- PubChem Database:[32] used for therapeutic and preventive molecules
- Ensembl Genome Browser: used for SNP genomic context
- Genotype-Tissue Expression (GTEx) project:[33] used for protein tissue expression data

To ensure fast and efficient performance, we implemented a caching mechanism that streamlines API calls by reducing latency, avoiding redundant queries, and improving productivity.

Another dynamic characteristic of our dataset is its continuous growth, as user queries for previously unrecorded pathologies trigger their automatic inclusion. Indeed, when a new pathology is queried, a dedicated folder is created to store the data retrieved via web scraping. Furthermore, for each newly integrated pathology, key information such as tissue-specific gene expression levels, characteristic SNPs plot, and 3D protein structures is automatically captured during tool operation.

By integrating data from the most trusted databases, machine learning-based predictions, and real-time data enrichment, our dataset brings together all essential molecular data needed to understand the mechanisms behind incurable pathologies, identifies potential therapeutic molecules to fight against them, and facilitates early diagnosis.

General structure of the program

PathoAnalyzer-I is implemented as a modular Python package. Its architecture is organized into core modules:

- **Api/:** Handles interactions with essential biological databases (GWAS Catalog, UniProt, PDB, STRINGdb, PubChem, Ensembl/GTEx).
- **Utils/:** Provides utilities for data analysis, file management, and display formatting.
- **Deciphering_incurable_pathologies/:** Contains the main analysis logic, pathology classes, and the primary entry point.
- **Dataset/:** Stores pathology-specific data.

This modular design ensures maintainability, scalability, and clear separation of concerns.

Our tool, relies on several Python libraries chosen for their performance and ability to process biological data efficiently. For user interaction and display, we employed *prompt_toolkit*[34] to build interactive command-line applications, *PrettyTable*[35] to generate tables in text format, and *Colorama*[36] to add colors to terminal outputs. For web access and automation, the program uses *Requests*[37] to perform HTTP queries, *Selenium*[38] for browser automation, and *webdriver_manager*[39] to manage browser drivers. Regarding data and image processing, *Pillow*[40] is used for handling pictures, while *FuzzyWuzzy*[41] supports fuzzy string matching. For visualization and network analysis, *Matplotlib* and *Plotly*[42] generate static and interactive graphics, and *NetworkX*[43] is dedicated to graph and network manipulation. Finally, for data manipulation and numerical computing, *Pandas* ensures efficient dataset management, and *NumPy* provides optimized tools for numerical analysis.

Program workflow

Leveraging its extensive and dynamic local dataset, various Python libraries, and its integration with multiple external biological data sources, PathoAnalyzer-I offers a comprehensive suite of analytical tools for chronic disease research.

After the user selects a pathology, the program searches for a locally available TSV file in the dataset that matches the entered name. If no compatible file is found, the program automatically queries the GWAS API to retrieve the corresponding TSV files. Once the data are available, the program displays an interactive menu composed of five options:

Option 1: pathology description. This option displays a detailed description of the selected pathology. The program queries the OLS API to retrieve pathology description. The information is then formatted and displayed in a readable console layout. This option serves as an introduction to the analysis, allowing the user to verify and understand the pathology before exploring its genetic and molecular aspects.

Option 2: SNP characteristics. This option analyzes SNPs associated with the pathology under three distinct scenarios, depending on the available data:

1. **Multi-population studies with shared SNPs**: This is the standard case, in which only SNPs shared across multiple populations (at least two) are considered. To rank these SNPs, a composite score is calculated as follows:

$$\text{Score} = 0.3 \times (-\log_{10}(\text{p})) + 0.3 \times (\text{sample\_size} / 1000) + 0.4 \times (\text{number\_of\_populations})$$

This score uses the same coefficient for two pieces of information—*P*-value and sample size (normalized by 1000 to ensure comparable scale)—to capture statistical significance, and a higher coefficient for the number of populations carrying the SNP, as this information enables consideration of the SNPs at an international level.

2. **Multi-population studies without shared SNPs**: This is the first particular case, in which GWAS data were generated from multiple populations (at least two), but no SNP is shared by at least two populations. In this case, a score is calculated to rank the SNPs as:

$$Score = 0.5 \times (sample\_size / 10000) + 0.5 \times (-\log_{10}(p))$$

This score does not take into account the number of populations carrying the SNP, as all SNPs are assigned a coefficient of 1.

3. **Single-population studies:** This is the second special case, in which GWAS results were generated from a single population. In this case, SNPs are ranked solely by *P*-value, as there is only one sample and one population.

The most accurate and effective scenario for studying a pathology is the first one, as it uses SNPs shared across multiple populations (at least two). In the absence of this condition, which depends on available data, the program offers the other two particulars' scenarios to users, which are based on adapted scoring. In all three cases, the program allows SNP visualization in both tabular and interactive graphical formats and provides detailed information for any selected SNP, including genomic context, associated genes, 3D protein structure, and tissue expression profiles.

Option 3: cell signaling pathways. This option extracts genes associated with characteristic SNPs (according to the previously described filter) and displays them in a gene–protein table. The analysis can be further enriched (based on the user's choice) by generating a protein–protein interaction network using the STRING database, which can be interactively visualized. Users can also explore detailed information for any selected gene.

Option 4: therapeutic molecules. This option focuses on identifying potential therapeutic molecules by retrieving chemical regulators of a selected gene or protein from PubChem. Each candidate molecule is evaluated through a double filter:

1. **Safety filter**: Eliminates molecules with hazardous Globally Harmonized System of Classification and Labelling of Chemicals (GHS) classifications—including carcinogenicity (H350, H350i, H351), mutagenicity (H340, H341, H342), reproductive toxicity (H360, H360F, H360D, H360FD, H360Fd, H360Df, H361, H361f, H361d, H361fd, H362), specific target organ toxicity (H370, H371, H372, H373), acute toxicity (H300, H301, H302, H304, H310, H311, H330, H331, H332), aquatic toxicity (H400, H410, H411, H412, H413), flammability hazards (H225, H226), skin and eye hazards (H314, H315, H317, H318, H319), respiratory and systemic hazards (H335, H336), and compounds with a "Danger" signal word accompanied by the Health Hazard pictogram (GHS08).
2. **ADMET filter**: assesses the remaining molecules using eliminatory and non-eliminatory criteria: the eliminatory rules—Lipinski's Rule of Five (MW $\leqslant$ 500, LogP $\leqslant$ 5, HBD $\leqslant$ 5, HBA $\leqslant$ 10), Veber rules (rotatable bonds $\leqslant$ 10, TPSA $\leqslant$ 140), and Egan rules (LogP $\leqslant$ 5.88, TPSA $\leqslant$ 131.6)—must all be satisfied, while the non-eliminatory Ghose filter (MW 160–480, LogP −0.4 to 5.6) and Muegge criteria (MW 200–600, LogP −2 to 5, HBD $\leqslant$ 5, HBA $\leqslant$ 10, rotatable bonds $\leqslant$ 15) generate only warnings. A quantitative bioavailability score (0.0–1.0) is calculated based on the percentage of passed rules to prioritize candidates.

After this double filtering, an interactive HTML report is generated to present the accepted molecules (those passing both filters) and the rejected ones (those failing at least one of the two filters).

Option 5: AI-based SNP prediction. This option reads TSV files containing predicted SNPs and their annotations for the selected pathology. The program displays the list of predicted SNPs and allows users to access detailed information for any selected SNP, including genomic context, location, and associated gene.

This analysis follows the workflow illustrated in Figure 2.

Results

By integrating bioinformatics and machine learning techniques, we developed a tool providing a comprehensive framework for *in silico* analysis of chronic diseases.

The program requires the user to input the name of the target disease, after which it displays an analysis menu, as shown in Figure 3. In cases of incorrect input, the program automatically identifies and suggests the most closely matching valid disease name.

Option 1. Pathology description. This option displays a detailed scientific description of the pathology.

For the rest of the features, we will focus on Alzheimer's disease as a case study:

Option 2. SNPs characteristics. This option provides the characteristic SNPs of the studied pathology (according to the principle explained above).

In the case of Alzheimer's disease, the top 10 SNPs are: rs6733839T, rs12590654-A, rs11771145-A, rs12151021-A, rs73223431T, rs2830489T, rs2526377-G, rs602602-A, rs871269T, and rs11218343C.

