An Applicability Domain Measures a Proximity, Not a Capacity

What a control device inherits from the abstraction it is plugged into

Some controls cannot audit the losses produced by the abstraction they inherit. The sentence carries a condition, and the condition is the whole argument. It does not say that no applicability domain can detect what a model discards. It says that a domain built downstream of the same abstraction cannot. Remove the shared abstraction and the claim dissolves, which is what makes it a claim rather than a lament.

The position is a chain. A featurization destroys information. A measure is computed downstream of it. An aggregation cannot refuse what the measure cannot see. An interface promotes the result into an authorization. Four links, one path, and the failure is located in the path rather than in any link.

What follows audits an artifact of this Institute :

> ToxTwin V2.4 is a toxicological prediction service: fourteen endpoints, twelve from the Tox21 panel plus Ames mutagenicity and hERG inhibition, served through a tri-router over two graph neural encoders. It is named, not anonymized. Five doctrinal texts of this corpus have argued that a control device cannot exceed the representation it is built on. This is the first with a measured artifact: not the sixth statement of the principle, its first measurement.

1. The case

Two SMILES strings, one character apart.

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

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

They are C17 epimers. Same formula, same mass, same topological graph, five stereocenters of which exactly one diverges. 17α-estradiol is a markedly weaker nuclear estrogen receptor agonist than 17β-estradiol (Kuiper GG, Carlsson B, Grandien K, Enmark E, Häggblad J, Nilsson S, et al.).

The divergence is documented, both forms are stable, and neither converts into the other in vivo.

Submitted to ToxTwin V2.4 on 15 July 2026, they return the same fourteen scores to the third decimal. NR-ER 0.5588 for both, hERG 0.8488 for both, applicability domain composite 0.4673 for both, is_in_domain: true for both. Green badge, twice.

Then a third string:

CC12CCC3c4ccc(O)cc4CCC3C1CCC2O

Composite 0.4673. Same fourteen scores. is_in_domain: true.

This third string is not a molecule. Its five stereocenters are unspecified, which makes it an equivalence class of stereoisomers, of which estradiol is one member. The device assigned it fourteen toxicity scores, declared it within the applicability domain, and produced a regulatory paragraph citing ICH S7B about "this molecule".

The applicability domain certified a query with no referent.

2. Where the probe looks

Second link. The question is not which metric the domain uses but which space it is computed in, and the answer needs no measurement.

The domain score is a weighted sum of three terms, carrying 0.30, 0.40 and 0.30. Three names, three apparent sources of confidence.

Start with the one anyone can check. The first term is a maximum Tanimoto similarity over Morgan fingerprints at radius 2 into 2048 bits. RDKit's fingerprint generator takes an includeChirality argument, and its default is False. That default is not an internal of this system. It is documented library behaviour, and three lines in any Python session reproduce what follows from it.

The other two terms are worse placed, and here the reader takes my word rather than his own. Reading the computation path establishes that the neighbour term and the density term are both functionals of a single intermediate: one twelve-dimensional projection retaining 80.2 percent of the variance of a 512-dimensional embedding. They are not two measurements. They are two normalizations of one vector at two incompatible scales, and 0.70 of the composite rests on a single point in a single space.

A word on the reference cloud, reported likewise, and the fact must be separated from what follows it. *Fact*: all three terms are computed against a corpus of 11,414 molecules, labelled for two endpoints, mutagenicity and cardiac channel inhibition. The twelve Tox21 labels live in a separate dataset and never enter the domain computation.

Consequence: the reference cloud is not the training corpus of twelve of the fourteen endpoints the domain authorizes. That is all the fact supports, and it does not license a

sentence about what the molecule resembles, since the latent space was trained on more than its labels declare.

One terminological point, then the word is retired. *Capacity* here does not mean model capacity, representational capacity, or VC dimension. It means the preservation of information: which properties of the object survive the featurization.

Two questions exist about any query. Does it resemble what I have seen, and which of its properties survive my encoding. The domain asks the first and is read as if it answered the second.

3. The common factor

First link, reached from below. The probe cannot signal the defect because probe and model are not in a dependency relation. They are siblings.

