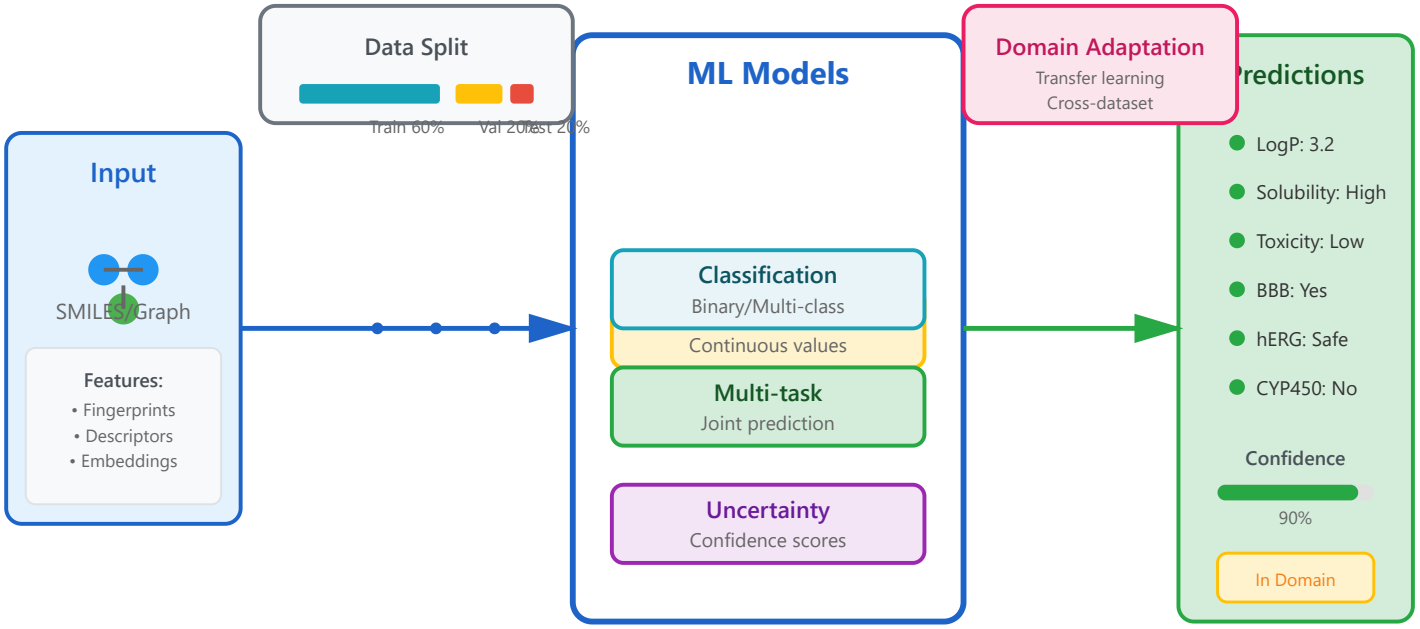
Property Prediction



Core Principles of Property Prediction

Molecular Representation

Feature Engineering

Converting chemical structures into machine-readable formats is the foundation of property prediction.

- **SMILES:** Text-based linear notation for molecules
- **Graph:** Atoms as nodes, bonds as edges
- **Fingerprints:** Binary vectors encoding structural features
- **Descriptors:** Calculated physicochemical properties

Extracting relevant molecular features that correlate with target properties.

- **Structural features:** Functional groups, ring systems
- **Topological indices:** Molecular connectivity
- **3D descriptors:** Spatial arrangement of atoms
- **Learned embeddings:** Neural network representations

Model Selection

Choosing appropriate algorithms based on the prediction task and data characteristics.

- **Regression:** For continuous properties (LogP, solubility)
- **Classification:** For categorical outcomes (toxic/non-toxic)
- **Multi-task:** Predicting multiple properties simultaneously
- **Ensemble methods:** Combining multiple models for robustness

Model Validation

Rigorous evaluation ensures model reliability and generalization capability.

- **Data splitting:** Train/validation/test sets
- **Cross-validation:** K-fold for robust assessment
- **External validation:** Testing on independent datasets
- **Applicability domain:** Defining model's valid range

Uncertainty Quantification

Estimating prediction confidence helps in decision-making and risk assessment.

- **Prediction intervals:** Range of plausible values
- **Ensemble variance:** Disagreement between models
- **Conformal prediction:** Statistically valid intervals
- **Domain distance:** Similarity to training data

Transfer Learning

Leveraging knowledge from related tasks improves predictions with limited data.

- **Pre-trained models:** Using large molecular databases
- **Fine-tuning:** Adapting to specific target properties
- **Multi-task learning:** Sharing representations across tasks
- **Domain adaptation:** Bridging different chemical spaces

Key Performance Metrics

Regression Metrics

R^2: Coefficient of determination (0-1)
RMSE: Root mean squared error

Classification Metrics

Accuracy: Overall correct predictions
ROC-AUC: Area under ROC curve

Model Interpretability

SHAP: Feature importance values
Attention: Relevant molecular substructures

MAE: Mean absolute error

F1-Score: Balance of precision/recall

Saliency maps: Critical atoms/bonds

Property Prediction Workflow

1. Data Collection & Curation
2. Molecular Featurization
3. Model Training
4. Validation & Optimization
5. Prediction & Deployment

Advanced Concepts

Data Quality & Bias

The quality of predictions is fundamentally limited by training data quality.

- **Data curation:** Removing errors and duplicates
- **Chemical diversity:** Ensuring broad coverage
- **Activity cliffs:** Similar structures, different properties
- **Imbalanced data:** Addressing class imbalance

Deep Learning Architectures

Modern neural networks capture complex structure-property relationships.

- **Graph Neural Networks:** Direct processing of molecular graphs
- **Transformers:** Attention-based sequence models
- **Message Passing:** Information flow across molecular structure
- **3D Convolution:** Learning from spatial conformations

Explainability & Trust

Understanding model decisions is crucial for scientific applications.

- **Structural alerts:** Known toxicophores and pharmacophores
- **Feature attribution:** Which features drive predictions
- **Counterfactual analysis:** What changes affect outcomes

Practical Applications

Property prediction accelerates drug discovery and materials design.

- **ADMET prediction:** Early filtering of candidates
- **Virtual screening:** Prioritizing compounds for synthesis
- **Lead optimization:** Guiding structural modifications

- **Model debugging:** Identifying failure modes

- **Safety assessment:** Identifying potential liabilities