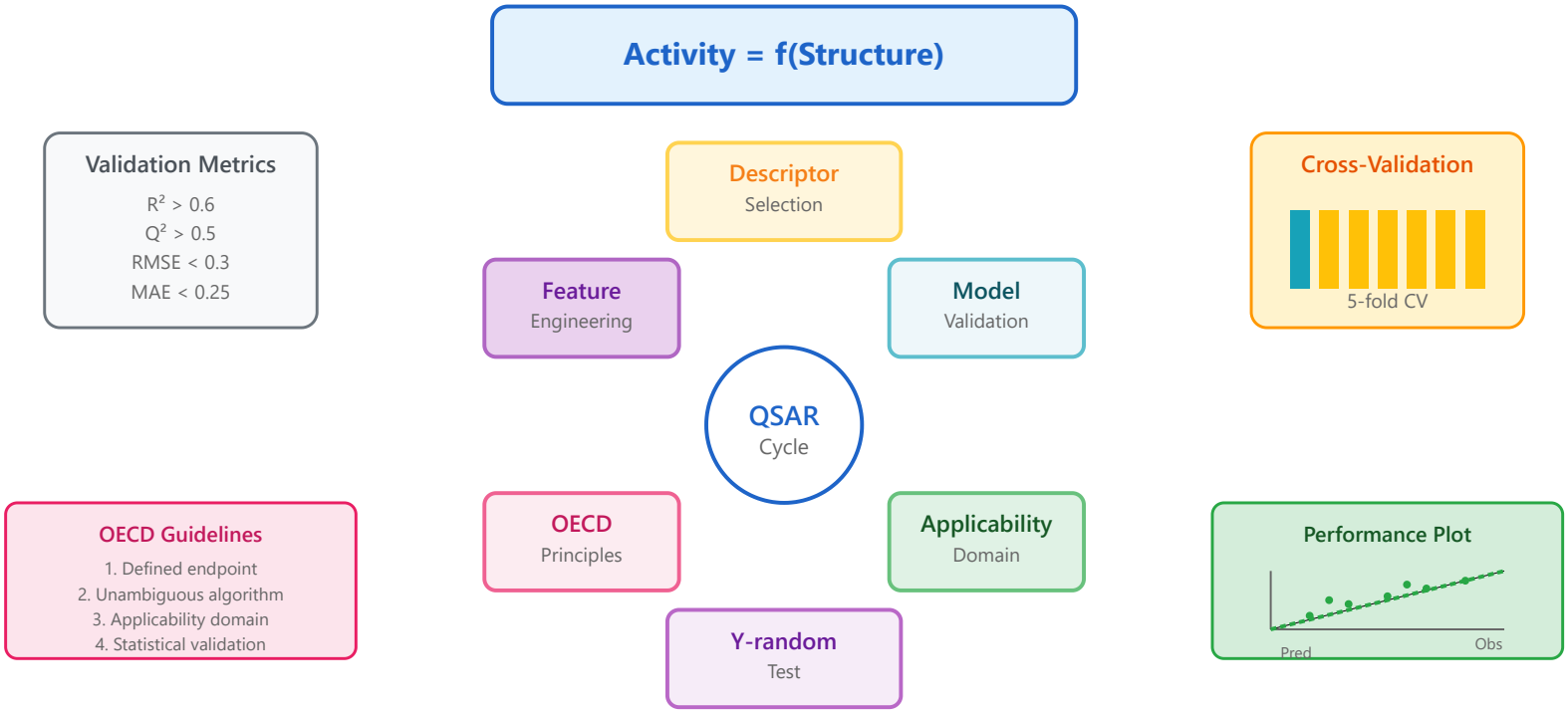
QSAR Modeling



Descriptor Selection

Molecular descriptors are numerical representations of chemical structures. Proper selection reduces dimensionality,

Model Validation

Statistical validation ensures model reliability through metrics like R^2 (goodness of fit), Q^2 (predictive ability), RMSE (error

removes redundant features, and identifies the most relevant structural properties that correlate with biological activity.

magnitude), and cross-validation. External test sets verify generalization to unseen compounds.

Applicability Domain

Defines the chemical space where predictions are reliable. Models should only predict compounds similar to training data. Distance-based, range-based, or probability-based methods identify out-of-domain structures.

Y-Randomization Test

Validates that model performance isn't due to chance correlation. Activity values are randomly shuffled while maintaining descriptors. A valid model shows significantly worse performance with randomized data.

OECD Principles

International guidelines for QSAR validation: (1) defined endpoint, (2) unambiguous algorithm, (3) defined applicability domain, (4) appropriate goodness-of-fit measures, and (5) mechanistic interpretation when possible.

Feature Engineering

Transforms raw descriptors into more informative features through scaling, normalization, polynomial features, or domain-specific transformations. Improves model performance and interpretability of structure-activity relationships.