Beyond the ADMET leaderboard: transferability, chemical space and evidence that survives new chemistry

Modern ADMET models can perform well on retrospective benchmarks yet fail when medicinal chemistry moves into new chemical space. Recent work on out-of-distribution validation, quantum-mechanical descriptors, uncertainty calibration and molecular language models suggests a more useful standard: characterize where a model works, where it extrapolates and how much confidence a decision should place in its output.

PharmaToxAI · 2026-09-28

Machine-learning models for absorption, distribution, metabolism, excretion and toxicity (ADMET) are now easy to build and increasingly difficult to evaluate honestly. Public benchmarks, pretrained molecular encoders and automated modelling pipelines can produce impressive performance tables. The harder question appears one step later: what happens when the next medicinal chemistry series does not resemble the molecules on which the model was trained?

That question is not a statistical technicality. ADMET models are normally deployed precisely because the compounds of interest have not yet been measured. A useful model must therefore do more than interpolate within a familiar dataset. It must support decisions when chemistry, assay history or the distribution of the endpoint changes. Recent studies are converging on an uncomfortable result: conventional validation can substantially overstate that ability.

The practical consequence is a change in emphasis. Model architecture still matters, but **transferability, applicability domain, calibration and explicit characterization of distribution shift are becoming as important as aggregate predictive accuracy**. This white paper examines that shift using recent primary evidence, including work on out-of-distribution (OOD) molecular prediction, quantum-mechanical descriptors, chemical language models and uncertainty calibration. It also considers how these developments relate to emerging regulatory thinking about computational evidence.

The regulatory connection needs to be stated carefully. Current FDA guidance on New Approach Methodologies (NAMs) does not qualify generic ADMET machine-learning models for regulatory use, and ICH M15 is not a QSAR guideline. What these documents do provide is a broader evidentiary logic: a model should be evaluated for a defined purpose, its limitations should be characterized, and the strength of evidence required should reflect the consequence of the decision it supports.

The benchmark problem

A molecular property model is usually trained and evaluated by partitioning an existing dataset into training, validation and test subsets. If those subsets are randomly sampled, closely related analogues can appear on both sides of the split. The resulting test therefore measures a mixture of memorization, local interpolation and genuine generalization.

Scaffold splitting was introduced partly to reduce this problem. Molecules are separated according to Bemis–Murcko scaffolds so that core structures in the test set differ from those used for training. Scaffold splits are now common in molecular machine-learning benchmarks and are frequently described as an approximation to prospective generalization.

Recent evidence shows that this approximation can still be weak.

Fooladi and colleagues trained more than 11,000 models across classical machine-learning, graph neural-network and pretrained graph approaches while comparing ten splitting strategies. Their analysis showed that scaffold and generic-scaffold splits were among the least challenging splits according to the chemical-distance measures they examined. Structurally similar compounds can possess different formal scaffolds, allowing chemically close examples to cross the train–test boundary. Across model families, harder distribution shifts produced similar relative degradation: modern graph models did not eliminate the underlying extrapolation problem.

This finding is important because it separates two questions that are often conflated:

1. **Can an algorithm fit this endpoint well?**
2. **Does the evaluation reproduce the chemical shift expected when the model is used?**

A high score answers the first only under the distribution defined by the split. It does not establish the second.

A 2026 cross-industry preprint led by Seal makes the deployment problem explicit. Rather than recommending a universally “best” splitter, the authors argue that validation should begin with the intended deployment scenario. Their framework tests chemical-series extrapolation, temporal drift, target-value extrapolation, local structure–

activity relationship discontinuities and representation sensitivity. In their ADME case study, failure arose from several distinct mechanisms rather than from one generic form of OOD shift.

This is a more useful way to think about applicability domain. A molecule is not simply “inside” or “outside” a model because its fingerprint similarity crosses an arbitrary threshold. It can be familiar in one dimension and unfamiliar in another: scaffold, physicochemical range, stereochemistry, ionisation state, endpoint value, assay provenance or mechanism.

