ToxTwin — Consolidated Technical Synthesis

Version V2.3 · April 2026 · Twingital Institute / Qualees

1. Overview

ToxTwin is a predictive toxicological scoring platform for pre-Phase 1 assessment, built on an end-to-end GNN (Graph Neural Network) pipeline. It covers 14 toxicological endpoints (12 Tox21 + Ames mutagenicity + hERG inhibition), complemented by a pharmacological interpretation layer powered by a local LLM. The entire system runs on sovereign infrastructure with no cloud dependency.

The platform went through six major development versions (V1.0→V2.3), marked by the rigorous discovery and correction of methodological biases. Version V2.3 is the first release with metrics validated on a strict 5-fold scaffold CV protocol.

Key V2.3 metrics:

Indicator	Value	Protocol
Tox21 Mean AUC	0.867 ± 0.043	5-fold scaffold CV
Ames AUC	0.843 ± 0.029	5-fold scaffold CV
hERG AUC	0.785 ± 0.053	5-fold scaffold CV
Targets reached (≥ threshold)	12/14	5-fold scaffold CV
Calibration ECE (14 ep.)	< 0.05	Isotonic Regression OOF
Endpoints	14	12 Tox21 + Ames + hERG

---

2. Trajectory V1.0 → V2.3 — Biases, Corrections, Gains

2.1 Timeline

Version	Period	Workstream	Key outcome
V1.0	Jan 2026	Initial GNN architecture, pre-training, Tox21 fine-tuning	AUC 0.857 reported — not reproducible

Version	Period	Workstream	Key outcome
V1.1	Feb 2026	Calibration, applicability domain	ECE 0.043 PASS — AD on random embeddings
V1.2	Feb–Mar 2026	hERG + Ames extension, Focal Loss	Ames AUC 0.968 — circular
V1.3	Mar–Apr 2026	Full audit, architecture correction, Hansen labels	True AUC: 0.594 ± 0.056
V2.0	Apr 2026	New corpus, pre-training, multi-task fine-tuning	Tox21 0.722, Ames 0.769, hERG 0.691
V2.1	Apr 2026	Training optimisation	Marginal gain
V2.2	Apr 2026	Multi-representation ensemble architecture	Selective per-endpoint gains
V2.3	Apr 2026	Per-endpoint selection mechanism + V2.3 calibration	Tox21 0.867, 12/14 targets reached

2.2 Biases Discovered and Corrected

Fundamental architecture bug (V1.0–V1.2). Silent incompatibility between the pre-trained checkpoint and the inference architecture. Loading with `strict=False` masked the issue: weights were ignored, producing meaningless embeddings. Consequence: all V1.0–V1.2 scores — predictions, AD, calibrators — were computed on random weights.

Data leakage (V1.0–V1.1). Partial contamination of the validation set through compound migration to the training set. Initial random (non-scaffold) split: estimated overestimation of 3–8% (Sheridan 2013). AUC 0.857 measured during training on a partially contaminated validation set.

Ames circularity (V1.2). Ames AUC 0.968 = structural artefact. SMARTS labels encode substructure alerts that the GNN learns to reproduce (generation rule, not biological phenotype). True AUC on experimental Hansen data: 0.560 ± 0.036.

Degenerate AD (V1.1–V1.2). Applicability domain components computed in an unsuitable dimensionality space with self-referencing thresholds.

2.3 Corrective Measures

Phase 0 — Foundations. Strict Bemis-Murcko scaffold split across the full corpus. InChIKey verification: train ∩ val ∩ test = ∅. Frozen holdout test set (single use). 5-fold CV baseline established.

Phase 1 — Corpus and encoder (V2.0). Pre-training corpus replaced with a reference drug-like corpus. Re-pre-training and multi-task fine-tuning across 14 endpoints.

Phase 2 — Ensemble architecture (V2.2). Fusion of complementary molecular representations. The fusion architecture, dimensionalities and projection mechanisms are Twingital Institute intellectual property.

Phase 3 — Per-endpoint selection (V2.3). The routing mechanism, selection logic and assignment table are Twingital Institute intellectual property.

---

3. V2.3 Architecture

3.1 Components

The V2.3 architecture combines multiple molecular representations (topological, attentional, substructural) through a fusion and per-endpoint selection mechanism.

Detailed component specifications — dimensionalities, training status, parameter counts — are Twingital Institute intellectual property.

3.2 Per-Endpoint Selection

The routing table, selection criteria and per-model distribution are Twingital Institute intellectual property.

3.3 Inference Pipeline

```
query (SMILES or name)
  → SMILES resolution (RDKit canonicalisation or PubChem REST)
  → Molecular featurisation
  → Encoding, fusion and selection (Twingital Institute intellectual property)
  → Probabilistic calibration
  → Applicability domain evaluation
```

---

4. Medallion Data Pipeline

4.1 Bronze — Raw Ingestion

Source	Volume	Content
PubChem	~100,000 compounds	SMILES, InChI, MW, physicochemical descriptors

Source	Volume	Content
ChEMBL v34	~100K compounds + activities	IC50, Ki, EC50, hERG data
Tox21	7,831 compounds	Binary labels, 12 assays
NER enrichment	~7,300 processed	LD50, LDLo, TDLo, NOAEL, target organs

Toxicological named entity extraction from PubChem monographs via local LLM.

4.2 Silver — Curation

Multi-phase curation pipeline: sanitisation, deduplication, filtering, label joining and enriched profile integration. Result: ~145,000 deduplicated compounds.

4.3 Gold — Training Dataset

Strict Bemis-Murcko scaffold split. InChIKey verification: train ∩ val ∩ test = ∅.

---

5. Applicability Domain

Composite AD score computed from independent components assessing structural similarity, latent space proximity, and regional density.

