

Un domaine d'applicabilité mesure une proximité, pas une capacité

Ce qu'un dispositif de contrôle hérite de l'abstraction sur laquelle il est branché

Certains contrôles ne peuvent pas auditer les pertes produites par l'abstraction dont ils héritent. La phrase porte une condition, et la condition est tout l'argument. Elle ne dit pas qu'aucun domaine d'applicabilité ne peut détecter ce qu'un modèle écarte. Elle dit qu'un domaine construit en aval de la même abstraction ne le peut pas. Retirez l'abstraction partagée et l'affirmation se dissout, ce qui en fait une affirmation plutôt qu'une déploration.

La position est une chaîne. Une featurization détruit de l'information. Une mesure est calculée en aval d'elle. Une agrégation ne peut pas refuser ce que la mesure ne voit pas. Une interface promeut le résultat en autorisation. Quatre maillons, un chemin, et la défaillance est située dans le chemin plutôt que dans un maillon.

Ce qui suit audite un artefact de cet Institut :

ToxTwin V2.4 est un service de prédiction toxicologique : quatorze endpoints, douze issus du panel Tox21 auxquels s'ajoutent la mutagénicité d'Ames et l'inhibition hERG, servis par un tri-routeur au-dessus de deux encodeurs à réseaux de neurones sur graphes. Il est nommé, non anonymisé. Cinq textes doctrinaux de ce corpus ont soutenu qu'un dispositif de contrôle ne peut pas excéder la représentation sur laquelle il est construit. Celui-ci est le premier avec un artefact mesuré : non pas le sixième énoncé du principe, sa première mesure.

1. Le cas

Deux chaînes SMILES, à un caractère près.

17 β -estradiol C[C@]12CC[C@@H]3c4ccc(O)cc4CC[C@H]3[C@@H]1CC[C@@H]2O
17 α -estradiol C[C@]12CC[C@@H]3c4ccc(O)cc4CC[C@H]3[C@@H]1CC[C@H]2O

Ce sont des épimères en C17. Même formule, même masse, même graphe topologique, cinq stéréocentres dont exactement un diverge. Le 17 α -estradiol est un agoniste nettement plus faible du récepteur nucléaire aux œstrogènes que le 17 β -estradiol (Kuiper GG, Carlsson B, Grandien K, Enmark E, Häggblad J, Nilsson S, et al.).

La divergence est documentée, les deux formes sont stables, et aucune ne se convertit en l'autre in vivo.

Soumises à ToxTwin V2.4 le 15 juillet 2026, elles retournent les mêmes quatorze scores à la troisième décimale. NR-ER 0,5588 pour les deux, hERG 0,8488 pour les deux, composite du domaine d'applicabilité 0,4673 pour les deux, is_in_domain: true pour les deux. Badge vert, deux fois.

Puis une troisième chaîne :

CC12CCC3c4ccc(O)cc4CCC3C1CCC2O

Composite 0,4673. Mêmes quatorze scores. is_in_domain: true.

Cette troisième chaîne n'est pas une molécule. Ses cinq stéréocentres sont non spécifiés, ce qui en fait une classe d'équivalence de stéréoisomères dont l'estradiol est un membre. Le dispositif lui a attribué quatorze scores de toxicité, l'a déclarée dans le domaine d'applicabilité, et a produit un paragraphe réglementaire citant ICH S7B à propos de « cette molécule ».

Le domaine d'applicabilité a certifié une requête sans référent.

2. Où regarde la sonde

Deuxième maillon. La question n'est pas quelle métrique le domaine utilise, mais dans quel espace elle est calculée, et la réponse n'exige aucune mesure.

Le score de domaine est une somme pondérée de trois termes, portant 0,30, 0,40 et 0,30. Trois noms, trois sources apparentes de confiance.

Commençons par celle que n'importe qui peut vérifier. Le premier terme est une similarité de Tanimoto maximale sur des empreintes Morgan de rayon 2 sur 2048 bits. Le générateur d'empreintes de RDKit accepte un argument includeChirality, et sa valeur par défaut est False. Ce défaut n'est pas un interne de ce système. C'est un comportement de bibliothèque documenté, et trois lignes dans n'importe quelle session Python reproduisent ce qui en découle.