Applicability domain should be treated as a map, not a gate

The classical QSAR concept of an applicability domain remains valuable, but contemporary molecular ML makes its implementation more complicated.

For a fingerprint model, one may estimate proximity to training chemistry using maximum Tanimoto similarity, leverage or nearest-neighbour distance. For a graph neural network or molecular language model, latent-space distances can also be calculated. None of these quantities is automatically equivalent to prediction reliability.

Three layers should therefore be separated.

Chemical support asks whether relevant structural and physicochemical features are represented in the training set. This can be examined with fingerprints, scaffolds, descriptors, clustering, local density and nearest-neighbour analysis.

Endpoint support asks whether the model has learned the part of the response surface relevant to the prediction. A molecule can be structurally close to training examples while occupying a sparse or discontinuous region of the activity landscape. Activity cliffs are an obvious example.

Predictive reliability asks how model error behaves conditional on those forms of support. This requires empirical calibration rather than assuming that distance is uncertainty.

For operational ADMET modelling, reporting a single applicability-domain flag discards useful information. A more informative output is a structured profile: nearest training analogues, similarity distribution, scaffold novelty, descriptor-range violations, local label density, ensemble disagreement or calibrated interval, and any known assay-domain limitations.

That profile also makes expert review more meaningful. Instead of asking whether a model “trusts” a molecule, the reviewer can inspect why the prediction is supported or weakly supported.

Quantum-mechanical descriptors: useful information when structure alone is not enough

One route to better transferability is to enrich the molecular representation with features that describe electronic behaviour rather than relying entirely on structural patterns.

Bose and colleagues tested this idea in four ADMET endpoints: human liver microsomal stability, permeability, solubility and hERG inhibition. Their 2026 study compared structural and quantum-mechanical (QM) descriptors both within familiar chemical space and under extrapolation. Adding QM information improved prediction across the four endpoints in their experiments, with particularly notable transferability gains for permeability and hERG. QM features also remained useful when training data were reduced.

The mechanistic interpretation is plausible. Permeability, ionisation-dependent behaviour, intermolecular interactions and channel binding are influenced by electronic properties that may not be fully represented by a finite set of structural fragments. A descriptor derived from electronic structure can, in principle, remain informative when a novel scaffold expresses familiar physicochemical behaviour through a different structural motif.

But the result should not be generalized beyond the experiment. It does not establish that QM descriptors universally improve ADMET prediction or solve OOD generalization. Their calculation introduces cost and methodological choices involving conformers, protonation states, level of theory and aggregation of conformational properties. A QM descriptor can also become another correlated feature if the dataset does not contain enough information to exploit it.

The practical implication is narrower and more useful: **representation should be tested under the type of shift expected at deployment**. If a representation improves random or conventional scaffold validation but provides no gain when chemical similarity is deliberately reduced, it has not demonstrated improved extrapolation.

Molecular language models change representation, not the validation problem

Large molecular encoders offer a different strategy. Instead of manually selecting descriptors, they learn representations from SMILES strings or molecular graphs and can then be fine-tuned for property prediction.

Lim and colleagues reported a 2026 DeBERTa-based SMILES encoder trained in a multi-task setting across 22 ADMET endpoints. The objective was to internalize ADMET-relevant structure–property information while retaining molecular syntax and structural

information. This type of multi-endpoint representation is attractive because ADMET properties are not independent: lipophilicity, solubility, permeability, metabolic stability and several safety liabilities share underlying structural and physicochemical determinants.

However, representation learning does not remove the need for deployment-oriented validation. A pretrained encoder can improve sample efficiency and average performance while remaining confidently wrong on chemistry poorly represented by its pretraining or fine-tuning data.

The same point is illustrated from another direction by M-JEPA, a 2026 self-supervised graph framework evaluated on Tox21. Its predictive self-supervised objective improved both discrimination and calibration relative to the matched supervised baseline under the authors' scaffold-split protocol. That is useful evidence for the representation-learning strategy. It is not evidence that scaffold splitting itself reproduces every prospective deployment scenario.