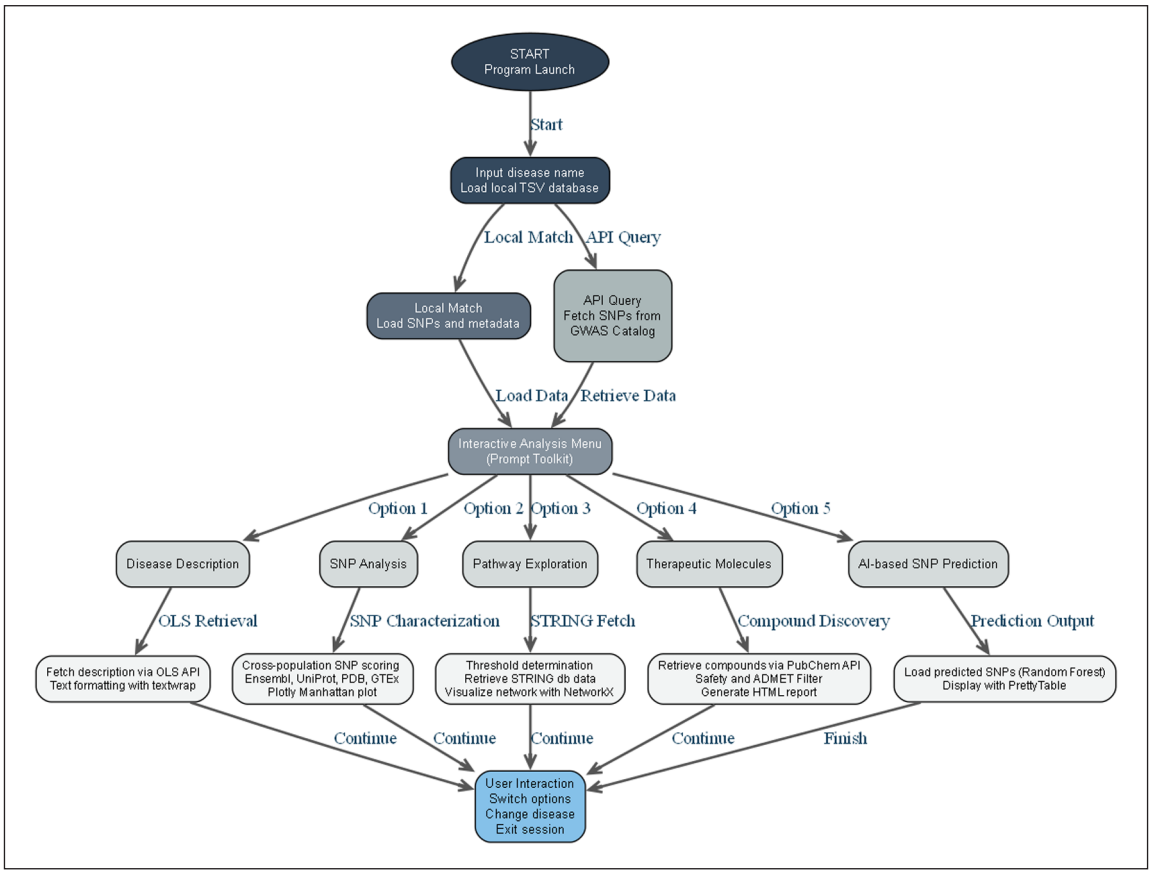

Figure 2. PathoAnalyzer-I workflow overview.

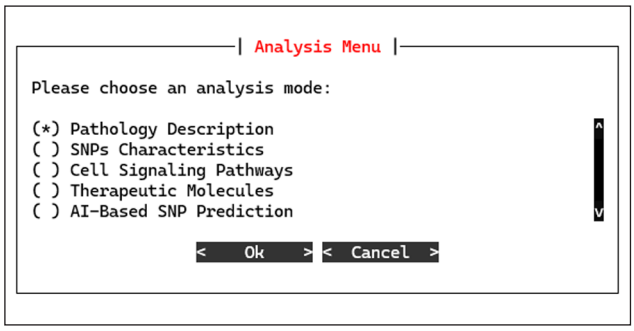

Figure 3. Analysis menu.

In addition to the tabular display, the user can choose the number of SNPs to display in graphical form (Figure 4).

For a detailed view of each biomarker, users can select an SNP to obtain its information. Taking SNP rs6733839 as an example, PathoAnalyzer-I displays risk allele (T), chromosomal location (chromosome 2, position 127135234), carrier gene (*BIN1*), variant type (intronic), gene function (the gene encodes a protein which regulates membrane dynamics, endocytosis, amyloid-beta production, and cytoskeletal organization, contributing to neuronal and cellular function), encoded protein information including 3D structure and amino acid sequence, and tissue-specific expression levels illustrated in Figure 5. In addition, the program provides the predicted impact of this mutation, including potential effects on splicing and regulatory elements within the intron.

Option 3. Cell signaling pathways. This option enables the exploration of a gene interaction network associated with the selected pathology.

Figure 6 illustrates a representative gene interaction network underlying Alzheimer's disease.

In addition, users can access information on any selected gene: encoded protein, its function, its 3D structure, and amino acid sequence.

Option 4. Therapeutic molecules. PathoAnalyzer-I also identifies potential therapeutic and preventive molecular candidates for chronic diseases. For Alzheimer's disease, the program proposes several compounds of interest including Dorsomorphin, Beta-Lapachone, Butyrate, and Curcumin.

Option5. AI-based SNP prediction. PathoAnalyzer-I is characterized by the integration of two machine learning models, enabling enhanced decoding of chronic pathologies.

The first model specialized in risk allele prediction achieves 77.6% overall accuracy on real data. Complementary validation on simulated data showed comparable performance with an average accuracy of 75.1% (SD = 0.62%) and a macro-averaged F1 score of 0.730. This agreement suggests that the model maintains stable behavior when applied to new datasets. The detailed performance on real data is presented in Table 4 below:

To illustrate the classification performance and the distribution of prediction errors across nucleotide classes, the corresponding confusion matrix is shown in Figure 7.

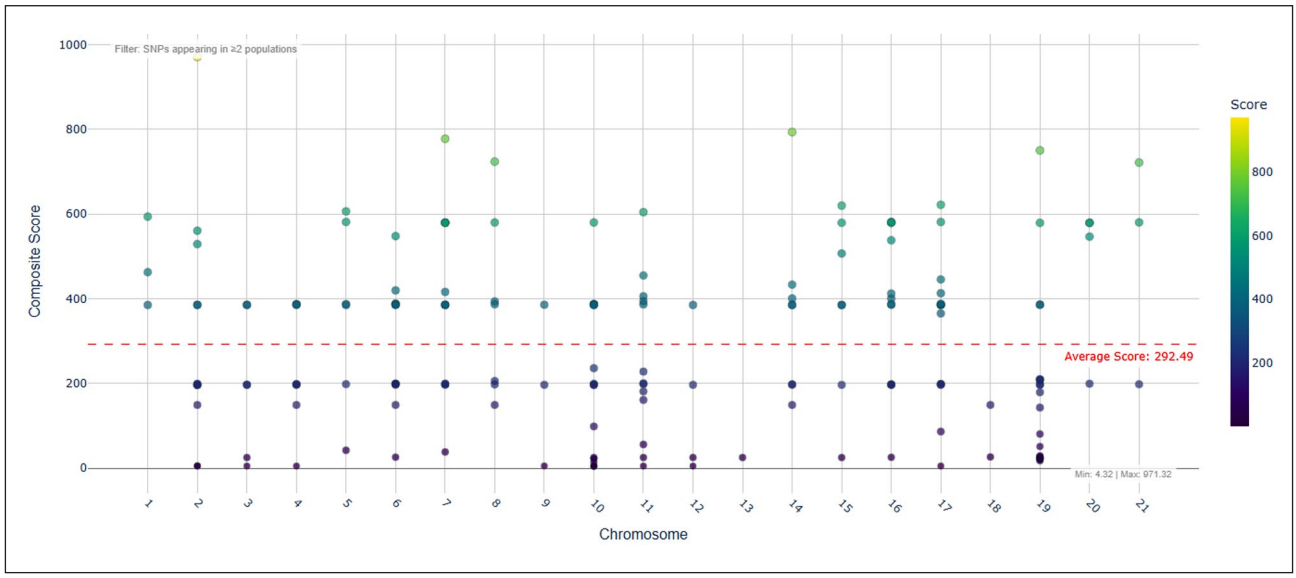

Figure 4. Characteristic SNPs plot of the genome-wide association study for Alzheimer's disease.
Each data point represents an individual single-nucleotide polymorphism (SNP). The y-axis shows the score of the association with the disease, and the x-axis shows the genomic position by chromosome.

