A Secure and Traceable Computational Evidence Platform for Drug–Biological System Analysis Using Machine Learning and Multi-Omics Data

Sangeetha Priya B.[1], Abirami M.[2], Janushiya J.[3], Sandhiya K.T.[4], Sankar S.[5], Saseedhar J.[6]
[1,2,3,4,5,6]*Department of Computer Science and Engineering, SNS College of Technology, Coimbatore, Tamil Nadu, India*

***Abstract*—Drug-candidate assessment is slow and expensive because molecular, biological and toxicity evidence is scattered across separate tools that rarely report model agreement, uncertainty or data integrity. This paper presents Cygen Vault, a secure and traceable computational evidence platform for drug–biological system analysis. The platform organizes drug molecular profiles and multi-omics biological response data, links drugs to targets, pathways and possible toxicity signals, and compares predictions from Random Forest, Support Vector Machine, Logistic Regression and XGBoost models. Instead of a binary safe/unsafe label, it reports risk signals, model consensus, confidence and uncertainty. An Evidence Passport records drug identity, inputs, data sources, model and dataset versions, results and timestamps together with cryptographic integrity values, while authentication, role-based access control, password hashing and audit logs protect the records. The workflow has ten stages and ends with prioritization of candidates for laboratory validation. Functional testing of the modules, the consensus logic and the integrity checks is described, and an evaluation protocol based on accuracy, precision, recall, F1-score and ROC-AUC is defined. Cygen Vault is a decision-support tool that supports, and does not replace, laboratory and clinical validation.**

***Index Terms*—Cybersecurity, drug discovery, evidence traceability, machine learning, multi-omics, toxicity prediction.**

I. INTRODUCTION

Early assessment of drug candidates requires researchers to examine molecular structure, biological targets, pathway effects and possible toxicity. Each of these steps normally relies on different datasets and tools, which makes the process expensive and slow. Machine learning has shown strong results for toxicity prediction, for example through deep neural networks trained on large chemical collections [1] and through community benchmarks such as the Tox21 Data Challenge [2]. In parallel, multi-omics data (genomic, transcriptomic, proteomic and metabolomic measurements) now describe how biological systems respond to a compound at several levels [3].

Despite this progress, most prediction tools return a score without a unified view of the supporting evidence, the agreement between models, the uncertainty of the result, or the integrity of the data that produced it. A result that cannot be traced back to its inputs, dataset version and model version is difficult to audit or reproduce. Sensitive research data also needs protection through access control and tamper detection. Recent work on securing data with blockchain-integrated encryption in the quantum era [4] and on distributed ledgers for biomedical data [5] shows the growing importance of verifiable data handling.

This paper presents Cygen Vault, a secure computational platform that connects these activities in one workflow. The main contributions are:

1) A structured representation of drug molecular profiles and species-specific biological response profiles.

2) A Drug → Target → Pathway → Biological Response → Possible Toxicity Signal analysis chain.
3) A multi-model prediction layer that reports consensus, confidence and uncertainty instead of a single safe/unsafe label.

4) An Evidence Passport with integrity verification, combined with authentication, role-based access control and audit logging.

Cygen Vault is intended as a decision-support concept for prioritizing candidates for laboratory investigation. It does not prove the safety, toxicity or clinical effectiveness of a drug.

II. RELATED WORK

A. Machine Learning for Toxicity Prediction

DeepTox showed that deep networks can outperform conventional methods on the Tox21 data [1], and the Tox21 challenge established public tasks covering nuclear receptor and stress response pathways [2]. Classical learners remain widely used on tabular molecular features: Random Forests [6], Support Vector Machines [7] and gradient-boosted trees [8]. Cygen Vault uses these learners as interchangeable members of an ensemble rather than relying on a single model.

B. Multi-Omics Evidence and Knowledge Bases

Multi-omics integration combines several molecular layers to explain disease and treatment response [3]. Curated resources such as DrugBank link drugs to targets, pathways and pharmacological data [9]. Cygen Vault organizes this type of evidence into a profile that is attached to each prediction.

C. Security, Traceability and AI-Enabled Sensing

Blockchain ledgers have been proposed to give biomedical and health records tamper evidence and traceability [5]. Related studies combine blockchain with steganographic encryption to protect data against quantum-era threats [4]. Role-based access control remains the standard model for restricting what each class of user may do with sensitive data [10]. Outside drug analysis, AI-enabled sensing platforms that join IoT data links with neural-network prediction of biomarkers have been reported [11]. Such systems also produce biological data that must be handled reliably. AI-based optimization is also being applied in other engineering domains, such as the design of solar absorbers for renewable energy systems [12], which shows the wider move towards data-driven design.