The structural term reads a Morgan fingerprint. The model reads a 163-dimensional atom featurization of the benchmark kind. The neighbour and density terms read the latent embedding of a third encoder, older than either of the two serving the fourteen predictions. Three computation paths, three representations, no shared line. And the same erasure: the model's featurization never reads the chiral tag, and the fingerprint generator never receives the argument that would make it do so. The blindness is not a shared architecture. It is a shared convention, inherited from library defaults and benchmark usage, which survived three changes of encoder without anyone deciding it.

The argument requires no apparatus. The information available to a computation is bounded by the representation it reads, and *bound* is meant in its information-theoretic sense: no function of a representation distinguishes what that representation maps to the same point. This is not a limitation of the computation but of what reached it. Stated precisely: the claim is not that the control computes the wrong function, it is that it computes a correct function over a sigma-algebra the upstream representation has already coarsened. The bare skeleton of section 1 is not a defect of the interface. It is an atom of that sigma-algebra, and both epimers live inside it. The probe returned the same number three times because it received the same object three times.

Three objections deserve to be raised here rather than in a comment thread.

1. The first: two encoders sharing primitives learn different geometries, so inferring a shared blind spot from shared primitives is illegitimate. Correct about geometry, irrelevant to the claim. Three generations of encoders learned three geometries from inputs whose chiral tag had already been removed. A layer does not reconstruct an absent bit.
2. The second is sharper. A computation can be indirectly sensitive to a property it does not encode, through correlation. True in general, and false here, for a reason

worth stating carefully. The molecules of section 1 are diastereomers, not enantiomers: their non-chiral properties, solubility and dipole moment among them, do differ, and a featurization carrying physicochemical descriptors would see something. This one carries none. It is purely topological, both epimers map to the identical graph, and the mutual information between what the probe reads and the property at issue is not small. It is zero, exactly. The bound is nothing, and it is nothing because the representation crushes all geometry, which is a condition and not a law.

3. The measurement confirms the reading and adds the third objection with it. On RDKit 2026.03.3, the Tanimoto similarity between the two epimers, computed with the probe's own configuration, is 1.0000. Computed with useChirality=True, on the same molecules, with the same library, it is 0.8182. The information was present. It was discarded by a default argument.

Which invites the repair, and the repair is the mistake. Set the argument to True, read the chiral tag, and the epimers separate. They remain siblings. The enriched representation still discards conformation, protonation state, tautomeric equilibrium, solvation, and whatever the model then fails to see, the probe fails to see with it, reporting a proximity with undiminished confidence. The criterion is not encode chirality. It is that the audit must not descend from the same factor as the model. Flipping the flag on both moves the boundary. It does not remove the bound. *Stereochemistry is the witness here, never the defendant*.

The standard says as much in its own way. The InChIKeys of the three states are VOXZDWNPVJITMN-ZBRFXRBCSA-N, VOXZDWNPVJITMN-SFFUCWETSA-N and VOXZDWNPVJITMN-UHFFFAOYSA-N, and any reader can recompute them. The connectivity block is identical, the stereochemistry block is not. IUPAC devotes an entire block of its identifier to stereochemistry because connectivity alone does not identify a molecule. The featurization computes the equivalent of the first block and calls it the molecule.

A control device detects only what its own bound permits. When its bound and the model's descend from the same abstraction, it inherits that abstraction's losses whatever its calibration, and it inherits them silently, because there is nothing left in its input to be silent about.

4. The promotion gap

Third and fourth links. The objection to concede is the strongest one, and it is a matter of record.

The applicability domain never promised the preservation of information. The canonical literature defines it as a domain of structural similarity: Dimitrov and colleagues [1], and

Sahigara and colleagues for the comparative survey [2], among whose co-authors is Kamel Mansouri, author of the standardization protocol section 5 returns to. Two of the three external sources cited here share an author, and the convention is centralized enough that one hand holds both its definition and its practice. Its question is whether the query resembles the training corpus, not whether the encoder can represent it.