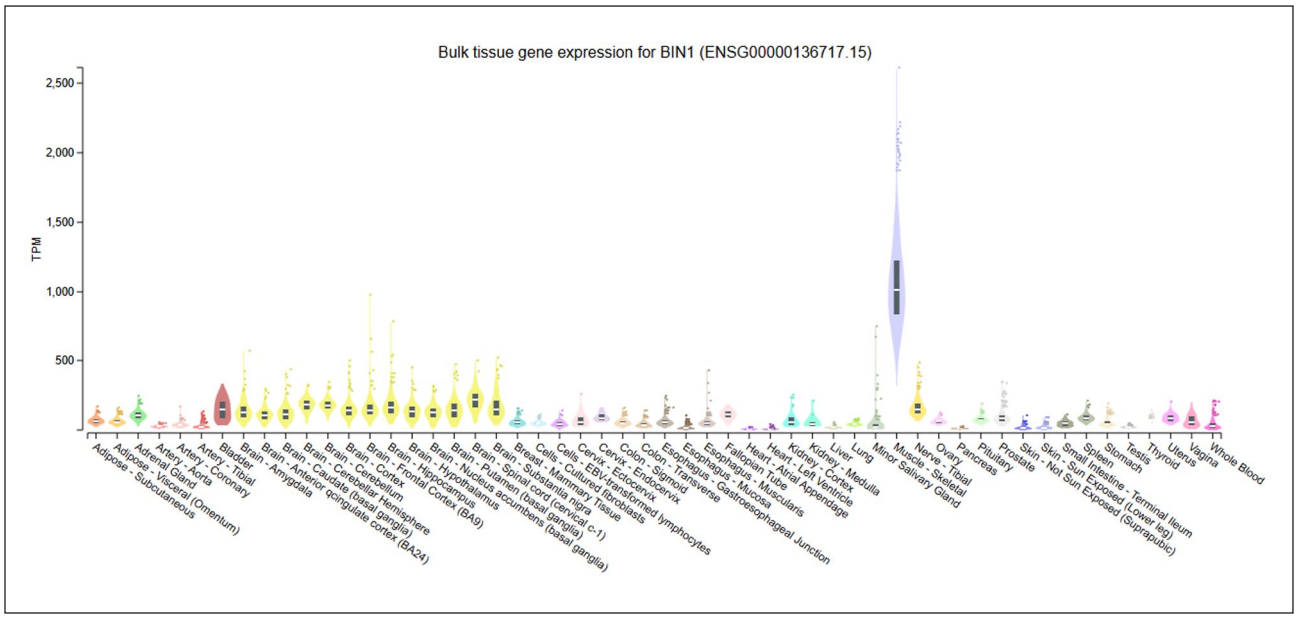

Figure 5. Tissue-specific expression level of BIN1, the carrier gene for SNP rs41289512.
The plot displays the transcript per million (TPM) values for the gene *BIN1* (ENSG00000136717.15) across a panel of healthy human tissues. Each data point represents the expression level measured in a bulk tissue sample. Samples are grouped and colored according to their tissue of origin. The y-axis indicates the expression level (TPM), while the x-axis lists the different tissue types.

The confusion matrix demonstrates that the model exhibits an overall satisfactory discriminative capability, with the majority of predictions correctly located along the main diagonal, reflecting accurate classification of the four nucleotides. For class A (adenine), 5997 predictions were correct, although notable misclassifications occurred toward T (622 instances) and G (434 instances). Similarly, cytosine (C) was accurately identified in 4267 cases, with errors mainly involving confusion with T (602 cases) and A (511 cases). Guanine (G) achieved 4207 correct predictions while being occasionally misclassified as A (624 cases) and C (354 cases). Finally, thymine (T) yielded 5511 correct classifications, with fewer confusions observed toward A (553 cases) and C (422 cases). Overall, the distribution of errors remained relatively balanced across classes, without evidence of substantial overprediction or underprediction biases. These results are consistent with the precision and recall metrics presented in the performance table (Table 4).

The second model of PathoAnalyzer-I, which specializes in SNP prediction, achieves 89.3% overall accuracy, as detailed in Table 5 below. Validation on simulated data confirmed this robustness with an average accuracy of 87.9%

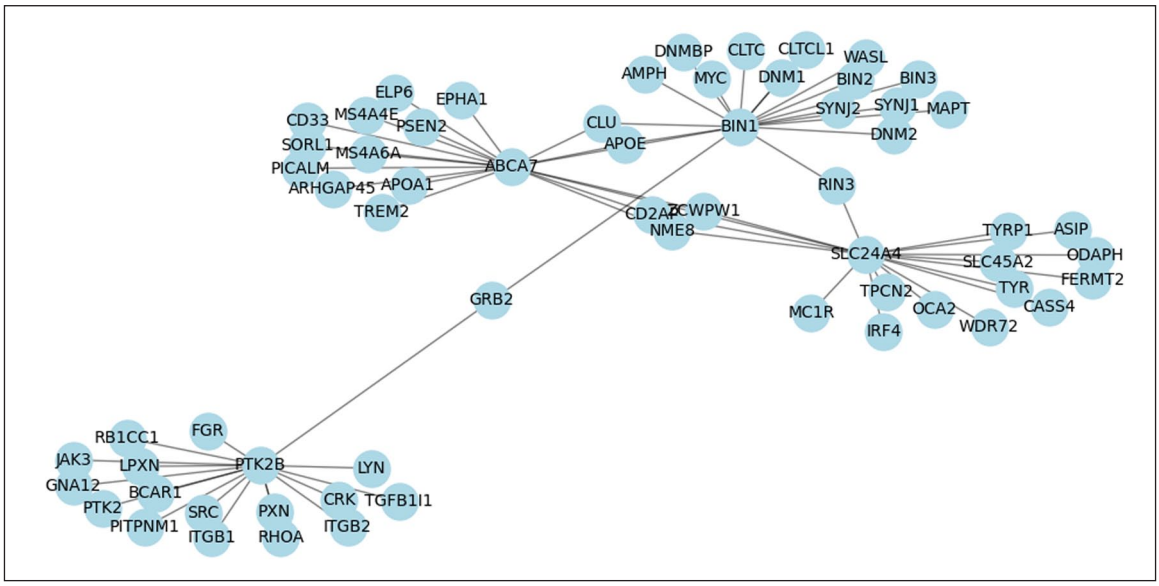

Figure 6. Expanded interaction network at threshold 17.
The network shows the gene–gene interactions detected at a significance threshold of 17 (this is the minimum threshold that allows the construction of a closed network in this case). Each node represents a gene, and each edge represents an interaction between two genes. Nodes are grouped into clusters that reflect shared pathways or biological functions. The layout highlights the connectivity patterns among interacting genes.

Table 4. Performance of the allele-prediction model by nucleotide base.

Nucleotide	Precision	Recall	F1 score	Support
A	0.78	0.81	0.79	7434
C	0.79	0.74	0.76	5782
G	0.78	0.74	0.76	5696
T	0.76	0.81	0.78	6839

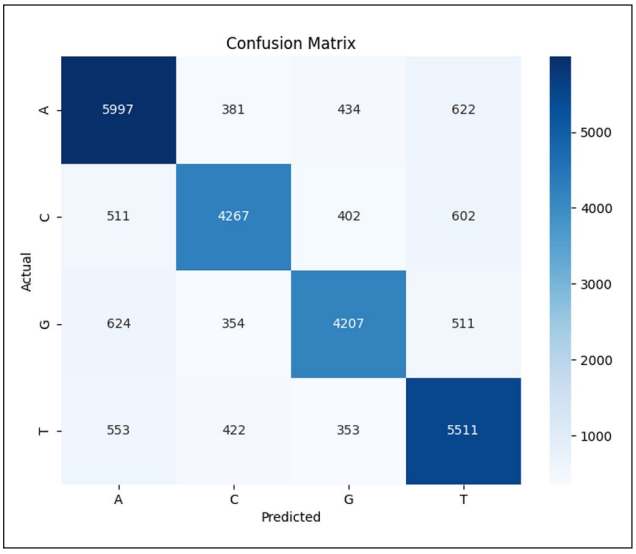

Figure 7. Confusion matrix of the risk allele prediction model.
The matrix displays the counts of correct and incorrect predictions made by the risk allele prediction model. Rows correspond to the actual nucleotides (A, C, G, T), and columns correspond to the predicted nucleotides. The color scale on the right indicates the number of observations in each cell, with darker colors representing higher counts.