D. Gap Addressed

Existing work treats prediction accuracy, evidence integration and data security largely as separate problems. Cygen Vault combines them by attaching consensus, uncertainty and a verifiable record to every analysis.

III. PROPOSED PLATFORM

A. Overview and Workflow

Cygen Vault contains six modules that follow the ten-stage workflow in Fig. 1. A drug candidate is entered and converted into molecular features. Biological response data are added, targets and pathways are analyzed, and the evidence is passed to the prediction layer. The prediction layer produces consensus, risk and confidence values. The result is sealed in an Evidence Passport and the candidate is ranked for laboratory validation. The layered architecture of the platform is shown in Fig. 2.

B. Module 1: Drug Molecular Profile

This module stores the chemical structure, molecular formula, molecular weight, molecular descriptors, molecular fingerprints, SMILES/InChI representations and known biological targets. Structures are encoded as SMILES strings [13] and converted into circular fingerprints [14], which give a fixed-length numerical input for the learners.

C. Module 2: Biological Response Profile

This module organizes DNA, RNA, protein, metabolite, pathway, biomarker and cellular-response information. Human, animal, bird and plant contexts are stored separately so that results from one biological context are not compared directly with another.

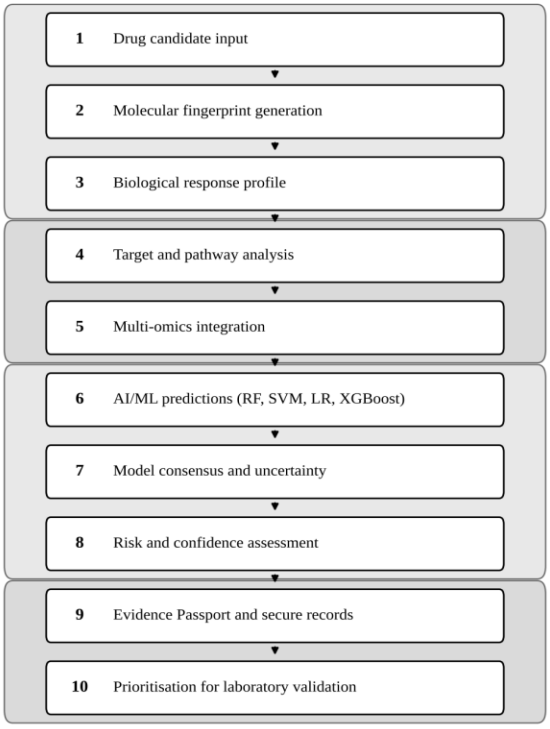

Fig. 1. Ten-stage Cygen Vault workflow. Shading groups the stages into input and profiling (1–3), biological analysis (4–5), prediction and assessment (6–8), and trust and output (9–10).

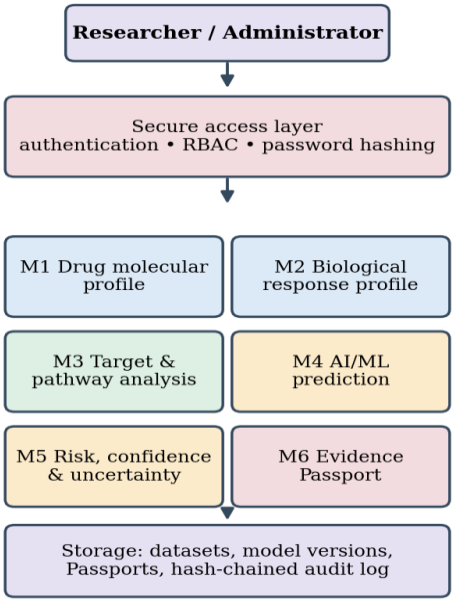

Fig. 2. Layered architecture of Cygen Vault: secure access layer, six analysis modules (M1–M6) and the storage layer.

D. Module 3: Target and Pathway Analysis

The module examines possible drug–target relationships and their effect on biological pathways, following the chain Drug → Target → Pathway → Biological Response → Possible Toxicity Signal. Known target annotations can be drawn from curated sources such as DrugBank [9]. The analysis chain is illustrated in Fig. 3.