The distinction is fundamental:

> Better molecular representations can improve prediction under distribution shift, but only an evaluation designed around distribution shift can demonstrate that improvement.

Confidence is a separate modelling problem

ADMET systems increasingly attach confidence values to predictions. Unfortunately, classification probability, ensemble variance and distance to training data are often treated as interchangeable measures of confidence. They are not.

Calibration asks whether stated confidence corresponds to observed error frequencies. Applicability-domain analysis asks whether a prediction is supported by relevant training chemistry. Uncertainty estimation attempts to quantify what is not known. These quantities are related, but each can fail independently.

A 2026 study by Guo examined uncertainty intervals under scaffold shift across six human protein targets, 20 scaffold-disjoint partitions per target and two model classes. Raw random-forest ensemble intervals intended to provide 90% coverage covered only about 9.5–11.4% of molecules. Split conformal methods brought average marginal coverage much closer to the nominal level. Yet conditional performance remained weaker for molecules least similar to training data and for the highest-activity region.

That result illustrates a crucial limitation of global calibration. A model can be calibrated on average while remaining systematically overconfident in precisely the region that matters most.

The same study also found that using lower confidence bounds to rank molecules did not automatically improve selection. In its tested settings, risk-aware ranking traded activity enrichment against uncertainty in ways that could worsen the practical selection objective.

This distinction between **marginal calibration**, **local reliability** and **decision utility** should become standard in ADMET model reports.

A second 2026 study, MARS, treats toxicity prediction explicitly as a joint prediction-and-reliability problem. The model retrieves multiple training anchors and uses their similarities, labels and disagreement to generate a reliability signal separate from the toxicity score. Across scaffold, strict OOD and cross-dataset transfer evaluations, the authors report improved calibration and error-detection performance relative to comparison approaches.

MARS is one proposed architecture rather than a general solution, but its framing is valuable. Confidence should be evaluated against actual errors under realistic shift, not inferred from the aesthetic properties of a model output.

What a stronger ADMET validation protocol looks like

The emerging evidence suggests that a useful validation package should contain several complementary tests rather than one headline metric.

1. Define the deployment question before splitting the data

The expected use determines the relevant shift. A model used to prioritize analogues within an established series has a different extrapolation burden from one intended to screen an external vendor library or propose new scaffolds.

Possible deployment scenarios include:

- interpolation within a medicinal chemistry series;
- prediction of new scaffolds;
- prospective prediction of compounds synthesized later in time;
- transfer between organizations or assay platforms;
- extrapolation toward unusually high or low endpoint values;
- prediction of chemical classes sparsely represented in training.

The split should reproduce the intended scenario as closely as the available data permit.

2. Preserve a simple baseline

Random forests or gradient-boosted trees on Morgan fingerprints and conventional physicochemical descriptors remain important controls. A more complex architecture should demonstrate a meaningful advantage over these baselines under the relevant deployment split, not merely on a convenient benchmark.

If model rankings collapse or reverse under harder shifts, that is scientific information rather than an inconvenience to be hidden.

3. Quantify chemical-space separation

A split name is not enough. Report the actual relationship between train and test chemistry.

Useful diagnostics include maximum training-set Tanimoto similarity, nearest-neighbour distributions, scaffold overlap, descriptor-space distances, cluster membership and visualizations of chemical-space coverage. The purpose is not to reduce domain analysis to a single similarity metric, but to show how difficult the test actually is.

4. Measure calibration as well as accuracy

For classification, AUROC or AUPRC should be accompanied by calibration metrics and reliability plots when probabilities are used for decisions. For regression, residual distributions and calibrated prediction intervals can be more informative than RMSE alone.

Calibration should be stratified by chemical similarity, endpoint range or other relevant domains. Good average calibration does not imply good local calibration.

5. Test the tails

Drug-discovery decisions often focus on the tails of a distribution: the most soluble compounds, the strongest inhibitors, the longest-lived molecules or the candidates predicted to avoid a toxicity liability.