(SD = 0.24%) and a macro-averaged F1 score of 0.831. On simulated data, the model showed excellent discrimination for ultra-rare variants (F1 = 0.985) and common variants (F1 = 0.864). while intermediate frequency classes remained more difficult to distinguish, reflecting the inherent challenge of separating adjacent frequency categories.

To provide a clearer understanding of the classification, the related confusion matrix is presented in Figure 8.

The confusion matrix shows that, for the COMMON_VARIANT class, 872 predictions were correct, with 13 instances misclassified as MODERATE_VARIANT. The MODERATE_VARIANT group had 3202 correct predictions, with 38 misclassified as RARE_VARIANT. The RARE_VARIANT category recorded 13 885 correct predictions, with 311 misclassified as VERY_RARE_VARIANT and 105 as ULTRA_RARE_VARIANT. VERY_RARE_VARIANT was correctly predicted in 14 615 instances, with 971 predicted as RARE_VARIANT and 816 as ULTRA_RARE_VARIANT. Finally, the ULTRA_RARE_VARIANT class had 3172 correct predictions, with 1189 misclassified as VERY_RARE_VARIANT.

Taken together, the two validation studies on simulated data showed low cross-validation variability (CV < 3%) and strong out-of-bag agreement (OOB error < 0.1%), indicating minimal overfitting and support the stability of the models. The close convergence between performances obtained on simulated and real data confirms that the models capture genuine biological signals and can be reliably applied to genomic data.

Overall, the PathoAnalyzer-I approach, through the selection of Option 5, enables the identification of novel biomarkers associated with chronic diseases. For Alzheimer's disease, the tool predicted 2593 SNPs, including rs12916T, rs6859219-A, rs10992900-G, rs192074881-A, rs6498168T, rs1269326T, rs9288520C, rs7671115T, rs7849973T, and

Table 5. Performance of the SNP variant classification model.

Class	Precision	Recall	F1 score	Support
COMMON_VARIANT	0.904	0.979	0.940	891
VARIANT	0.900	0.977	0.937	3278
RARE_VARIANT	0.910	0.957	0.933	14510
VERY_RARE_VARIANT	0.906	0.883	0.894	16560
ULTRA_RARE_VARIANT	0.774	0.665	0.715	4771

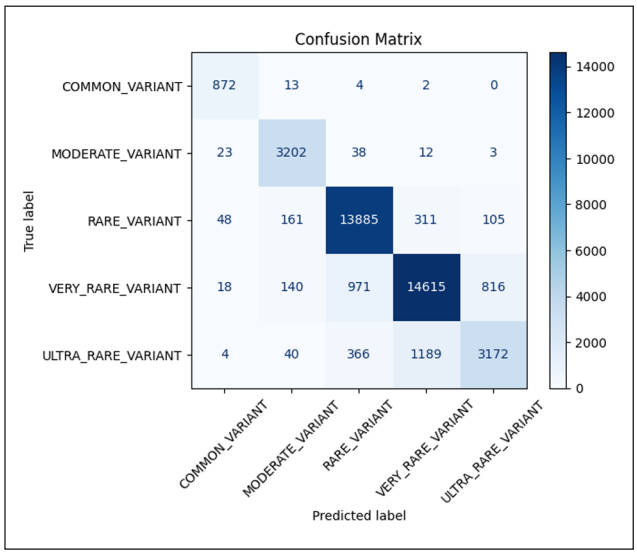

Figure 8. Confusion matrix of the SNP–pathology association prediction model.
The figure presents the confusion matrix of the Random Forest model developed to predict new SNP–pathology associations. Rows represent the true variant frequency classes (COMMON_VARIANT, MODERATE_VARIANT, RARE_VARIANT, VERY_RARE_VARIANT, ULTRA_RARE_VARIANT) and columns represent the predicted classes. The color bar indicates the number of SNPs in each cell, with darker shades reflecting higher counts of predictions.

rs1044069T. For any predicted SNP, the program provides detailed information; for example, rs192074881-A is located on chromosome 9, within the Fanconi anemia complementation group C (*FANCC*) gene at position 95 292 358, and is a missense variant.

In addition, the tool predicts missing risk alleles for SNPs already confirmed to be associated with chronic diseases. For instance, the risk allele for rs1060743 (linked to Alzheimer's disease)[44] was previously unknown (rs1060743-?), but the model successfully predicted it as rs1060743T.

Discussion

Advantages and limitations

Our Python-based software provides an integrated platform for *in silico* analyses of chronic pathologies.

Designed to optimize computational workflows, it enables rapid data processing and offers an intuitive interface suitable for users with varying levels of expertise.

By combining key analytical modules within a single environment, the tool reduces the need for multiple external applications and supports efficient bioinformatics investigations of chronic disease mechanisms.

Ease of use. Our software is distinguished by its simplicity and ease of use, compatible with all levels of bioinformatics expertise, enabling seamless adoption by researchers. In this regard, Galaxy[45] is another tool known for its user-friendliness, offering an accessible platform for users without programming skills.

Similarly, integrated platforms like the cBio Cancer Genomics Portal (cBioPortal)[46] deliver comprehensive analytical capabilities through intuitive web interfaces. However, leveraging these functionalities typically necessitates navigating multi-layered menus and configuring multistep analytical workflows posing onboarding challenges for users lacking bioinformatics training.

In contrast, command-line frameworks such as Genome Analysis Toolkit (GATK)[47] demand in-depth knowledge of sequencing pipelines and configuration parameters, posing barriers for experimental biologists without computational backgrounds.

Ultimately, ease of use remains a core criterion for selecting *in silico* analysis tools.

Comprehensive functionality. Integrated bioinformatics platforms have emerged to address the complexity of multi-omics analyses, yet significant gaps persist in chronic disease research. For example, cBioPortal offers robust cancer genomics exploration but provides limited coverage of nononcological chronic conditions and therapeutic insights. In contrast, our software delivers comprehensive analyses of 531 carefully curated chronic pathologies (scalable through program utilization) through a specialized workflow.

DisGeNET[48] operates as a specialized knowledge base for gene-disease associations. While valuable, its framework lacks molecular interaction context essential for mechanistic insights. PathoAnalyzer-I bridges this gap by integrating protein interaction networks and drug discovery within a unified analytical environment.

Where platforms like Open Targets[44] uses multiple modules requiring users to navigate between different contexts, PathoAnalyzer-I offers a streamlined workflow that systematically connects genetic markers to protein networks and therapeutic candidates.

PathoAnalyzer-I's distinctive architecture integrates six core capabilities:

1. Pathology Characterization: Comprehensive description of chronic disease.
2. SNP Characteristics Analysis: Identification and prioritization of pathology-associated SNPs using population-aware scoring strategies, integrating statistical significance and sample size to ensure robust and clinically relevant SNP selection.
3. Multi-Level SNP Annotation:
   - Genomic localization and context
   - Risk Allele
   - Functional consequences

- Associated gene/protein details
- 3D protein structure visualization
- Tissue-specific expression patterns

4. Network-Based Mechanistic Analysis: Construction of gene interaction networks using adjustable interaction thresholds.
5. Therapeutic and Preventive Molecule Screening: Identification of bioactive compounds targeting pathology-associated genes and proteins.
6. AI-Enhanced SNP Prediction: Prediction of novel SNP–disease associations using machine learning.

In summary, while existing solutions offer valuable specialized capabilities, PathoAnalyzer-I provides an integrative framework specifically optimized for chronic disease research. By combining multi-source data integration with computational predictions in an intuitive interface, we accelerate the translation of genetic insights into actionable therapeutic strategies.

Predictive capability. PathoAnalyzer-I's two machine learning models demonstrated robust predictive performance and substantial potential to expand both the annotation of SNPs already linked to chronic diseases and the catalog of SNPs associated with these pathologies.

