ToxTwin V2.4 — Le tri-routeur, ou pourquoi un seul modèle ne suffit jamais

Note technique · Avril 2026 · Twingital Institute

De V2.3 à V2.4 : enrichissement des corpus Ames et hERG, architecture tri-modèle sélective, et les leçons méthodologiques d'un cycle d'amélioration contrôlé.

La thèse en une phrase

Améliorer un pipeline multi-endpoint ne consiste pas à entraîner un meilleur modèle mais à identifier quel modèle est meilleur *pour chaque endpoint*, et à le prouver sur un protocole qui ne triche pas.

Contexte : ce que V2.3 avait établi

ToxTwin V2.3 couvre 14 endpoints toxicologiques (12 Tox21, Ames mutagénicité, hERG inhibition) via un pipeline GNN end-to-end validé en 5-fold scaffold CV strict. Le router V2.3 sélectionnait, par endpoint, le meilleur entre deux modèles : un GINEConv fine-tuné (V2.0a) et un ensemble multi-représentation GINEConv + AttentiveFP + Morgan ECFP6 (V2.2).

Métriques V2.3 : Tox21 mean 0.867, Ames 0.843, hERG 0.785. Douze endpoints sur quatorze atteignaient leur seuil cible. Les deux manquants :

- Ames (−0.027 sous cible)
- hERG (−0.045)

posaient une question précise : le levier est-il l'architecture ou les données ?

Diagnostic V2.4 : le levier, ce sont les données (mais pas n'importe lesquelles)

Premier échec : la carcinogénicité n'est pas la mutagénicité

L'AID 1259411 de PubChem, étiqueté « Ames » dans plusieurs méta-analyses, est en réalité un assay de *carcinogénicité in vivo multi-espèces*. Ses labels ne corrèlent que partiellement avec la mutagénicité bactérienne. L'intégration naïve de ces 547 composés dans le fine-tuning multi-tâche a produit une régression sur tous les endpoints, y compris Ames lui-même : de 0.769 à 0.698 en V2.0a.

La leçon est banale mais coûteuse : un composé « Ames positif » dans un assay carcinogénicité in vivo et un composé Ames positif au sens ICH S2R1 (Salmonella reverse

Jérôme Vetillard · VP R&D & CPO, Qualees / Twingital Institute

mutation, TA98/TA100 ± S9) ne portent pas la même information. Les confondre injecte du bruit, pas du signal.

Correction : le dataset ISS, source primaire réglementaire

Le corpus retenu provient de l'Istituto Superiore di Sanità (ISS), distribué via Mendeley Data. Ses caractéristiques décisives : labels par souche bactérienne (TA98, TA100, TA102, TA1535, TA1537), curation multi-étapes documentée, et harmonisation selon la convention ICH S2R1 (positif si au moins une souche positive). Après sanitization RDKit et déduplication InChIKey contre le corpus Hansen existant, 1 511 composés nouveaux ont été intégrés, avec un ratio positifs/négatifs quasi-équilibré (1.1:1).

Le deuxième levier : hERG via ChEMBL

Pour hERG, le réservoir est ChEMBL v34. La cible CHEMBL240 (KCNH2) agrège 22 273 activités IC50/Ki. Après filtrage nM, binarisation au seuil 10 µM, sanitization et déduplication, 6 485 composés nouveaux ont été intégrés. Effet collatéral bénéfique : le corpus hERG passe d'un déséquilibre 65/35 à un ratio 48.5/51.5, rééquilibrage qui améliore la calibration du modèle.

Ce que les features conformationnelles 3D n'apportent pas

Hypothèse testée : le canal hERG étant pharmacologiquement sensible à la forme 3D du ligand dans son vestibule, l'ajout de descripteurs conformationnels (NPR, USR, USRCAT, 79 dimensions) devrait améliorer la prédiction.

Résultat sur 5-fold scaffold CV :

Configuration	**Ames AUC**	**hERG AUC**
Morgan ECFP6 seul	0.841	0.750
3D seul	0.763	0.679
Morgan + 3D	0.842	0.764

Le gain marginal sur hERG (+0.014) et nul sur Ames (+0.001) indique que le GNN et les fingerprints topologiques capturent déjà l'essentiel de l'information structurale exploitable à cette échelle de corpus. L'information conformationnelle existe, mais elle est redondante avec ce que le modèle d'ensemble extrait par d'autres voies. Ce n'est pas le levier.