Les deux autres termes sont plus mal placés, et ici le lecteur me croit sur parole plutôt que sur la sienne. La lecture du chemin de calcul établit que le terme de voisinage et le terme de densité sont tous deux des fonctionnelles d'un unique intermédiaire : une projection en douze dimensions retenant 80,2 pour cent de la variance d'un embedding de dimension 512. Ce ne sont pas deux mesures. Ce sont deux normalisations d'un même vecteur à deux échelles incompatibles, et 0,70 du composite repose sur un point unique dans un espace unique.

Un mot sur le nuage de référence, rapporté de même, et le fait doit être séparé de ce qui le suit. Fait : les trois termes sont calculés contre un corpus de 11 414 molécules, étiquetées pour deux endpoints, la mutagénicité et l'inhibition du canal cardiaque. Les douze étiquettes Tox21 vivent dans un jeu de données distinct et n'entrent jamais dans le calcul du domaine. Conséquence : le nuage de référence n'est pas le corpus d'entraînement de douze des quatorze endpoints que le domaine autorise. C'est tout ce que le fait soutient, et il n'autorise pas une phrase sur ce à quoi la molécule ressemble, puisque l'espace latent a été entraîné sur davantage que ce que ses étiquettes déclarent.

Un point terminologique, puis le mot est retiré. Capacité ici ne signifie pas capacité du modèle, capacité de représentation, ni dimension de Vapnik-Chervonenkis. Cela signifie la préservation de l'information : quelles propriétés de l'objet survivent à la featurization.

Deux questions existent à propos de toute requête. Ressemble-t-elle à ce que j'ai vu, et lesquelles de ses propriétés survivent à mon encodage. Le domaine pose la première et se lit comme s'il répondait à la seconde.

3. Le facteur commun

Premier maillon, atteint par le bas. La sonde ne peut pas signaler le défaut parce que la sonde et le modèle ne sont pas dans une relation de dépendance. Ce sont des frères.

Le terme structurel lit une empreinte Morgan. Le modèle lit une featurization atomique en 163 dimensions du genre benchmark. Les termes de voisinage et de densité lisent l'embedding latent d'un troisième encodeur, plus ancien que les deux qui servent les quatorze prédictions. Trois chemins de calcul, trois représentations, aucune ligne partagée. Et le même effacement : la featurization du modèle ne lit jamais le tag chiral, et le générateur d'empreintes ne reçoit jamais l'argument qui le lui ferait faire. L'aveuglement n'est pas une architecture partagée. C'est une convention partagée, héritée des valeurs par défaut des bibliothèques et de l'usage des benchmarks, qui a survécu à trois changements d'encodeur sans que personne ne la décide.

L'argument n'exige aucun appareillage. L'information disponible pour un calcul est bornée par la représentation qu'il lit, et bornée est entendu au sens de la théorie de l'information : aucune fonction d'une représentation ne distingue ce que cette représentation envoie sur le même point. Ce n'est pas une limitation du calcul mais de ce qui lui est parvenu. Énoncé précisément : l'affirmation n'est pas que le contrôle calcule la mauvaise fonction, c'est qu'il calcule une fonction correcte sur une tribu que la représentation amont a déjà grossie. Le squelette nu de la section 1 n'est pas un défaut de l'interface. C'est un atome de cette tribu, et les deux épimères y vivent. La sonde a retourné le même nombre trois fois parce qu'elle a reçu le même objet trois fois.

Trois objections méritent d'être soulevées ici plutôt que dans un fil de commentaires.

La première : deux encodeurs partageant des primitives apprennent des géométries différentes, donc inférer un angle mort partagé de primitives partagées est illégitime. Correct sur la géométrie, sans pertinence pour l'affirmation. Trois générations d'encodeurs ont appris trois géométries à partir d'entrées dont le tag chiral avait déjà été retiré. Une couche ne reconstruit pas un bit absent.