Before discussing the utility of these predictions, it is essential to first address their reliability. Indeed, the missing risk allele prediction model achieved robust discriminative accuracy across all four nucleotides, with an overall accuracy of 77.6%, and demonstrated balanced precision and recall across classes (Table 4). The confusion matrix (Figure 7) further confirms that most predictions were correctly assigned, with adenine (A) and thymine (T) exhibiting particularly high recall (5997 and 5511 correct predictions, respectively), indicating that the model effectively captures nucleotide-specific patterns.

As shown in figure 7, the higher recall for A and T can be partly attributed to the intrinsic nucleotide composition bias of the human genome, which is enriched in A/T bases (~59%) compared to C/G (~41%).[49] This compositional asymmetry biases the model toward A/T bases, explaining the comparatively higher prediction accuracy for these nucleotides. Misclassifications were primarily observed between similar nucleotide pairs (A/T and C/T), reflecting the current limitations of the machine learning model in distinguishing nucleotides with closely related profiles, and suggesting potential avenues for improving feature representation and reducing recurrent errors.

Thanks to the robust accuracy of risk-allele predictions, numerous SNPs that were previously only partially annotated can now be meaningfully exploited. Knowing that an SNP lacking a risk allele is merely a genomic location and cannot be used in risk-allele–dependent analyses, such as those assessing genetic predisposition, this model substantially expands the pool of SNPs available for such analyses by providing accurate risk-allele predictions.

The second machine learning model of PathoAnalyzer-I, specialized in predicting novel SNP–chronic pathology associations, achieved an overall accuracy of 89.3% and demonstrates high precision and recall across all classes (Table 5). The stratification of SNPs by occurrence frequency directly contributed to these results. The confusion matrix (Figure 8) shows that predictions for SNP–pathology associations, relying on SNPs with the closest occurrence frequencies, are more prone to ambiguity due to subtle differences in SNP frequency, which limits the discriminative resolution of the model. Despite these constraints, these results demonstrate the model's robustness across all classes and ensure that the predictions remain reliable for downstream analyses of chronic diseases.

The predictions generated by the second model can be transformative on several levels.

First, rare chronic pathologies are often characterized by a limited number of validated biomarkers,[50] particularly SNPs, creating substantial ambiguity and delaying progress in their characterization. With PathoAnalyzer-I, however, the set of SNPs associated with these rare diseases becomes sufficiently enriched to support deeper biological insights and accelerate research on their molecular mechanisms.

Second, many chronic diseases are defined by silent biomarkers variants that show no clear involvement in disease pathways. Here, the predictive power of PathoAnalyzer-I is particularly valuable, as it uncovers novel SNP–disease associations that open entirely new avenues for scientific inquiry.

Combined, the predictions from both models, annotation and biomarker identification, provide the scientific community with powerful resources to advance the understanding of chronic diseases and ultimately address their incurability.

Current limitations and challenges of PathoAnalyzer-I. Like most modern bioinformatics platforms, PathoAnalyzer-I leverages distributed data resources to deliver comprehensive analyses. While this architecture ensures exceptional data freshness, it inherently creates dependencies on several external databases, such as the GWAS Catalog, PDB, and UniProt. As a result, when external APIs are temporarily unavailable, certain analytical modules may become inaccessible. To address this, PathoAnalyzer-I incorporates a resilience strategy: Intelligent Caching, whereby frequently accessed datasets (e.g. GTEx expression profiles) are automatically stored locally to reduce external query loads and improve responsiveness. Prospects remain for definitively addressing this limitation in future developments.

Another limitation is that predictive results are not available for pathologies outside the initial dataset, which comprises 531 chronic diseases, as predictions have been precomputed exclusively for this reference set. This limitation is partially mitigated by the fact that this set encompasses a wide range of chronic pathologies. Extending prediction capabilities to additional or newly identified diseases remains a priority for future releases, with plans to integrate real-time prediction modules and further enhance model performance.

The predicted SNPs (Option 5) are probabilistic and statistical in nature and do not represent experimentally validated biological or clinical findings. They should be

considered as indicators that depend on the study and population context rather than fixed biological features. In addition, the same allele may be pathogenic in one population and neutral in another. This further confirms the need for *in vitro* validation to move from statistical inference to biological confirmation.

Furthermore, although PathoAnalyzer-I applies a dual filtering step to candidate therapeutic molecules based on safety and Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) criteria, the safety assessment has certain limitations. While the current list of hazardous GHS codes captures the most critical hazards, further enrichment is required to cover a broader range of safety concerns. In addition, the clinical relevance of these compounds remains to be established. Further *in silico* validation (e.g. molecular docking and molecular dynamics simulations), followed by *in vitro* experiments, is required before any biological interpretation or translational application can be considered.

Comparison with existing SNP tools

PathoAnalyzer-I is a Python-based platform for the analysis of chronic diseases, integrating early diagnosis, therapeutic guidance, and signaling pathway investigation. Unlike established online version of tools such as SIFT,[51] PhD-SNP,[52] and I-Mutant 2.0,[53] which focus primarily on the functional or structural consequences of non-synonymous variants, PathoAnalyzer-I implements a disease-centered integrative framework. Specifically, SIFT evaluates amino acid substitutions based on sequence conservation, PhD-SNP predicts the likelihood of an SNP being disease-associated with an accuracy of approximately 74%, and I-Mutant 2.0 estimates mutation-induced protein stability changes, reporting accuracies of 80% and 77% depending on input type. In contrast, PathoAnalyzer-I combines experimentally validated SNP–disease associations from the GWAS Catalog with predictions generated by a machine learning model achieving an accuracy of 89.3%, surpassing the reported performance of PhD-SNP. The platform considers both intragenic and intergenic SNPs, supports gene network analysis, identifies relevant signaling pathways, and proposes potential therapeutic molecules. While a direct comparison with the aforementioned tools is limited by differences in objectives and input data, PathoAnalyzer-I demonstrates superior predictive performance, broader genomic coverage, and a more disease-oriented framework. As a result, it provides a systemic, multi-level view of chronic diseases and enables direct SNP–disease link, offering a more comprehensive and relevant solution for their study.

Deciphering chronic disease

After presenting the structure and functionalities of PathoAnalyzer-I in detail, a central question arises: *Does this program truly enable the deciphering of chronic diseases?*

In the context of *in silico* approaches, "deciphering" encompasses several dimensions: pre-diagnosis and diagnosis, understanding the molecular mechanisms underlying a disease, and identifying therapeutic and preventive avenues. To answer this question, we examine the case of Alzheimer's disease as a concrete example to demonstrate the capacity of PathoAnalyzer-I to decipher chronic pathologies.

Prediagnosis and risk estimation. PathoAnalyzer-I identifies genetic biomarkers (SNPs) that can be exploited to detect genetic predisposition to a disease before the onset of clinical symptoms. This information paves the way for preventive strategies, particularly through the control of gene expression in SNP-carrying genes. Nutritional interventions could achieve this, representing a promising avenue in preventive epigenetics. Unlike simple detection methods, the tool calculates the SNP–pathology association score by taking into account sample size, *P*-value, and population.

For example, in Alzheimer's disease, the top 5 SNPs among the 175 meeting the criteria, ranked in descending order, are: rs6733839T, rs12590654-A, rs11771145-A, rs12151021-A, and rs73223431T.

Importantly, probable genetic risk variants converge on key genes, including *BIN1*, *NIFKP9* (rs6733839T), *SLC24A4* (rs12590654-A), *EPHA1-AS1* (rs11771145-A), *ABCA7* (rs12151021-A), and *PTK2B* (rs73223431T) for which multiple studies have consistently reported significant associations with Alzheimer's disease.[54-57]

Finally, several studies highlight the importance of SNPs as early biomarkers of chronic diseases.[58-60] In parallel, numerous works emphasize gene expression control as a preventive epigenetic strategy.[61,62]

Confirmed diagnosis. These biomarkers can also serve in confirmed diagnosis among individuals already presenting clinical symptoms. Based on PathoAnalyzer-I, a subject displaying neurological symptoms in addition to the combination of SNPs (e.g. rs6733839T, rs12590654-A, rs11771145-A, rs12151021-A, and rs73223431T) can be reliably identified as an Alzheimer's patient. This diagnostic approach is supported by several studies, such as that of Zhang et al.[63]

Understanding pathological mechanisms. Cellular signaling is a cornerstone for deciphering pathological mechanisms and identifying metabolites implicated in disease processes. However, in chronic pathologies, well-established signaling pathways and gene interaction networks are often lacking, particularly in Alzheimer's disease.[64]

PathoAnalyzer-I overcomes this limitation by generating a gene interaction network for Alzheimer's disease (Figure 6). This network highlights 11 central genes *BIN1*, *SLC24A4*, *ABCA7*, *PTK2B*, *CLU*, *APOE*, *GRB2*, *RIN3*, *NME8*, *ZCWPW1*, *CD2AP*.