Jérôme Vetillard · VP R&D & CPO, Qualees / Twingital Institute

L'architecture tri-routeur

La découverte opérationnelle de V2.4 est que les données Ames ISS et les données hERG ChEMBL ne bénéficient pas au même modèle d'ensemble. Un ensemble entraîné sur le corpus Ames enrichi (V2.4b) produit le meilleur score Ames mais un score hERG inférieur. Un ensemble entraîné sur le corpus hERG enrichi (V2.4d) produit l'inverse.

La solution est structurelle, pas paramétrique : un *tri-routeur* qui sélectionne, pour chaque endpoint, le modèle optimal parmi trois :

1. V2.0a (GINEConv multi-tâche),
2. V2.4b (ensemble optimisé Ames),
3. V2.4d (ensemble optimisé hERG).

Table de routage V2.4 :

Modèle	**Endpoints**
V2.4b	NR-AR-LBD, NR-AhR, NR-ER, NR-ER-LBD, NR-PPAR-γ, SR-ARE, SR-HSE, SR-MMP, SR-p53, Ames
V2.4d	NR-AR, NR-Aromatase, SR-ATAD5, hERG

L'ensemble V2.4b sert 10 endpoints, l'ensemble V2.4d en sert 4. L'encodeur V2.0a reste disponible en fallback mais n'est sélectionné par aucun endpoint dans le routing optimal.

Résultats V2.4

Métrique	**V2.3**	**V2.4**	**Delta**
Tox21 mean AUC	0.867	0.898	+0.031
Ames AUC	0.843	0.853	+0.010
hERG AUC	0.785	0.800	+0.015
Cibles atteintes (≥ seuil)	12/14	12/14	=

Protocole inchangé : 5-fold scaffold CV strict Bemis-Murcko, vérification InChIKey train ∩ val = ∅, holdout gelé non utilisé pour la sélection.

Les deux endpoints manquants restent Ames (0.853 vs cible 0.87, gap −0.017) et hERG (0.800 vs cible 0.83, gap −0.030). Les gaps se sont réduits de 37 % (Ames) et 33 % (hERG) par rapport à V2.3.

Jérôme Vetillard · VP R&D & CPO, Qualees / Twingital Institute

Ce que V2.4 enseigne

Un modèle unique est un compromis, pas un optimum. Le fine-tuning multi-tâche produit un encodeur généraliste, mais la spécialisation par endpoint (données spécifiques, tête spécifique, sélection par routeur) surpasse systématiquement l'encodeur partagé. Le tri-routeur ne choisit pas le « meilleur modèle » mais le meilleur modèle *pour cette tâche*.

Les données battent l'architecture, mais pas n'importe quelles données. L'injection de données mal étiquetées (carcinogénicité vs. mutagénicité) dégrade les performances même si le volume augmente. La curation (vérification de la source primaire, harmonisation des labels selon les conventions réglementaires, contrôle du ratio positifs/négatifs) est le travail qui produit le gain, pas le téléchargement.

Les features supplémentaires ne compensent pas les données manquantes. Les descripteurs 3D, malgré leur justification pharmacologique, n'apportent qu'un gain marginal quand le modèle topologique est déjà correctement entraîné. Le plafond de performance actuel est un plafond de couverture chimique, pas de capacité représentationnelle.

Limites

Le tri-routeur introduit une complexité opérationnelle : trois modèles chargés en mémoire GPU au lieu de deux, et une table de routage qui devra être revalidée à chaque évolution des données. La sélection par endpoint repose sur un val set unique par fold, pas sur une comparaison inter-folds (la variance de sélection n'est pas quantifiée). Les métriques V2.4 ne sont pas directement comparables aux métriques V2.3 pour les endpoints dont le corpus a changé (Ames, hERG), bien que le protocole de validation reste identique. Enfin, le holdout gelé n'a pas été utilisé : il reste disponible pour une validation finale pré-déploiement.

Infrastructure

L'ensemble du pipeline s'exécute sur infrastructure souveraine (GPU dédié, pas de dépendance cloud pour le compute). L'API REST V2.4 expose le tri-routeur via les mêmes endpoints que V2.3. Aucune modification côté client n'est nécessaire (le champ routing dans la réponse API indique désormais V2.4b ou V2.4d au lieu de V2.0a ou V2.2.)

Jérôme Vetillard · VP R&D & CPO, Qualees / Twingital Institute