A model can have acceptable global RMSE while performing poorly in those tails. Target-value extrapolation should therefore be tested explicitly when the model will be used for optimization.

6. Report failure modes

Applicability-domain failures, activity cliffs, stereochemical ambiguities, unusual charge states, assay inconsistencies and sparse chemical classes should be visible in the report.

This is particularly important for negative predictions. In safety assessment, “predicted inactive” and “supported prediction of inactivity” are not equivalent statements.

Regulatory relevance: evidence quality rather than algorithm preference

Regulators are not currently prescribing a preferred molecular machine-learning architecture for general ADMET prediction. The more relevant development is that regulatory frameworks are becoming increasingly explicit about the relationship between **context of use, validation and decision consequence**.

FDA's March 2026 draft guidance, *General Considerations for the Use of New Approach Methodologies in Drug Development*, defines NAMs broadly enough to include in silico approaches. Its proposed validation framework emphasizes four features: context of use, human biological relevance, technical characterization and fitness for purpose. The guidance also states that a NAM can be useful within a weight-of-evidence assessment even when it has not undergone formal qualification, provided its suitability for the specific use is adequately established.

This does not mean that an ADMET ML benchmark satisfying these concepts becomes regulatory evidence automatically. The FDA document is a draft guidance focused on NAMs submitted in drug development, not a certification scheme for QSAR software.

The useful connection is conceptual. A computational model intended to support a consequential decision should state what decision it supports, characterize the evidence relevant to that decision and demonstrate performance under conditions that resemble the proposed use.

ICH M15, finalized at Step 4 in January 2026 and implemented by major regulators during 2026, provides a parallel framework for model-informed drug development. Its scope is broader than ADMET QSAR and includes pharmacometric and mechanistic modelling. Nevertheless, its emphasis on model purpose, risk, evaluation and documentation reinforces the same direction: the credibility of a model is conditional on what is being asked of it.

For QSAR specifically, the established OECD validation principles remain directly relevant. A defined endpoint, unambiguous algorithm, defined applicability domain, appropriate measures of fit/robustness/predictivity and mechanistic interpretation where possible are still a useful foundation. Modern ML does not make these principles obsolete. It makes their implementation harder.

From model cards to evidence packages

For practical predictive toxicology, the output of an ADMET pipeline should therefore be more than a CSV containing a predicted value.

A defensible prediction package can include:

- standardized input structure and protonation/stereochemical assumptions;
- model and training-data version;
- endpoint definition and assay provenance;
- prediction and calibrated uncertainty;
- nearest relevant training analogues;
- chemical-space and applicability-domain diagnostics;
- known representation or assay limitations;
- evidence from complementary models or experimental systems;
- an explicit statement of intended use.

For early discovery, much of this information can be generated automatically. The goal is not to impose regulatory documentation on every screening calculation. It is to preserve enough provenance that a prediction can later be interrogated rather than treated as an unexplained number.

This also changes how computational and experimental methods can be combined. A model that identifies a compound as both high-risk and poorly supported by training chemistry is a rational candidate for targeted experimental follow-up. Conversely, a prediction supported by several close analogues, consistent orthogonal models and a well-characterized assay domain may justify a different experimental priority.

The computational model is then functioning as part of an evidence-generation strategy rather than as a replacement for measurement.

Implications for model development

Several practical conclusions follow from the recent literature.

First, **benchmark improvement is not synonymous with deployment improvement**. New encoders, graph architectures and descriptors should be evaluated under splits that reproduce the intended chemical shift.

Second, **applicability domain is multidimensional**. Fingerprint similarity is useful, but it should not be mistaken for uncertainty or used as the sole criterion for model acceptance.

Third, **uncertainty must be validated**. A confidence score that has not been tested against errors under shift is another model output, not evidence of reliability.

Fourth, **physically meaningful information may improve transferability**. The recent QM-descriptor results provide a concrete example, particularly for endpoints

where electronic interactions are directly relevant. Whether the added computational cost is justified remains endpoint- and workflow-dependent.