On the other hand, the currently established mechanism of Alzheimer's disease involves the amyloid precursor protein (*APP*), which undergoes amyloidogenic cleavage by β- and γ-secretases, producing amyloid-β (Aβ) peptides. These peptides progressively accumulate in the extracellular

neuronal space, forming toxic aggregates known as amyloid plaques. Concurrently, tau proteins undergo abnormal phosphorylation, leading to neurofibrillary tangles inside neurons. Together, plaques and tangles trigger early synaptic dysfunction, impairing neuronal communication, and initiate chronic neuroinflammation and oxidative stress. These disturbances ultimately converge on neuronal apoptosis, which underpins the severe cognitive and behavioral symptoms of this chronic disease.[65-67]

Considering this, the central genes identified by PathoAnalyzer-I play pivotal roles in Alzheimer's pathogenesis and should be taken into account when interpreting the disease mechanism:

Amyloid metabolism, lipid transport, and immune clearance. *APOE*, *CLU*, and *ABCA7* regulate lipid homeostasis and amyloid-β (Aβ) dynamics in the brain. The *APOE ε4* allele reduces Aβ clearance and promotes aggregation, while *CLU* acts as an extracellular chaperone controlling Aβ solubility and deposition. *ABCA7* contributes to microglial phagocytosis of Aβ and modulates inflammatory responses, thereby linking amyloid accumulation to immune activation.[68-70]

Endosomal trafficking and APP processing. *BIN1*, *CD2AP*, and *RIN3* orchestrate endocytic trafficking, APP internalization, and vesicular recycling. Dysfunction of this pathway represents an early pathogenic event in Alzheimer's disease, favoring amyloidogenic processing and endosomal stress, which precede plaque formation and synaptic impairment.[71-73]

Synaptic signaling, calcium homeostasis, and tau-related degeneration. *PTK2B* and *GRB2* regulate intracellular signaling pathways essential for synaptic plasticity, neuronal survival, and stress responses, while *SLC24A4* maintains calcium homeostasis. In Alzheimer's disease, *GRB2* modulates ERK1/2 signaling and helps protect the neuronal cytoskeleton against β-amyloid–induced damage. Dysregulation of these mechanisms may promote synaptic failure, calcium overload, tau hyperphosphorylation, and cytoskeletal destabilization.[74-77]

Structural integrity and transcriptional vulnerability. *NME8* and *ZCWPW1* contribute to cytoskeletal organization and epigenetic regulation of neuronal gene expression.[78] Alterations in these processes may reduce neuronal resilience and amplify vulnerability to neurodegeneration, reinforcing the progressive nature of Alzheimer's disease.

These mechanisms illustrate that Alzheimer's pathogenesis emerges from the convergence of amyloid dysregulation, endosomal trafficking failure, immune activation, synaptic dysfunction, and transcriptional vulnerability, each modulated by the key genes *APOE*, *CLU*, *ABCA7*, *BIN1*, *CD2AP*, *RIN3*, *PTK2B*, *GRB2*, *SLC24A4*, *NME8*, and *ZCWPW1*. Genetic predisposition, in combination with age-related and environmental risk factors, perturbs the expression and function of these interconnected genes, thereby accelerating pathological onset and promoting progressive neuronal degeneration.

Therapeutic molecules. To date, no curative treatments exist for Alzheimer's disease. Current strategies aim to alleviate symptoms and slow disease progression.[79,80] However, the identification of effective therapeutic strategies is closely linked to a comprehensive understanding of the gene networks and signaling pathways underlying chronic diseases. To achieve this, we must identify regulators for the 11 genes that play a central role in the cellular signaling pathway proposed by PathoAnalyzer-I, as already described—*BIN1*, *SLC24A4, ABCA7, PTK2B, CLU, APOE, GRB2, RIN3, NME8, ZCWPW1, CD2AP*—PathoAnalyzer-I proposed gene regulators after a double filtering step (Safety and ADMET), which made it possible to retain potential therapeutic molecules, some of which are summarized in Table 6.

Some of these molecules, such as curcumin (from turmeric), deguelin (from certain legumes like Derris spp.), and protocatechuic acid (from fruits and vegetables), are of dietary origin, offering preventive potential in addition to therapeutic perspectives.

Another important point concerns Beta-Lapachone, which regulates two key genes (*APOE* and *RIN3*) in the identified networks, highlighting a strong potential for therapeutic involvement. Moreover, some studies confirm its therapeutic relevance to Alzheimer's disease.[81,82]

The case study of Alzheimer's disease demonstrates the ability of PathoAnalyzer-I to decipher chronic diseases. The program provides an integrated framework for comprehensive analysis of chronic pathologies, combining molecular-data-driven insights with predictive approaches to support research and translational applications. By uniting multi-source data integration with systematic analyses, the software enables the development of targeted strategies for understanding and managing chronic diseases.

Conclusion

PathoAnalyzer-I is a comprehensive bioinformatics platform designed for the in-depth analysis of chronic pathologies. It integrates large-scale genetic and molecular datasets with advanced machine learning models, enabling precise biomarker annotation, prediction of risk alleles, identification of causal molecular mechanisms, and suggestions of potential therapeutic and preventive interventions. By combining these functionalities within a single environment, the tool provides researchers with rapid, reliable, and holistic insights into disease biology.

The case of Alzheimer's disease illustrates the capabilities of PathoAnalyzer-I. The platform identified key SNPs, such as rs6733839T, for early risk assessment and confirmed diagnosis, and highlighted central genes including *BIN1*, *APOE*, *SLC24A4*, *ABCA7*, and *PTK2B*, which are involved in amyloid processing, synaptic dysfunction, and cellular vulnerability. Moreover, it proposed therapeutic molecules, such as Beta-Lapachone, targeting

Table 6. Selected therapeutic modulators of central genes (*BIN1, SLC24A4, ABCA7, PTK2B, CLU, APOE, GRB2, RIN3, NME8, ZCWPW1, CD2AP*) identified by PathoAnalyzer-I in the context of Alzheimer's disease.

Gene	Stimulators	Inhibitors
BIN1		2-(Biphenyl-4-ylsulfonyl) N-hydroxybenzamide
SLC24A4	2-(3-Trifluoromethylphenyl)histamine	
ABCA7	-	- 2(R)-Hydroxy-2-methylbutyronitrile-beta-D-glucopyranoside
PTK2B	Benzo(a)pyrene-7,8-dione	Clofibrate
APOE	Beta-Lapachone, Beta-Naphthoflavone	3-(4-Dimethylaminobenzylidenyl)-2-indolinone
CLU	2,5-Dimethoxy-4-iodoamphetamine	2,6-Dimethoxyphenol
GRB2	Protocatechuic acid	Curcumin, Deguelin
RIN3	Beta-Lapachone	4-Methylpyrazole, Ocaperidone
NME8		
ZCWPW1	Cgp-52608	
CD2AP	4-(6-(4-(1-Piperazinyl)phenyl)pyrazolo (1,5-a)pyrimidin-3-yl)quinoline	

theses central pathways and dietary compounds, including curcumin, offering preventive benefits through modulation of gene expression. This example demonstrates the platform's ability to connect genetic data, molecular mechanisms, and intervention strategies in a coherent framework.

Overall, the tool provides the scientific community with a powerful resource to advance the understanding and management of chronic diseases. The platform facilitates comprehensive research into chronic pathologies, ultimately contributing to strategies that improve disease management and public health outcomes.

ORCID iD

Ali Aguerd https://orcid.org/0009-0004-6819-2130

Ethical Considerations

This study exclusively used publicly available molecular and genomic datasets and did not involve human participants, human tissue, or animals. Consequently, no ethical approval or consent was required.