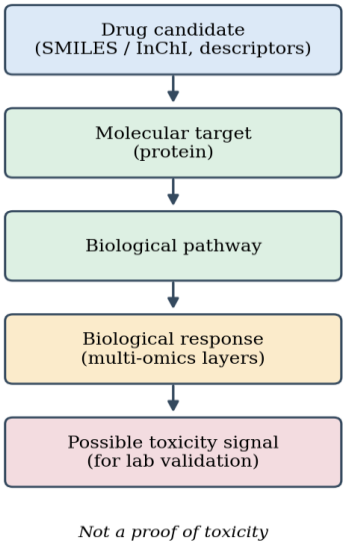

Fig. 3. Drug → Target → Pathway → Biological Response → Possible Toxicity Signal analysis chain.

E. Module 4: AI/ML Prediction

Four learners are applied to the molecular and biological features (Table I). Comparing several models allows the platform to measure how consistent their predictions are.

TABLE I Learners used in the prediction layer.

Model	Type	Role in platform
Random Forest [6]	Ensemble of decision trees	Non-linear baseline; feature importance
SVM [7]	Maximum-margin kernel classifier	Handles high-dimensional fingerprints
Logistic Regression	Linear probabilistic model	Interpretable baseline
XGBoost [8]	Gradient-boosted trees	High-accuracy learner for tabular features

F. Module 5: Risk, Confidence and Uncertainty

For a sample x, let M be the number of models, $p_m(x)$ the class-probability vector of model m, and $\hat{y}_m$ its predicted class. The majority class is $\hat{y}^*$. The consensus score and the predictive uncertainty are

$$C(x) = (1/M)\ \Sigma_m\ 1[\hat{y}_m = \hat{y}^*] \quad (1)$$

$$U(x) = -\Sigma_k\ \bar{p}_k \log_2 \bar{p}_k, \quad \bar{p} = (1/M)\ \Sigma_m\ p_m(x) \quad (2)$$

where $1[\cdot]$ equals 1 when the condition holds and 0 otherwise, and $\bar{p}_k$ is the mean probability of class k. A high C(x) with a low U(x) indicates a stable prediction. The platform reports these values with an evidence-strength indicator and a model-applicability note, so a drug is never labelled simply safe or unsafe. The usefulness of these indicators depends on validation against real experimental data. Fig. 4 illustrates the two typical cases.

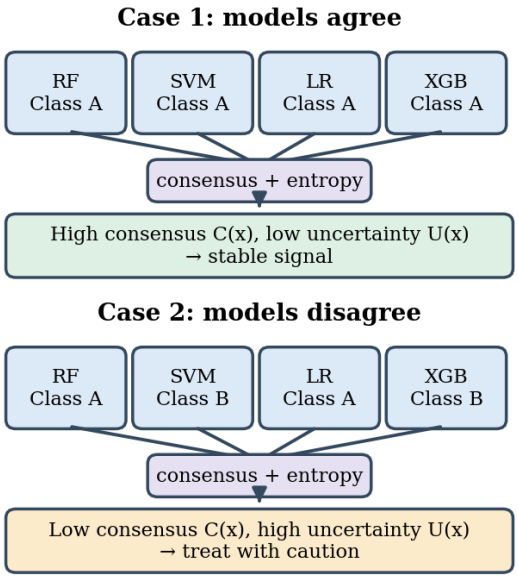

Fig. 4. Qualitative example of model consensus and uncertainty. The figure contains no measured values.

G. Data Preparation and Feature Pipeline

Before prediction, every input passes through a fixed preparation pipeline so that results are comparable and reproducible. Structures are standardized, duplicates are removed, and invalid SMILES strings are rejected with a logged reason. Molecular descriptors and fingerprints are computed, and numerical features are scaled where a learner requires it. Missing biological values are flagged and are not silently filled, because an imputed value can look like real evidence. Records are tagged with their biological context (human, animal, bird or plant), and the data are split into training and test sets within each context. Each preparation step is recorded with the version of the code that performed it, so the same input always produces the same feature vector.

The biological evidence is kept in layers that match the omics levels: DNA, RNA, protein, metabolite and pathway. A layer that has no data for a drug is marked as absent in the evidence profile. The prediction layer receives this absence flag as an input, and the final report states which layers supported the result.

IV. EVIDENCE PASSPORT AND SECURITY

A. Evidence Passport

The Evidence Passport (Module 6) is the traceable record of one analysis. It stores the drug identity, input data, data sources, model and dataset versions, results and timestamps. A digest of these fields is computed with SHA-256,

$$h = \text{SHA-256}(id \,\|\, D \,\|\, V \,\|\, R \,\|\, t) \quad (3)$$

where id is the drug identity, D the input data, V the model and dataset versions, R the results and t the timestamp. Any later change to the stored fields produces a different digest, so tampering is detected when the passport is verified.

Table II lists the fields held in one passport.

