

From QSAR models to regulatory evidence: how OECD and ICH frame computational toxicology

A practical view of how OECD validation principles and ICH M7 shape the use of QSAR models as defensible regulatory evidence, with emphasis on applicability domain, complementary methodologies and evidence integration.

PharmaToxAI · 2026-09-10

Computational toxicology has moved well beyond its original role as a rapid screening tool. In several regulatory settings, quantitative structure–activity relationships, read-across strategies and broader New Approach Methodologies now contribute to formal evidence packages. The important shift is not that prediction has replaced experiment, but that computational evidence is increasingly expected to be transparent, traceable and defensible for a defined decision context.

For PharmaToxAI, this distinction is central. The value of a QSAR model is not determined solely by predictive performance. A model can achieve strong cross-validation metrics and still be difficult to defend if the endpoint is poorly defined, the training domain is unclear, the prediction falls outside that domain, or the model cannot be documented in a way that supports expert review. OECD and ICH guidance approach this problem from different regulatory angles, but the underlying logic is consistent: computational predictions become useful regulatory evidence only when their strengths and limitations are explicit.

OECD: from model performance to regulatory credibility

The OECD framework for regulatory QSAR has long emphasized that model validation must address more than numerical accuracy. Its guidance identifies five core principles: a defined endpoint, an unambiguous algorithm, a defined applicability domain, appropriate measures of goodness-of-fit, robustness and predictivity, and, where possible, a mechanistic interpretation.

These principles are intended to support regulatory assessment rather than to prescribe a single modeling technology. That distinction matters because the regulatory question is rarely “which algorithm performs best?” The practical question is whether a given prediction can be considered reliable for a specific chemical, endpoint and purpose.

The OECD (Q)SAR Assessment Framework (QAF) makes this operational by providing a systematic structure for assessing individual models, individual predictions and results that combine multiple predictions. The second edition, published in 2024, added a dedicated reporting format for results based on multiple predictions, reflecting the growing importance of transparent evidence integration rather than single-model outputs.

This becomes particularly relevant in applied toxicology. A numerical prediction is rarely sufficient on its own. A defensible assessment should make clear whether the target compound is represented by the chemistry used to train the model, whether relevant structural features are covered, whether the prediction is interpolative or extrapolative, and how uncertainty has been handled.

The OECD QSAR Toolbox supports this broader logic by combining profilers, databases, grouping strategies and read-across workflows. In the most recent OECD Guidance on Grouping of Chemicals, published in 2025, the scope expands further to include New Approach Methodologies such as adverse outcome pathways, omics, high-throughput screening and QSAR models as tools for developing groups and substantiating similarity. The regulatory value therefore lies increasingly in structured evidence integration, not in isolated predictions.

In practical terms, the question changes from:

| Does the model predict toxicity?

into something closer to:

| Is this prediction scientifically defensible for this chemical, endpoint and regulatory decision?

That is a substantially higher standard.

ICH M7: a concrete example of QSAR in pharmaceutical regulation

ICH M7(R2) provides one of the clearest pharmaceutical examples of QSAR being used as an explicit part of a regulatory assessment strategy. The guideline addresses DNA-reactive mutagenic impurities in pharmaceuticals and defines how mutagenic risk should be assessed and controlled.

For impurities without adequate mutagenicity data, the guideline calls for two complementary QSAR methodologies: one expert rule-based and one statistical-based. The models should follow the general OECD validation principles. The purpose of using complementary methodologies is not simply redundancy; it is to reduce dependence on the failure modes of a single modeling paradigm.

ICH M7 also establishes a clear hierarchy between computational and experimental evidence. When both complementary QSAR approaches show no structural alerts within an acceptable assessment context, the impurity can be classified as having no mutagenic concern. If a relevant alert is present, additional expert review, control measures or bacterial mutagenicity testing may be warranted. A properly conducted negative bacterial mutagenicity result can overrule the structure-based concern.

The accompanying ICH M7(R2) Questions and Answers provide further detail on out-of-domain and non-coverage cases. An out-of-domain result from one of the two required QSAR models does not automatically justify a non-mutagenic classification. Additional assessment may include read-across to structurally similar analogues with experimental data, expert evaluation of DNA-reactivity potential, or use of another validated model of the same methodological class that produces an in-domain prediction.

This is important because it makes applicability domain a decision variable rather than a technical footnote.

Applicability domain is part of the result

Machine-learning software will usually generate a prediction for any molecule that can be converted into the required descriptors. That does not mean the prediction is scientifically meaningful.

Consider a model trained mainly on conventional drug-like organic chemistry. It may still return a numerical score for a structurally unusual compound that lies far from the training distribution. From a purely computational perspective the calculation succeeded; from a regulatory perspective the relevant issue is whether the result can be trusted.

A regulatory-grade workflow should therefore answer at least four questions:

1. Is the query compound represented within the chemical space used to develop the model?
2. Are the structural and mechanistic features relevant to the endpoint represented?
3. Is the prediction interpolative or extrapolative?
4. Is there a transparent characterization of uncertainty or model confidence?