La deuxième est plus acérée. Un calcul peut être indirectement sensible à une propriété qu'il n'encode pas, par corrélation. Vrai en général, et faux ici, pour une raison qui mérite d'être énoncée avec soin. Les molécules de la section 1 sont des diastéréoisomères, non des énantiomères : leurs propriétés non chirales, dont la solubilité et le moment dipolaire, diffèrent effectivement, et une featurization portant des descripteurs physicochimiques verrait quelque chose. Celle-ci n'en porte aucun. Elle est purement topologique, les deux épimères sont envoyés sur le graphe identique, et l'information mutuelle entre ce que la sonde lit et la propriété en jeu n'est pas petite. Elle est nulle, exactement. La borne n'est rien, et elle n'est rien parce que la représentation écrase toute géométrie, ce qui est une condition et non une loi.

La mesure confirme la lecture et amène avec elle la troisième objection. Sur RDKit 2026.03.3, la similarité de Tanimoto entre les deux épimères, calculée avec la configuration propre à la sonde, vaut 1,0000. Calculée avec `useChirality=True`, sur les mêmes molécules, avec la même bibliothèque, elle vaut 0,8182. L'information était présente. Elle a été écartée par un argument par défaut.

Ce qui invite à la réparation, et la réparation est l'erreur. Mettez l'argument à `True`, lisez le tag chiral, et les épimères se séparent. Ils restent frères. La représentation enrichie écarte toujours la conformation, l'état de protonation, l'équilibre tautomérique, la solvatation, et quoi que le modèle échoue ensuite à voir, la sonde échoue à le voir avec lui, en rapportant une proximité avec une confiance intacte. Le critère n'est pas d'encoder la chiralité. C'est que l'audit ne doit pas descendre du même facteur que le modèle. Basculer le drapeau sur les deux déplace la frontière. Cela ne retire pas la borne. La stéréochimie est ici le témoin, jamais l'accusé.

La norme le dit à sa manière. Les InChIKeys des trois états sont VOXZDWNPVJITMN-ZBRFXRBCSA-N, VOXZDWNPVJITMN-SFFUCWETSA-N et VOXZDWNPVJITMN-UHFFFAOYSA-N, et tout lecteur peut les recalculer. Le bloc de connectivité est identique, le bloc de stéréochimie ne l'est pas. L'IUPAC consacre un bloc entier de son identifiant à la stéréochimie parce que la connectivité seule n'identifie pas une molécule. La featurization calcule l'équivalent du premier bloc et l'appelle la molécule.

Un dispositif de contrôle ne détecte que ce que sa propre borne permet. Quand sa borne et celle du modèle descendent de la même abstraction, il hérite des pertes de cette abstraction quelle que soit sa calibration, et il en hérite silencieusement, parce qu'il ne reste rien dans son entrée sur quoi se taire.

4. L'écart de promotion

Troisième et quatrième maillons. L'objection à concéder est la plus forte, et elle est une question de fait établi.

Le domaine d'applicabilité n'a jamais promis la préservation de l'information. La littérature canonique le définit comme un domaine de similarité structurelle : Dimitrov et ses collègues [1], et Sahigara et ses collègues pour la revue comparative [2], parmi les co-auteurs de laquelle figure Kamel Mansouri, auteur du protocole de standardisation sur lequel la section 5 revient. Deux des trois sources externes citées ici partagent un auteur, et la convention est assez centralisée pour qu'une même main tienne sa définition et sa pratique. Sa question est de savoir si la requête ressemble au corpus d'entraînement, non si l'encodeur peut la représenter.

La forme la plus aiguë de cette objection concède davantage et exige davantage. Le domaine n'a pas tort. Il a raison. Il rapporte, correctement, que le modèle traite ces deux molécules à l'identique et qu'il a vu des choses leur ressemblant. Le défaut réside dans la featurization, et le domaine certifie la conformité à un espace qui était inadéquat avant qu'il n'arrive.