TABLE II Fields of the Evidence Passport.

Field	Content
Drug identity	Name, SMILES/InChI, internal identifier
Input data	Molecular profile and biological response profile used
Sources	Origin and version of each dataset or database
Model versions	Learner, hyperparameters, training-data version
Results	Predictions, consensus, uncertainty, risk and confidence
Integrity	SHA-256 digest and timestamp

B. Access Control and Audit

Users must authenticate before using the platform. Passwords are stored only as salted hashes. Role-based access control [10] limits each role, such as researcher or administrator, to the actions it needs. Every access and change is written to an audit log.

The log entries are linked so that each entry contains the hash of the previous one,

$$H_i = \text{SHA-256}(e_i \parallel H_{i-1}) \quad (4)$$

where e_i is the i-th log entry. This hash-chaining gives tamper evidence in the same way as blockchain-based record keeping [4], [5], without requiring a distributed network. The passport and the audit chain are summarized in Fig. 5. The digests could later be anchored to an external ledger if stronger guarantees are needed.

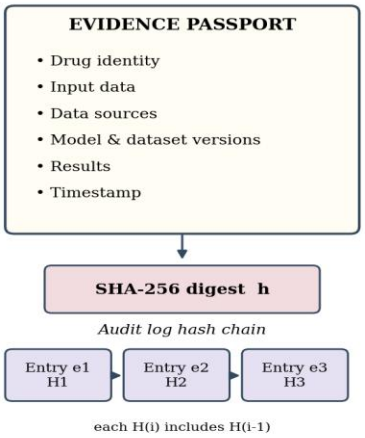

Fig. 5. Evidence Passport fields, SHA-256 digest and hash-chained audit log.

C. Threat Model

The security design addresses the risks that matter most for stored research evidence (Table III). The platform is aimed at a research group or institution, so the threats considered are accidental or deliberate misuse by users and tampering with stored records, not attacks on the network infrastructure. Fig. 6 maps each threat to its mitigation.

TABLE III Threats and mitigations.

Threat	Mitigation
Unauthorized access	Authentication and role-based access control
Altered results or inputs	Passport digest check (3)
Edited or deleted log entries	Hash-chained audit log (4)
Stolen credentials from storage	Salted password hashing
Results that cannot be reproduced	Dataset, model and code version records
Mixing of species contexts	Context tags and per-context splits

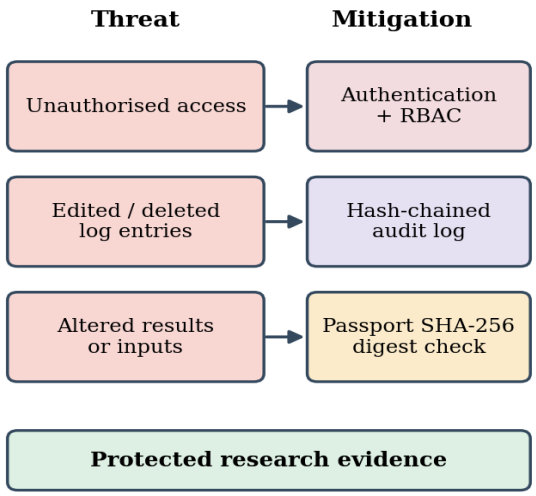

Fig. 6. Threats considered and the matching security mitigation.

V. TESTING AND EVALUATION

A. Functional Testing

The modules were tested as listed in Table IV. The tests check that each module works correctly and that the security features behave as designed.

TABLE IV Functional tests described for the platform.

Test area	What is checked
Module tests	Unit and functional tests of each module
Data validation	Drug molecular data and biological profiles
Model execution	AI/ML execution and model-consensus checks
Target/pathway	Drug–target–pathway relationships
Evidence Passport	Passport generation and integrity verification
Security	Authentication, access permissions, audit logging

B. Evaluation Protocol

The prediction layer is evaluated with accuracy, precision, recall, F1-score and ROC-AUC. With true positives TP, false positives FP and false negatives FN, precision is $P = TP/(TP + FP)$, recall is $R = TP/(TP + FN)$, and

$$F1 = 2PR/(P + R) \quad (5)$$

Models are to be trained on public toxicity data such as the Tox21 collection [2], using stratified train/test splits and cross-validation, and compared against single-model baselines [1]. The numerical values must come from actual experiments. No performance figures are claimed in this paper; reporting them is the next stage of the work.