Author Contributions

Ali Aguerd: Conceptualization; Methodology; Software; Writing—original draft; Data curation.
Faïza Bennis: Writing—review & editing; Supervision; Validation.
Fatima Chegdani: Validation; Writing—review & editing; Supervision; Project administration.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The Python code, datasets, and user guide for the PathoAnalyzer-I platform are publicly available in the Zenodo repository at: https://doi.org/10.5281/zenodo.19031470

The materials have been deposited as a 7-Zip archive (PathoAnalyzer-I.7z) containing the following components:

- **deciphering_incurable_pathologies** – Core Python scripts required to run the PathoAnalyzer-I platform.
- **requirements.txt** – List of Python libraries required to install and execute the program.
- **Prediction** – Scripts used to build and apply machine-learning models for genetic prediction tasks:
  - **RiskAllele-Prediction** – model training and prediction of risk alleles.
  - **SNP-Prediction –** model training and prediction of SNPs associated with diseases.
- **Validation** – Two Python scripts used to validate the performance of the prediction models.
- **dataset** – TSV files with curated data on 531 pathologies.
- **User Guide** – A Word document describing installation steps and usage of the platform.

Project specifications

Operating systems: Platform independent (Windows, macOS, Linux)

Programming language: Python

Best practices

To ensure optimal use of PathoAnalyzer-I, users should:

Install Python and the libraries listed in *requirements.txt*.

Run the *main.py* script located in the *deciphering_incurable_pathologies* folder.

Use English scientific terminology for pathologies, and preferably adopt the nomenclature provided by the GWAS Catalog to ensure consistency and accurate interoperability with both local and external databases.

Maintain a stable Internet connection for reliable and rapid access to external databases.

When using an integrated development environment (IDE) such as PyCharm, a dark theme is recommended, as expressions

displayed in light colors may be difficult to read on a light background.

Supplemental Material

Supplemental material for this article is available online.

References

1. Watkins D, Ahmed S, Pickersgill S. Best investments in chronic, noncommunicable disease prevention and control in low- and lower–middle-income countries. *J Benefit-cost Anal*. 2023;14:255-271. doi:10.1017/bca.2023.25
2. Non communicable diseases. Accessed March 29, 2025. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
3. Hacker K. The burden of chronic disease. *Mayo Clin Proc Innov Qual Outcomes*. 2024;8:112-119. doi:10.1016/j.mayocpiqo.2023.08.005
4. Cvijovic M, Polster A. Network medicine: facilitating a new view on complex diseases. *Front Bioinform*. 2023;3:1163445. doi:10.3389/fbinf.2023.1163445
5. Baião AR, Cai Z, Poulos RC, et al. A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. *Brief Bioinform*. 2025;26:bbaf355. doi:10.1093/bib/bbaf355.
6. Papadaki E, Kakkos I, Vlamos P, et al. Recent web platforms for multi-omics integration unlocking biological complexity. *Appl Sci*. 2024;15:329. doi:10.3390/app15010329
7. Arbitrio M, Milano M, Lucibello M, et al. Bioinformatic challenges for pharmacogenomic study: tools for genomic data analysis. *Front Pharmacol*. 2025;16:1548991. doi:10.3389/fphar.2025.1548991
8. Trost B, Loureiro LO, Scherer SW. Discovery of genomic variation across a generation. *Hum Mol Genet*. 2021;30:R174-R186. doi:10.1093/hmg/ddab209
9. Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25:1422-1423. doi:10.1093/bioinformatics/btp163
10. McKinney W. *Data Structures for Statistical Computing in Python*. SciPy; 2010.
11. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32. doi:10.1023/A
12. Cerezo M, Sollis E, Ji Y, et al. The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Res*. 2025;53:D998-D1005. doi:10.1093/nar/gkae1070
13. Dyer SC, Austine-Orimoloye O, Azov AG, et al. Ensembl 2025. *Nucleic Acids Res*. 2025;53:D948-D957. doi:10.1093/nar/gkae1071
14. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585:357-362. doi:10.1038/s41586-020-2649-2
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
16. Category Encoders — Category encoders 2.8.1 documentation. Accessed July 03, 2025. https://contrib.scikit-learn.org/category_encoders/
17. Imbalanced-learn documentation — Version 0.13.0. Accessed July 03, 2025. https://imbalanced-learn.org/stable/
18. Waskom M. Seaborn: statistical data visualization. *J Open Source Softw*. 2021;6:3021. doi:10.21105/joss.03021
19. Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci Eng*. 2007;9:90-95. doi:10.1109/MCSE.2007.55
20. Python documentation. Os — miscellaneous operating system interfaces. Accessed July 03, 2025. https://docs.python.org/3/library/os.html
21. Python documentation. Datetime — basic date time types. Accessed July 03, 2025. https://docs.python.org/3/library/datetime.html
22. Python documentation. Logging — logging facility for Python. Accessed July 03, 2025. https://docs.python.org/3/library/logging.html
23. Python documentation. sys — system-specific parameters functions. Accessed July 03, 2025. https://docs.python.org/3/library/sys.html
24. Python documentation. hashlib — secure hashes message digests. Accessed July 03, 2025. https://docs.python.org/3/library/hashlib.html
25. Python documentation. warnings — warning control. Accessed July 03, 2025. https://docs.python.org/3/library/warnings.html
26. Python documentation. collections — container datatypes. Accessed July 03, 2025. https://docs.python.org/3/library/collections.html
27. Joblib: running Python functions as pipeline jobs — joblib 1.5.1 documentation. Accessed July 03, 2025. https://joblib.readthedocs.io/en/stable/
28. McLaughlin J, Lagrimas J, Iqbal H, Parkinson H, Harmse H. OLS4: a new ontology lookup service for a growing interdisciplinary knowledge ecosystem. *Bioinformatics*. 2025;41:btaf279. doi:10.1093/bioinformatics/btaf279
29. Burley SK, Bhatt R, Bhikadiya C, et al. Updated resources for exploring experimentally-determined PDB structures and computed structure models at the RCSB protein data bank. *Nucleic Acids Res*. 2025;53:D564-D574. doi:10.1093/nar/gkae1091
30. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res*. 2025;53:D609-D617. doi:10.1093/nar/gkae1010
31. Szklarczyk D, Nastou K, Koutrouli M, et al. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res*. 2025;53:D730-D737. doi:10.1093/nar/gkae1113
32. Kim S, Chen J, Cheng T, et al. PubChem 2025 update. *Nucleic Acids Res*. 2025;53:D1516-D1525. doi:10.1093/nar/gkae1059
33. The GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369:1318-1330. doi:10.1126/science.aaz1776
34. Python Prompt Toolkit 3.0 — prompt_toolkit 3.0.50 documentation. Accessed July 03, 2025. https://python-prompt-toolkit.readthedocs.io/en/master/
35. Python. Prettytable: a simple Python library for easily displaying tabular data in a visually appealing ASCII table format. Accessed July 03, 2025. https://github.com/prettytable/prettytable
36. Python. Colorama: cross-platform colored terminal text [OS Independent]. Accessed July 03, 2025. https://github.com/tartley/colorama
37. Requests: HTTP for Humans™ — requests 2324 documentation. Accessed July 03, 2025. https://requests.readthedocs.io/en/latest/
38. Selenium with Python — selenium Python bindings 2 documentation. Accessed July 03, 2025. https://selenium-python.readthedocs.io/