The sharpest form of it concedes more and demands more. The domain is not wrong. It is right. It reports, correctly, that the model treats these two molecules identically and has seen things resembling them. The defect lies in the featurization, and the domain certifies conformity to a space that was inadequate before it arrived.

This is exactly the argument, and it arrives one link too early. A correct report of conformity to an inadequate space, published as is_in_domain: true, stops being a correct report. The measure says the model is equally blind to both molecules. The field says the model may be used on both. Nothing in the pipeline performs that inference, and everything downstream depends on it. The promise is not made where the measure is computed. It is made twice at the interface.

It is made first by the aggregation. Once information has survived the representation, a second and independent limitation appears, and it belongs to the arithmetic rather than to the encoding. A weighted sum with a threshold has a property calibration cannot remove.

Lemma of non-veto. Let $s = \Sigma\, w_i x_i$ with $x_i \in [0,1]$, $\Sigma w_i = 1$, and a decision $s \geq \theta$. Term i can impose a refusal only if $1 - w_i < \theta$, that is, only if $w_i > 1 - \theta$. Consequently, for all n terms to hold a veto, $\theta > (n-1)/n$ is required.

ToxTwin serves weights of 0.30, 0.40 and 0.30, and a threshold of 0.40. The minimum weight for a veto is 0.60, no term reaches it, and with three terms every term would hold a veto only above a threshold of 0.667.

The measured instance is arithmetic on the decomposition the API already returns. For 17β-estradiol the Tanimoto and KDE terms alone contribute 0.4024. Set the kNN term to zero, the heaviest of the three, the one that measures position in the model's own latent space, and let it vote that the molecule is outside. The composite is still 0.4024. The molecule is still admitted. *The aggregation cannot refuse. It can only say less.*

It is made second by the field. is_in_domain is a boolean where the measure is continuous. A proximity becomes an authorization by crossing a comparison operator, which is where a similarity turns into a permission to use. This is an instance of what Article VI of this corpus named the promotion gap, and it is the first one the Institute has found in its own product.

And when the refusal does occur, it refuses nothing. Cisplatin returns a composite of 0.05 and is_in_domain: false, its kNN and KDE terms both at zero, the 0.05 entirely structural.

The fourteen scores are emitted anyway, with an interpretive paragraph. The domain is consultative. The interface is not.

5. This is not an implementation fault

The tempting reading is that a particular team made particular mistakes and that better engineering closes the matter. The architecture forbids it, in three steps.

ToxTwin serves fourteen endpoints from two routed encoders, seven each, routed per endpoint. Each encoder induces its own latent space, therefore its own metric, therefore its own neighbourhood relation. A molecule near the training corpus under one encoder may be far from it under the other, and no scalar expresses both. A unitary domain score over a routed ensemble therefore has two options. It selects one space, and it is silent about the seven endpoints served by the other. Or it aggregates across spaces, and the lemma of the previous section applies to the aggregation. Both branches fail, and the second fails by a theorem rather than by an oversight.

Someone had to select a space, and nothing in the response indicates which. That choice is local and argues nothing here. The absence of any legitimate choice is the point, it survives every recalibration, and it follows from ensembling itself. How common composed models have become is a matter of impression, and this text offers no count.

The external corroboration converges, and its limits travel with it. von Borries and colleagues [3] report that the standardization protocol of Mansouri et al. removes inorganic and organometallic compounds before modeling, and that uncertainty estimates on those compounds are overconfident. They read this as a coverage defect to be closed with better descriptors. Their mean Jaccard distance of 0.81 for non-standardized compounds against 0.57 for standardized ones sits close to a figure measured here on carboplatin, and the proximity is a convergence of regime, not an identity: different corpora, different fingerprints, different protocols. The field's reference implementations declare their perimeter without interrogating their space.

Weaver and Gleeson [4] computed the domain within the model's own descriptor space and said so plainly. This is not an oversight to point at twenty years later. It was a paradigm, and the assumption became implicit to the point of no longer being discussed.

6. The principle and the criterion

The generalization is available and it is narrow enough to be useful.