Fifth, **pretraining and multi-task learning are promising but do not relax validation requirements**. Better representation and better validation solve different problems.

Finally, **model documentation should be designed around decisions**. The relevant question is not whether a model is “AI”, interpretable, mechanistic or state of the art in isolation. It is whether its evidence is adequate for the specific decision being made.

Conclusions

The central weakness in many ADMET modelling workflows is no longer the absence of powerful algorithms. It is the gap between retrospective performance and prospective use.

Recent studies show that common scaffold splits can provide weaker distribution shifts than expected, that calibration can fail dramatically outside familiar chemistry, and that even calibrated uncertainty can deteriorate in the least-supported regions of chemical space. At the same time, new representations are producing useful advances: quantum-mechanical descriptors can improve transferability for selected ADMET endpoints, and molecular language or self-supervised graph models can learn richer cross-endpoint representations.

These findings are compatible rather than contradictory. Better representations are valuable, but they need harder tests.

For PharmaToxAI, the practical standard is therefore straightforward: a predictive model should be accompanied by an explicit description of **where its evidence comes from, how far the query compound lies from that evidence, how uncertainty behaves under comparable shifts, and what decision the prediction is intended to support**.

That standard is more demanding than a leaderboard. It is also much closer to the way predictive toxicology and ADMET modelling are actually used.

References

1. Bose A, Anselmetti GLR, Degroote M, Moll N, Santagati R, Streif M, Weskamp N, Tkatchenko A. Improving the Stability and Transferability of Effective ADMET Models by Adding Quantum Mechanical Descriptors. *Journal of Chemical Information and Modeling*. 2026;66(5):2488–2500. doi: 10.1021/acs.jcim.5c02491.

2. Fooladi H, Vu TNL, Mathea M, Kirchmair J. Evaluating Machine Learning Models for Molecular Property Prediction: Performance and Robustness on Out-of-Distribution Data. *Journal of Chemical Information and Modeling*. 2025;65(19):9871–9891. doi: 10.1021/acs.jcim.5c00475.

3. Seal S, Zalte AS, Araripe DA, et al. Model Validation Protocols for Machine Learning in Small Molecule Drug Discovery. *bioRxiv*. Posted August 24, 2026. doi: 10.64898/2026.08.19.745868.

4. Lim JH, Kim M, Han Y, Lee JY. Improving predictive performance for molecular ADMET properties using a chemical language model. *Bulletin of the Korean Chemical Society*. 2026;47(6):756–768. doi: 10.1002/bkcs.70177.

5. Iyer K. M-JEPA: Predictive Self-Supervised Learning for Molecular Graphs with Scaffold-Shift Evaluation on Tox21. *Journal of Chemical Information and Modeling*. 2026;66(14):8008–8021. doi: 10.1021/acs.jcim.6c00828.

6. Guo Y. Scaffold-shift uncertainty calibration in molecular activity prediction: A multi-target benchmark of coverage, efficiency, and risk-aware selection. *Computational Biology and Chemistry*. 2026. Article 109371. doi: 10.1016/j.compbiolchem.2026.109371.

7. Wu S, Zhang S, Ling Y, Wu X. MARS: Multi-anchor reasoning for reliable toxicity prediction under distribution shift. *Computational Biology and Chemistry*. 2026;124(Pt 2):109213. doi: 10.1016/j.compbiolchem.2026.109213.

8. U.S. Food and Drug Administration. *General Considerations for the Use of New Approach Methodologies in Drug Development: Draft Guidance for Industry*. March 2026. FDA guidance PDF.

9. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *ICH M15: General Principles for Model-Informed Drug Development*. Step 4 adopted January 29, 2026. ICH M15 via EMA.

10. Organisation for Economic Co-operation and Development. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. OECD Series on Testing and Assessment No. 69. OECD Publishing. OECD (Q)SAR project.

Online edition: https://pharmatoxai.ar/whitepapers/admet-transferability-chemical-space/