39. Pirogov S. Python. Webdriver -manager: library provides the way to automatically manage drivers for different browsers [MacOS, Microsoft: Windows, POSIX, Unix]. Accessed July 03, 2025. https://github.com/SergeyPirogov/webdriver_manager
40. Pillow. Pillow (PIL Fork). Accessed July 03, 2025. https://pillow.readthedocs.io/en/stable/index.html
41. Python. SeatGeek. Seatgeek/fuzzywuzzy. June 30, 2025. Accessed July 03, 2025. https://github.com/seatgeek/fuzzywuzzy
42. Plotly. Accessed July 03, 2025. https://plotly.com/python/
43. Hagberg AA, Schult DA, Swart PJ. *Exploring Network Structure, Dynamics, and Function using NetworkX*. SciPy; 2008.
44. Ochoa D, Hercules A, Carmona M, et al. Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res*. 2021;49:D1302-D1310. doi:10.1093/nar/gkaa1027
45. Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res*. 2024;52:W83-W94. doi:10.1093/nar/gkae410
46. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2:401-404. doi:10.1158/2159-8290.CD-12-0095
47. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11101-11111. doi:10.1002/0471250953.bi1110s43
48. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2020;48:D845-D855. doi:10.1093/nar/gkz1021
49. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409:860-921. doi:10.1038/35057062
50. Núñez-Carpintero I, Rigau M, Bosio M, et al. Rare disease research workflow using multilayer networks elucidates the molecular determinants of severity in congenital myasthenic syndromes. *Nat Commun*. 2024;15:1227. doi:10.1038/s41467-024-45099-0
51. Ng PC. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2013;31:3812-3814. doi:10.1093/nar/gkg509
52. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*. 2006;22:2729-2734. doi:10.1093/bioinformatics/btl423
53. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*. 2005;33:W306-W310. doi:10.1093/nar/gki375
54. Rajabli F, Benchek P, Tosto G, et al. Multi-ancestry genome-wide meta-analysis of 56,241 individuals identifies known and novel cross-population and ancestry-specific associations as novel risk loci for Alzheimer's disease. *Genome Biol*. 2025;26:210. doi:10.1186/s13059-025-03564-z
55. Sanders KL, Manuel AM, Liu A, Leng B, Chen X, Zhao Z. Unveiling gene interactions in Alzheimer's disease by integrating genetic and epigenetic data with a network-based approach. *Epigenomes*. 2024;8:14. doi:10.3390/epigenomes8020014
56. Tan MS, Yang YX, Xu W, et al; Alzheimer's Disease Neuroimaging Initiative. Associations of Alzheimer's disease risk variants with gene expression, amyloidosis, tauopathy, and neurodegeneration. *Alzheimers Res Ther*. 2021;13:15. doi:10.1186/s13195-020-00755-7
57. Guo Y, Sun CK, Tang L, Tan MS. Microglia PTK2B/Pyk2 in the pathogenesis of Alzheimer's disease. *Curr Alzheimer Res*. 2023;20:692-704. doi:10.2174/1115672050299004240129051655
58. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nat Rev Methods Primer*. 2021;1:1-21. doi:10.1038/s43586-021-00056-9
59. Fujita M, Liu X, Iwasaki Y, et al. Population-based screening for hereditary colorectal cancer variants in Japan. *Clin Gastroenterol Hepatol*. 2022;20:2132-2141. doi:10.1016/j.cgh.2020.12.007
60. Ahmed H, Soliman H, Elmogy M. Early detection of Alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree. *Comput Biol Med*. 2022;146:105622. doi:10.1016/j.compbiomed.2022.105622
61. Abdul QA, Yu BP, Chung HY, Jung HA, Choi JS. Epigenetic modifications of gene expression by lifestyle and environment. *Arch Pharm Res*. 2017;40:1219-1237. doi:10.1007/s12272-017-0973-3
62. Agrawal P, Kaur J, Singh J, et al. Genetics, nutrition, and health: a new frontier in disease prevention. *J Am Nutr Assoc*. 2024;43:326-338. doi:10.1080/27697061.2023.2284997
63. Zhang C, Zheng T, Ma Q, et al. Logical analysis of multiple single-nucleotide-polymorphisms with programmable DNA molecular computation for clinical diagnostics. *Angew Chem Int Ed*. 2022;61:e202117658. doi:10.1002/anie.202117658
64. Liu E, Zhang Y, Wang J-Z. Updates in Alzheimer's disease: from basic research to diagnosis and therapies. *Transl Neurodegener*. 2024;13:45. doi:10.1186/s40035-024-00432-x
65. Sharma A, Rudrawar S, Bharate SB, Jadhav HR. Recent advancements in the therapeutic approaches for Alzheimer's disease treatment: current and future perspective. *RSC Med Chem*. 2025;16:652-693. doi:10.1039/D4MD00630E
66. Kashif M, Sivaprakasam P, Vijendra P, Waseem M, Pandurangan AK. A recent update on pathophysiology and therapeutic interventions of Alzheimer's disease. *Curr Pharm Des*. 2023;29:3428-3441. doi:10.2174/0113816128264355231121064704
67. Alkhalifa AE, Al Mokhlf A, Ali H, Al-Ghraiybah NF, Syropoulou V. Anti-amyloid monoclonal antibodies for Alzheimer's disease: evidence, ARIA risk, and precision patient selection. *J Pers Med*. 2025;15:437. doi:10.3390/jpm15090437
68. Chen Y, Jin H, Chen J, Li J, Găman M-A, Zou Z. The multifaceted roles of apolipoprotein E4 in Alzheimer's disease pathology and potential therapeutic strategies. *Cell Death Discov*. 2025;11:312. doi:10.1038/s41420-025-02600-y
69. Yu Y, Wang C, Wang B, et al. Clusterin regulates the mechanisms of neuroinflammation and neuronal circuit impairment in Alzheimer's disease. *Int J Mol Sci*. 2025;26:7271. doi:10.3390/ijms26157271
70. Santos-García I, Bascuñana P, Brackhan M, et al. The ABC transporter A7 modulates neuroinflammation via NLRP3 inflammasome in Alzheimer's disease mice. *Alzheimers Res Ther*. 2025;17:30. doi:10.1186/s13195-025-01673-2

71. Saha T, Karati D. A comprehensive review on novel opportunities for Alzheimer therapy by targeting BIN1. *Egypt J Med Hum Genet*. 2025;26:134. doi:10.1186/s43042-025-00759-8
72. Vandal M, Janmaleki M, Rea I, et al. CD2AP at the junction of nephropathy and Alzheimer's disease. *Mol Neurodegener*. 2025;20:63. doi:10.1186/s13024-025-00852-x
73. Meshref M, Ghaith HS, Hammad MA, et al. The role of RIN3 gene in Alzheimer's disease pathogenesis: a comprehensive review. *Mol Neurobiol*. 2024;61:3528-3544. doi:10.1007/s12035-023-03802-0
74. Astillero-Lopez V, Villar-Conde S, Gonzalez-Rodriguez M, et al. Proteomic analysis identifies HSP90AA1, PTK2B, and ANXA2 in the human entorhinal cortex in Alzheimer's disease: potential role in synaptic homeostasis and Aβ pathology through microglial and astroglial cells. *Brain Pathol*. 2024;34:e13235. doi:10.1111/bpa.13235
75. Nizzari M, Venezia V, Repetto E, et al. Amyloid precursor protein and presenilin1 interact with the adaptor GRB2 and modulate ERK 1,2 signaling. *J Biol Chem*. 2007;282:13833-13844. doi:10.1074/jbc.M610146200
76. Majumder P, Roy K, Singh BK, Jana NR, Mukhopadhyay D. Cellular levels of growth factor receptor bound protein 2 (Grb2) and cytoskeleton stability are correlated in a neurodegenerative scenario. *Dis Model Mech*. 2017;10:655-669. doi:10.1242/dmm.027748
77. Akbor MM, Kim J, Nomura M, Sugioka J, Kurosawa N, Isobe M. A candidate gene of Alzheimer diseases was mutated in senescence-accelerated mouse prone (SAMP) 8 mice. *Biochem Biophys Res Commun*. 2021;572:112-117. doi:10.1016/j.bbrc.2021.07.095
78. Giri M, Lü Y, Zhang M. Genes associated with Alzheimer's disease: an overview and current status. *Clin Interv Aging*. 2016;11:665-681. doi:10.2147/CIA.S105769
79. Rajanna S, Gundale PP, Dev Mahadevaiah A. Advancements in the treatment of Alzheimer's disease: a comprehensive review. *Dement Neuropsychol*. 2025;19:e20240204. doi:10.1590/1980-5764-DN-2024-0204
80. Passeri E, Elkhoury K, Morsink M, et al. Alzheimer's disease: treatment strategies and their limitations. *Int J Mol Sci*. 2022;23:13954. doi:10.3390/ijms232213954
81. Mokarizadeh N, Karimi P, Erfani M, Sadigh-Eteghad S, Fathi Maroufi N, Rashtchizadeh N. β-Lapachone attenuates cognitive impairment and neuroinflammation in beta-amyloid induced mouse model of Alzheimer's disease. *Int Immunopharmacol*. 2020;81:106300. doi:10.1016/j.intimp.2020.106300
82. Mokarizadeh N, Karimi P, Kazemzadeh H, et al. An evaluation on potential anti-inflammatory effects of β-lapachone. *Int Immunopharmacol*. 2020;87:106810. doi:10.1016/j.intimp.2020.106810