The OECD framework treats applicability-domain assessment as part of prediction evaluation, and ICH M7 explicitly addresses what to do when one of the required models returns an out-of-domain or non-coverage result.

This has an important consequence for model selection. A model with slightly lower global benchmark performance may be more useful in regulatory work than a nominally stronger black-box model if the former offers better traceability, domain definition and interpretation of uncertainty.

Regulatory QSAR is closer to evidence integration than benchmark optimization

Academic machine learning often focuses on maximizing global performance metrics. Regulatory toxicology asks a different set of questions.

A useful model should be reproducible, chemically interpretable where possible, appropriately validated and linked to a defined decision. In that context, several elements become as important as algorithm selection:

- curated and traceable training data;
- explicit endpoint definitions;
- external or scaffold-aware validation strategies;
- applicability-domain assessment;
- uncertainty characterization;
- transparent handling of conflicting predictions;
- expert review;
- predefined criteria for escalation to experimental testing.

This is why classical methods such as Random Forest, gradient boosting, rule-based expert systems and other interpretable models remain highly relevant. Regulatory utility is not necessarily proportional to architectural complexity.

A deep neural network may outperform a classical QSAR model on a benchmark and still be less useful in practice if training-data provenance is unclear or if the reliability of an individual compound prediction cannot be assessed.

OECD and ICH point toward a broader evidence model

OECD and ICH operate in different regulatory domains, but their approaches show a useful convergence.

OECD increasingly places QSAR within grouping, read-across, adverse outcome pathway and NAM-based evidence frameworks. ICH M7 demonstrates how computational predictions can become an operational part of pharmaceutical risk assessment when the methodology, applicability domain and decision rules are explicit.

The common principle is that computational evidence gains value when it is contextualized.

A modern toxicology workflow should therefore not stop at:

```
structure → model → prediction
```

A more defensible sequence is:

```
structure
  ↓
chemical characterization
  ↓
complementary computational models
  ↓
applicability domain + uncertainty
  ↓
mechanistic and expert interpretation
  ↓
experimental evidence when required
  ↓
integrated assessment
```

This is much closer to the direction of contemporary regulatory science.

Implications for platforms such as PharmaToxAI

A platform intended to support early toxicology or preclinical decisions should be designed around traceability rather than prediction alone.

For each result, a useful technical report should ideally document:

- the model and version used;
- the endpoint being predicted;
- the relationship between the query compound and the training domain;
- relevant structural alerts;
- agreement or disagreement among complementary models;
- uncertainty or confidence;
- supporting mechanistic or experimental evidence;

- the conditions under which additional testing is warranted.

This architecture is compatible with both the OECD validation framework and the operational logic seen in ICH M7.

The practical consequence is significant. A QSAR platform should not behave like a calculator that returns a toxicity score. It should behave more like an evidence system that explains why a prediction can—or cannot—be trusted.

Where the field is moving

The regulatory direction is increasingly toward combinations of computational models, mechanistic information and human-relevant experimental methods. The OECD 2025 grouping guidance explicitly incorporates NAMs such as omics, high-throughput screening, adverse outcome pathways and QSAR into the same evidentiary landscape.

The challenge over the next few years will therefore not simply be to build more accurate models. It will be to make computational toxicology auditable, reproducible and sufficiently transparent to support real decisions.

For drug development, this points toward deeper integration between predictive toxicology and experimental workflows rather than a simple replacement model.

A strong computational workflow should be able to answer not only:

| What does the model predict?

but also:

| Why should this prediction be trusted, for this compound, in this decision context?

That second question is where regulatory QSAR is ultimately judged.

Primary regulatory sources

1. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. OECD Series on Testing and Assessment No. 69. OECD Publishing. <https://doi.org/10.1787/9789264085442-en>
2. OECD. *(Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions, Second Edition*. OECD Series on Testing and Assessment No. 405. OECD Publishing, 2024. <https://doi.org/10.1787/bbdac345-en>

3. OECD. *Quantitative Structure-Activity Relationships Project*. OECD Environment, Health and Safety Programme. <https://www.oecd.org/en/topics/sub-issues/assessment-of-chemicals/quantitative-structure-activity-relationships-project.html>
4. OECD. *OECD QSAR Toolbox*. <https://www.oecd.org/en/data/tools/oecd-qsar-toolbox.html>
5. OECD. *Guidance on Grouping of Chemicals, Third Edition*. OECD Series on Testing and Assessment No. 418. OECD Publishing, 2025. <https://doi.org/10.1787/b254a158-en>
6. ICH. *M7(R2): Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk*. https://database.ich.org/sites/default/files/ICH_M7%28R2%29_Guideline_Step4_2023_0216_0.pdf
7. ICH M7 Implementation Working Group. *M7(R2) Questions and Answers*. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-m7r2-questions-and-answers-assessment-and-control-dna-reactive-mutagenic-impurities-pharmaceuticals-limit-potential-carcinogenic-risk-step-5_en.pdf

Online edition: <https://pharmatoxai.ar/whitepapers/qsar-regulatory-evidence-oecd-ich/>