Principle of representational separation. The information available to a governance device is bounded by the information contained in the representation it is plugged into. When that representation and the controlled system's representation descend from a common abstraction, the bound inherits that abstraction's losses, whatever the device's

calibration. An audit becomes informative about those losses only by being plugged into a representation that retains information the model destroys.

From which the operational criterion follows, as a necessary condition and nothing more:

Common factor criterion. An audit can be informative about a property destroyed by the model only if it does not share the factor that destroys it.

The ordering of informativeness this rests on is old. Blackwell established the comparison of statistical experiments in 1953 [5], and the useful consequence is the elementary one: a quantity computed downstream of a channel carries no more information about the state than the channel carried.

The criterion is necessary and not sufficient, and the distinction matters more than the criterion. An audit sharing no factor with the model may still be too noisy to detect the property, or unable to compute it. Failing the criterion is disqualifying. Passing it guarantees nothing. And sufficiency, where it can be argued at all, is indexed on a named property, which yields the concession that costs the most and is worth what it costs: *there is no property-agnostic audit*. An audit is always the audit of an enumerated list.

The operational corollary is where this stops being an argument and becomes a specification. A validity port should not expose a proximity. It should expose which properties of the object survive the featurization. For either epimer of section 1, such a port would report five stereocenters present and zero encoded, without running the model, without a reference corpus, without calibration, without a test set. That is a property of the model, not of the query. It is deterministic, and it is computed by reading the call graph.

The objection to a port is regress. If it must read a representation to report what another lost, what audits the port. The answer is positional. It needs no richer representation, it needs an earlier one. The parser still holds the five stereocenters at the moment the featurizer declines to read them, so the port compares the parser's object against the featurizer's computation path and reports the difference. It is not plugged in beside the model. It is plugged in before it.

The corollary also dissolves the composition problem rather than arguing with it. If a port exposes which properties survive a featurization, and each routed model has its own featurization, then ports are per model by construction. There is no single domain to compute, because there was never a single question to answer.

Which is how the reading behind this article was done.

7. Domain of validity

The thesis holds for applicability domains whose computation descends from a factor they share with the model and which destroys the property at issue. It does not hold for an audit that retains information the model discards. The refuting observation is nameable: an applicability domain computed on three-dimensional, quantum or geometric descriptors, therefore retaining what the model's featurization removes, and correctly flagging the unencodable. It would bound this thesis, which is the intended effect.

The lemma of non-veto holds for weighted composite scores with a threshold, not for single-measure domains: leverage, Mahalanobis distance, range-based methods. Those have other problems, not this one.

The validity port falls under its own principle, and the concession is owed. It reports the featurization's losses relative to what the parser retained, and the parser is an abstraction too, having already discarded conformation, protonation state, solvation. A port cannot enumerate what nothing upstream of it recorded. It is not richer than the model. It is earlier, and earlier is all the criterion asks.

Nothing here asserts that ToxTwin is wrong about 17β-estradiol. It asserts that the experimental divergence is real and documented, that the device returns one number, and that it is therefore wrong about one of the two without being able to say which. Nothing here asserts anything about metals, and nothing here claims a general impossibility of self-knowledge: the obstacle is positional, it concerns where the control is plugged in, and that is why it is repairable.

On provenance, since the subject demands it, and the demand cuts here too. Every figure was measured or read on 15 July 2026, against a running service and its source. The evidence then splits, and the split should be visible rather than smoothed. One half any reader reproduces: the fingerprint primitive is blind by default, the epimers collide at 1.0000 under it and separate at 0.8182 without it, the connectivity block is shared by all three states, and the service returns one number for all three. The other half is reported, read in source not published here: that two of the three terms descend from one vector, that the control is computed in a latent space serving no prediction, that the reference corpus carries two of the fourteen endpoints. The thesis rests on the first half. The second explains the mechanism, and a reader who declines it keeps the fact and loses the reason, which is a smaller article and not a different one.

One quantity remains unmeasured: the gap between the two epimers, which the secondary literature reports inconsistently and which only the primary source settles. It is also the one quantity the argument does not need.

An instrument that cannot tell you what it failed to see will tell you that it saw nothing wrong.

References

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