C'est exactement l'argument, et il arrive un maillon trop tôt. Un rapport correct de conformité à un espace inadéquat, publié sous la forme `is_in_domain: true`, cesse d'être un rapport correct. La mesure dit que le modèle est également aveugle aux deux molécules. Le champ dit que le modèle peut être utilisé sur les deux. Rien dans le pipeline n'effectue cette inférence, et tout en aval en dépend. La promesse n'est pas faite là où la mesure est calculée. Elle est faite deux fois à l'interface.

Elle est faite d'abord par l'agrégation. Une fois que l'information a survécu à la représentation, une seconde limitation, indépendante, apparaît, et elle appartient à l'arithmétique plutôt qu'à l'encodage. Une somme pondérée avec un seuil possède une propriété que la calibration ne peut pas retirer.

Lemme de non-veto. Soit $s = \sum w_i x_i$ avec $x_i \in [0,1]$, $\sum w_i = 1$, et une décision $s \geq \theta$. Le terme i ne peut imposer un refus que si $1 - w_i < \theta$, c'est-à-dire seulement si $w_i > 1 - \theta$. Par conséquent, pour que les n termes détiennent tous un veto, il faut $\theta > (n-1)/n$.

ToxTwin sert des poids de 0,30, 0,40 et 0,30, et un seuil de 0,40. Le poids minimal pour un veto est 0,60, aucun terme ne l'atteint, et avec trois termes chaque terme ne détiendrait un veto qu'au-dessus d'un seuil de 0,667.

L'instance mesurée est une arithmétique sur la décomposition que l'API retourne déjà. Pour le 17 β -estradiol, les termes Tanimoto et KDE contribuent à eux seuls 0,4024. Mettez le terme kNN à zéro, le plus lourd des trois, celui qui mesure la position dans l'espace latent propre au modèle, et laissez-le voter que la molécule est hors domaine. Le

composite vaut toujours 0,4024. La molécule est toujours admise. L'agrégation ne peut pas refuser. Elle peut seulement en dire moins.

Elle est faite ensuite par le champ. `is_in_domain` est un booléen là où la mesure est continue. Une proximité devient une autorisation en traversant un opérateur de comparaison, et c'est là qu'une similarité se change en permission d'usage. C'est une instance de ce que l'Article VI de ce corpus a nommé l'écart de promotion, et c'est la première que l'Institut ait trouvée dans son propre produit.

Et quand le refus survient, il ne refuse rien. Le cisplatine retourne un composite de 0,05 et `is_in_domain: false`, ses termes kNN et KDE tous deux à zéro, les 0,05 entièrement structurels. Les quatorze scores sont émis quand même, avec un paragraphe interprétatif. Le domaine est consultatif. L'interface ne l'est pas.

5. Ce n'est pas un défaut d'implémentation

La lecture tentante est qu'une équipe particulière a commis des erreurs particulières et qu'une meilleure ingénierie clôt l'affaire. L'architecture l'interdit, en trois temps.

ToxTwin sert quatorze endpoints à partir de deux encodeurs routés, sept chacun, routés par endpoint. Chaque encodeur induit son propre espace latent, donc sa propre métrique, donc sa propre relation de voisinage. Une molécule proche du corpus d'entraînement sous un encodeur peut en être loin sous l'autre, et aucun scalaire n'exprime les deux. Un score de domaine unitaire sur un ensemble routé a donc deux options. Il sélectionne un espace, et il est muet sur les sept endpoints servis par l'autre. Ou il agrège à travers les espaces, et le lemme de la section précédente s'applique à l'agrégation. Les deux branches échouent, et la seconde échoue par un théorème plutôt que par une inadvertance.

Quelqu'un a dû sélectionner un espace, et rien dans la réponse n'indique lequel. Ce choix est local et n'argumente rien ici. L'absence de tout choix légitime est le point, elle survit à toute recalibration, et elle découle de l'ensemblerage lui-même. À quel point les modèles composés sont devenus courants relève de l'impression, et ce texte n'en propose aucun décompte.