The components, weights, normalisation formulas and decision thresholds are Twingital Institute intellectual property.

Decision: `is_in_domain` = True/False based on a calibrated composite threshold.

---

6. Probabilistic Calibration

Each endpoint has a calibrator trained on out-of-fold predictions from the 5-fold CV protocol. Post-calibration ECE < 0.05 across all endpoints.

---

7. LLM Interpretation Layer

Local LLM transforming raw scores into pharmacologically grounded toxicological reports.

The anti-hallucination mechanisms, prompt architecture, generation parameters and knowledge base structure are Twingital Institute intellectual property.

Dual output: human-readable structured report + machine-readable structured data. SSE streaming to mask latency.

---

8. REST API V2.3

8.1 Endpoints

Method	Endpoint	Description
`POST`	`/v1/score-toxicity`	Full 14-endpoint prediction + AD
`POST`	`/v1/interpret`	LLM interpretation
`GET`	`/health`	Status + validation metrics
`GET`	`/v1/endpoints`	Detailed per-endpoint list

8.2 Error Codes

Code	Internal code	Cause
400	`invalid_smiles`	Syntactically invalid SMILES
400	`missing_query`	`query` field absent
422	`resolution_failed`	Name not resolved via PubChem
422	`molecule_too_large`	Molecule > 500 heavy atoms
503	`model_unavailable`	Model not loaded or GPU error
504	`interpret_timeout`	LLM interpretation timeout

---

9. Infrastructure

Deployment on sovereign physical server (dedicated GPU, no cloud dependency for compute). API exposed via encrypted tunnel. Static frontend on CDN. Magic link authentication. Application-level rate limiting.

---

10. Detailed V2.3 Performance (14 Endpoints)

Endpoint	Baseline	V2.3 (5-fold)	Target	Status
NR-AR	0.584	0.755	0.70	✓

Endpoint	Baseline	V2.3 (5-fold)	Target	Status
NR-AR-LBD	0.535	0.787	0.70	✓
NR-AhR	0.644	0.902	0.85	✓
NR-Aromatase	0.603	0.887	0.85	✓
NR-ER	0.605	0.854	0.78	✓
NR-ER-LBD	0.645	0.925	0.78	✓
NR-PPAR-γ	0.571	0.752	0.72	✓
SR-ARE	0.630	0.859	0.78	✓
SR-ATAD5	0.687	0.940	0.80	✓
SR-HSE	0.507	0.898	0.78	✓
SR-MMP	0.495	0.909	0.85	✓
SR-p53	0.622	0.934	0.85	✓
Ames Hansen	0.560	0.843	0.87	✗ (−0.027)
hERG	0.548	0.785	0.83	✗ (−0.045)
Tox21 MEAN	**0.594**	**0.867**	**0.80**	✓

---

11. Benchmarks vs Published Literature

Endpoint	Published reference	Reference AUC	V2.3	Gap
Ames mutagenicity	DeepAmes (Xu et al. 2021)	0.889	0.843	−0.046
hERG inhibition	CardioTox (Karim et al. 2023)	0.872	0.785	−0.087
NR-AhR	Tox21 challenge winner (2014)	0.911	0.902	−0.009
SR-MMP	AttentiveFP (Xiong et al. 2020)	0.848	0.909	+0.061
NR-Aromatase	GROVER (Rong et al. 2020)	0.882	0.887	+0.005

---

12. Technical Limits V2.3

Limit	Value	Description
Maximum molecule size	500 heavy atoms	Beyond this, quasi-systematically out of domain

Limit	Value	Description
Supported atoms	C H N O S P F Cl Br I	Other atoms: degraded featurisation
Transition metals	Not supported	Pt, Ru, Au, etc.: outside scope (planned V3.0)
Peptides	Not recommended	> 5 amino acids: likely out of domain

---

13. Security and Compliance

Regulatory positioning: ToxTwin is not a medical device (no direct diagnostic/ therapeutic use). Pre-clinical toxicological scoring falls outside the "high-risk" category (EU AI Act, Annex III). Pipeline designed proactively to meet high-risk requirements: audit trail, model versioning, uncertainty and applicability domain exposed in every API response.

---

14. V2.4 Plan — Quality Control

14.1 Internal Validation (Completed)

Strict 5-fold scaffold CV, controlled inter-fold variance, InChIKey contamination = 0, ECE < 0.05 calibration, consistent routing.

14.2 Robustness Tests (To Be Conducted)

Inter-run reproducibility, SMILES representation invariance, SAR sensitivity on known pairs, applicability domain coverage, structure-activity monotonicity.

14.3 External Validation (To Be Conducted)

Frozen holdout test set (single use). Prospective external validation (partner agreement or ECHA). Published benchmarks (DeepAmes, CardioTox).

14.4 Improvement Axes

Priority	Action
P1	Ames and hERG corpus enrichment
P3	DILI, ClinTox, Carcinogens extension
P3	Transition metal support (V3.0)

---

15. References

- DiMasi JA et al. *J Health Econ.* 2016;47:20-33.
- Mayr A et al. DeepTox. *Front Environ Sci.* 2016;3:80.
- Xiong Z et al. *J Med Chem.* 2020;63(16):8749-8760.
- Rong Y et al. *NeurIPS.* 2020.
- Sheridan RP. *JCIM.* 2013;53(4):783-790.
- Wu Z et al. MoleculeNet. *Chem Sci.* 2018;9(2):513-530.
- Xu C et al. *Nat. Comm.* 2021.
- Karim A et al. CardioTox. *JCIM.* 2023.
- Regulation (EU) 2017/745 (MDR).
- Regulation (EU) 2024/1689 (EU AI Act).