C. Illustrative Workflow

The following example is qualitative and uses no measured values. A researcher enters a candidate compound. The platform standardizes its structure, computes the fingerprint, and attaches the human-context biological profile. Target and pathway analysis then lists the proteins and pathways linked to the compound. The four learners each return a class probability. If three of the four models agree and the entropy in (2) is low, the compound is reported as a stable signal with high consensus. If the models disagree or the entropy is high, the report marks the result as uncertain and recommends that the researcher treat it with caution. In both cases the passport is sealed, and a reviewer can later reopen it and confirm that nothing has changed.

D. Comparison with Conventional Pipelines

Table V compares the proposed platform with a conventional pipeline that uses a single prediction tool, using the features discussed in this paper. The comparison is by design features only and does not compare measured accuracy.

TABLE V Design comparison.

Feature	Single tool	Cygen Vault
Models compared	One	Four
Output	Score or label	Risk, confidence, uncertainty
Evidence view	Separate	Unified profile
Species contexts	Often mixed	Kept separate
Integrity record	None	Passport and audit chain
Access control	Tool dependent	Role-based

VI. DISCUSSION AND LIMITATIONS

A. Value and Limits

The main value of Cygen Vault is that every prediction is delivered with its evidence, its model agreement, its uncertainty and a verifiable record. This supports reproducibility and audit, which single-score tools do not provide. The platform has limits. It is a prioritization aid, not a source of proof; its predictions depend on the quality and coverage of the training data; and cross-species comparisons are deliberately restricted. The security design protects stored records but does not replace institutional data-governance policies. Finally, the claims about prediction quality remain to be confirmed by the experiments described in Section V-B.

B. Responsible Use

Cygen Vault ranks candidates for further study. Its outputs must not be used as a clinical decision or as evidence that a drug is safe. Because biological data may be sensitive, access should follow institutional ethics and data-protection rules, and any work involving human or animal samples needs the usual ethical approval.

VII. CONCLUSION AND FUTURE WORK

This paper presented Cygen Vault, a secure and traceable platform that joins molecular profiling, multi-omics evidence, multi-model prediction, uncertainty assessment and cryptographically verifiable records. Future work includes quantitative benchmarking on public datasets, stronger multi-omics integration, explainable AI methods such as Shapley-value attribution [15], larger and more diverse datasets, improved privacy mechanisms, and laboratory validation of prioritized candidates.

DECLARATION

A generative AI tool was used solely for grammar checking of the manuscript text. The authors reviewed all content and take full responsibility for the manuscript.

REFERENCES

[1] A. Mayr, G. Klambauer, T. Entertainer, and S. Hochreiter, "DeepTox: Toxicity prediction using deep learning," Front. Environ. Sci., vol. 3, Art. no. 80, 2016.

[2] R. Huang et al., "Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs," Front. Environ. Sci., vol. 3, Art. no. 85, 2016.

[3] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," Genome Biol., vol. 18, no. 1, Art. no. 83, 2017.

[4] D. Kaleeswaran, S. Priyadharshini, L. Mubaraali, and A. Rajaram, "Quantum adaptive pixel shuffling and blockchain integrated steganographic encryption for securing data in the quantum era," Int. J. Pattern Recognit. Artif. Intell., Art. no. 2650017, 2026.

[5] T. T. Kuo, H. E. Kim, and L. Ohno-Machado, "Blockchain distributed ledger technologies for biomedical and health care applications," J. Amer. Med. Informat. Assoc., vol. 24, no. 6, pp. 1211–1220, 2017.

[6] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[7] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2016, pp. 785–794.

[9] D. S. Wishart et al., "DrugBank 5.0: A major update to the DrugBank database for 2018," Nucleic Acids Res., vol. 46, no. D1, pp. D1074–D1082, 2018.

[10] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Role-based access control models," Computer, vol. 29, no. 2, pp. 38–47, 1996, doi: 10.1109/2.485845.

[11] L. Mubaraali, S. SathishKumar, R. Sarvanan, and A. U. Begum, "Wearable plasmonic sweat analyzers with AI-IoT fusion for rapid electrolyte and stress biomarker profiling," Plasmonics, vol. 21, no. 8, pp. 3371–3388, 2026.

[12] R. Rasappan Senthil Kumar, U. Arun Kumar, A. Flah, and H. Kraiem, "Computational design and AI optimization of a symmetry-engineered ultra-broadband metasurface solar absorber for renewable energy systems," Energy Rep., vol. 16, Art. no. 109524, 2026, doi: 10.1016/j.egyr.2026.109524.

[13] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci., vol. 28, no. 1, pp. 31–36, 1988.

[14] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," J. Chem. Inf. Model., vol. 50, no. 5, pp. 742–754, 2010.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017.