La corroboration externe converge, et ses limites voyagent avec elle. von Borries et ses collègues [3] rapportent que le protocole de standardisation de Mansouri et al. retire les composés inorganiques et organométalliques avant modélisation, et que les estimations d'incertitude sur ces composés sont surconfiantes. Ils y lisent un défaut de couverture à combler par de meilleurs descripteurs. Leur distance de Jaccard moyenne de 0,81 pour les composés non standardisés contre 0,57 pour les standardisés est proche d'un chiffre mesuré ici sur le carboplatine, et la proximité est une convergence de régime, non une identité : corpus différents, empreintes différentes, protocoles différents. Les

implémentations de référence du champ déclarent leur périmètre sans interroger leur espace.

Weaver et Gleeson [4] ont calculé le domaine dans l'espace de descripteurs propre au modèle et l'ont dit clairement. Ce n'est pas une inadvertance à pointer vingt ans plus tard. C'était un paradigme, et l'hypothèse est devenue implicite au point de ne plus être discutée.

6. Le principe et le critère

La généralisation est disponible et elle est assez étroite pour être utile.

Principe de séparation représentationnelle. L'information disponible pour un dispositif de gouvernance est bornée par l'information contenue dans la représentation sur laquelle il est branché. Quand cette représentation et celle du système contrôlé descendent d'une abstraction commune, la borne hérite des pertes de cette abstraction, quelle que soit la calibration du dispositif. Un audit ne devient informatif sur ces pertes qu'en étant branché sur une représentation qui retient de l'information que le modèle détruit.

D'où découle le critère opérationnel, comme condition nécessaire et rien de plus :

Critère du facteur commun. Un audit ne peut être informatif sur une propriété détruite par le modèle que s'il ne partage pas le facteur qui la détruit.

L'ordonnancement d'informativité sur lequel cela repose est ancien. Blackwell a établi la comparaison des expériences statistiques en 1953 [5], et la conséquence utile est l'élémentaire : une quantité calculée en aval d'un canal ne porte pas plus d'information sur l'état que le canal n'en portait.

Le critère est nécessaire et non suffisant, et la distinction importe davantage que le critère. Un audit ne partageant aucun facteur avec le modèle peut encore être trop bruité pour détecter la propriété, ou incapable de la calculer. Échouer au critère est disqualifiant. Le passer ne garantit rien. Et la suffisance, là où elle peut être argumentée, est indexée sur une propriété nommée, ce qui livre la concession qui coûte le plus et qui vaut ce qu'elle coûte : il n'existe pas d'audit agnostique aux propriétés. Un audit est toujours l'audit d'une liste énumérée.

Le corollaire opérationnel est l'endroit où ceci cesse d'être un argument et devient une spécification. Un port de validité ne devrait pas exposer une proximité. Il devrait exposer quelles propriétés de l'objet survivent à la featurization. Pour l'un ou l'autre épimère de la section 1, un tel port rapporterait cinq stéréocentres présents et zéro encodé, sans exécuter le modèle, sans corpus de référence, sans calibration, sans jeu de test. C'est une propriété du modèle, non de la requête. C'est déterministe, et cela se calcule en lisant le graphe d'appels.

L'objection à un port est la régression. S'il doit lire une représentation pour rapporter ce qu'une autre a perdu, qui audite le port. La réponse est positionnelle. Il n'a pas besoin d'une représentation plus riche, il a besoin d'une représentation antérieure. Le parseur détient encore les cinq stéréocentres au moment où le featurizer décline de les lire, donc le port compare l'objet du parseur au chemin de calcul du featurizer et rapporte la différence. Il n'est pas branché à côté du modèle. Il est branché avant lui.

Le corollaire dissout aussi le problème de composition plutôt que d'argumenter avec lui. Si un port expose quelles propriétés survivent à une featurization, et que chaque modèle routé a sa propre featurization, alors les ports sont par modèle par construction. Il n'y a pas de domaine unique à calculer, parce qu'il n'y a jamais eu de question unique à laquelle répondre.

C'est ainsi que la lecture derrière cet article a été conduite.

7. Domaine de validité

La thèse vaut pour les domaines d'applicabilité dont le calcul descend d'un facteur qu'ils partagent avec le modèle et qui détruit la propriété en jeu. Elle ne vaut pas pour un audit qui retient de l'information que le modèle écarte. L'observation réfutante est nommable : un domaine d'applicabilité calculé sur des descripteurs tridimensionnels, quantiques ou géométriques, retenant donc ce que la featurization du modèle retire, et signalant correctement l'inencodable. Elle bornerait cette thèse, ce qui est l'effet recherché.

Le lemme de non-veto vaut pour les scores composites pondérés avec seuil, non pour les domaines à mesure unique : leverage, distance de Mahalanobis, méthodes fondées sur les intervalles. Ceux-là ont d'autres problèmes, pas celui-ci.

Le port de validité tombe sous son propre principe, et la concession est due. Il rapporte les pertes de la featurization relativement à ce que le parseur a retenu, et le parseur est une abstraction lui aussi, ayant déjà écarté la conformation, l'état de protonation, la solvation. Un port ne peut pas énumérer ce que rien en amont de lui n'a enregistré. Il n'est pas plus riche que le modèle. Il est plus tôt, et plus tôt est tout ce que le critère demande.

Rien ici n'affirme que ToxTwin a tort à propos du 17 β -estradiol. Il est affirmé que la divergence expérimentale est réelle et documentée, que le dispositif retourne un seul nombre, et qu'il a donc tort à propos de l'une des deux sans pouvoir dire laquelle. Rien ici n'affirme quoi que ce soit sur les métaux, et rien ici ne revendique une impossibilité générale de la connaissance de soi : l'obstacle est positionnel, il concerne l'endroit où le contrôle est branché, et c'est pourquoi il est réparable.

Sur la provenance, puisque le sujet l'exige, et l'exigence tranche ici aussi. Chaque chiffre a été mesuré ou lu le 15 juillet 2026, contre un service en fonctionnement et son code

source, lors de travaux pour passer de ToxTwin 2.4 à ToxTwin 3.0 par ajout de familles moléculaires et de endpoints.

La preuve se scinde alors, et la scission doit être visible plutôt que lissée. Une moitié, tout lecteur la reproduit : la primitive d'empreinte est aveugle par défaut, les épimères collisionnent à 1,0000 sous elle et se séparent à 0,8182 sans elle, le bloc de connectivité est partagé par les trois états, et le service retourne un seul nombre pour les trois. L'autre moitié est rapportée, lue dans un code source non publié ici : que deux des trois termes descendent d'un même vecteur, que le contrôle est calculé dans un espace latent ne servant aucune prédiction, que le corpus de référence porte deux des quatorze endpoints.

La thèse repose sur la première moitié. La seconde explique le mécanisme, et un lecteur qui la décline garde le fait et perd la raison, ce qui fait un article plus court et non un article différent.

Une quantité reste non mesurée : l'écart entre les deux épimères, que la littérature secondaire rapporte de façon inconsistante et que seule la source primaire tranche. C'est aussi la seule quantité dont l'argument n'a pas besoin.

Un instrument qui ne peut pas vous dire ce qu'il a manqué de voir vous dira qu'il n'a rien vu d'anormal.

Références

1. Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemelä J, et al. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J Chem Inf Model*. 2005;45:839-49. doi:10.1021/ci0500381
2. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*. 2012;17(5):4791-810. doi:10.3390/molecules17054791
3. von Borries K, Beckwith KV, Goodman JM, Chiu WA, Jolliet O, Fantke P. Uncertainty-aware machine learning to predict non-cancer human toxicity for the global chemicals market. *Nat Commun*. 2026;17(1):647. doi:10.1038/s41467-025-67374-4
4. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model*. 2008;26:1315-26. doi:10.1016/j.jmgm.2008.01.002
5. Blackwell D. Equivalent comparisons of experiments. *Ann Math Stat*. 1953;24(2):265-72. doi:10.1214/aoms/1177729032