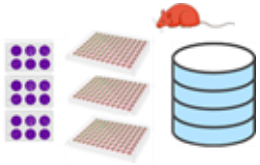
Leveraging Cell Painting Morphological Profiles for Machine Learning–Driven Bioactivity Prediction

Srijit Seal

Learning Objectives

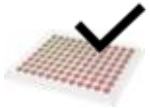
- Explore Cell Painting features and assay endpoints
- Standardize and prepare data for machine learning
- Train and tune a Random Forest classifier
- Evaluate model performance using metrics
- Visualize PCA, ROC curves, and feature importance
- Apply the workflow to different endpoints and interpret results

in vitro assays and *in vivo* endpoints as a proxy labels



in vivo and *in vitro* assays

1. Relevance to human toxicity
2. Mechanistic insights
3. Chemical space coverage
4. Translational potential



Proxy endpoint representing toxicity (e.g. hERG inhibition)

strong correlation

$\propto$

Human toxicity (e.g. Cardiotoxicity)

Evidence of absence vs **Absence of evidence**

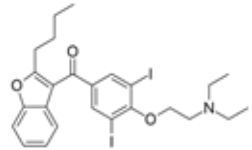
Cardiotoxicity

ML model

Semi-causal

- Blocks hERG?
- Reactive metabolites?
- Mitochondrial toxicity?
- Changes in morphology?
- C_{max} in human?

With Mechanistic Insights



Chemical Structure (e.g. Amiodarone)

ML model

Structural features

Cardiotoxicity

Without Mechanistic Insights

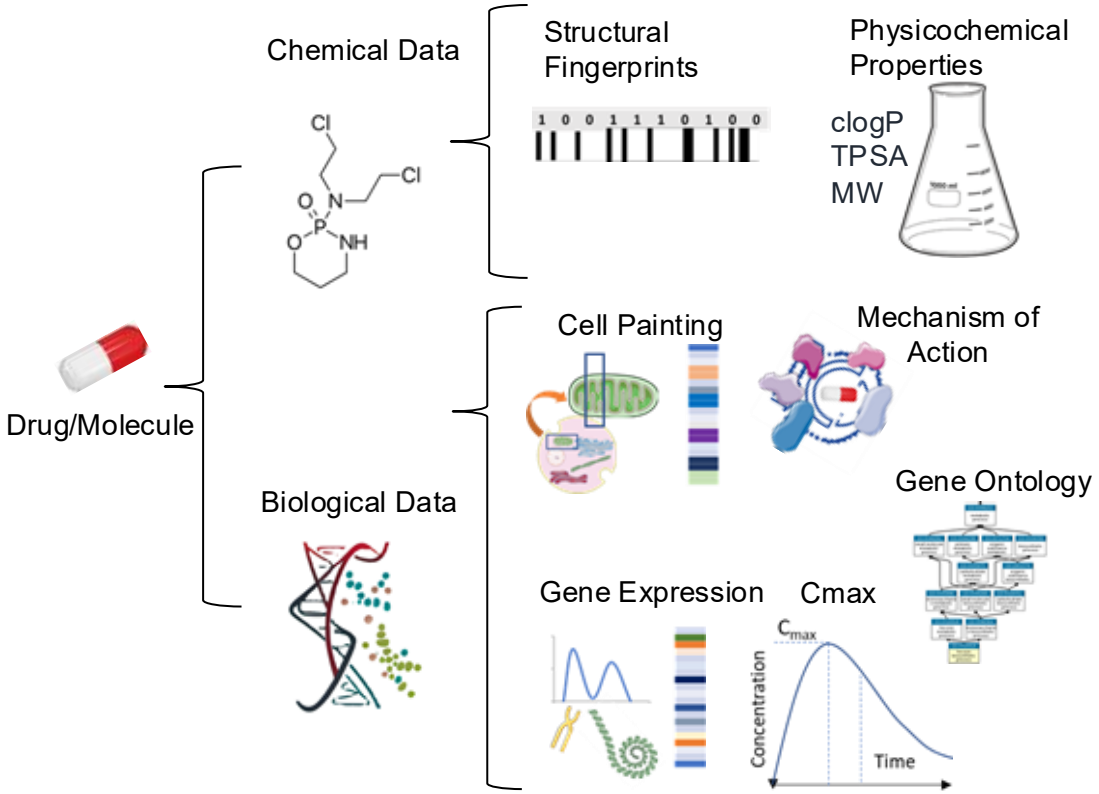
What do kind of descriptors do we need?

Mirrors **real-world drug discovery** challenges and practical drug development outcomes.

Preprocessed molecular features from a **wide range of modalities**, such as structural features, cell imaging, and gene expression.

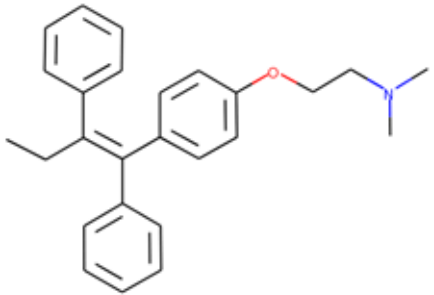
Out-of-distribution splits: Recommended dataset splits for validation.

Support models to **improve applicability**: Multitask and transfer learning models



Descriptors

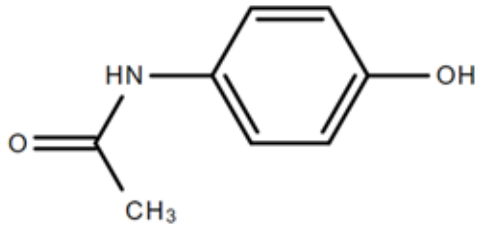
- Provide an *information-preserving* representation of input data (e.g. structures) for the model
- Either knowledge-based (e.g. reactive groups), or (usually) ‘trial and error’
- *Can* be learned from data, but only *if there is enough data, and we can meaningfully label!*



0100101010000…

Fingerprints, pharmacophores, surface properties, substructures/ functional groups, shapes, physchem properties *etc.*

Paracetamol



Paracetamol

Formula C8H9NO2

Identifiers

IUPAC name: N-(4-hydroxyphenyl)acetamide

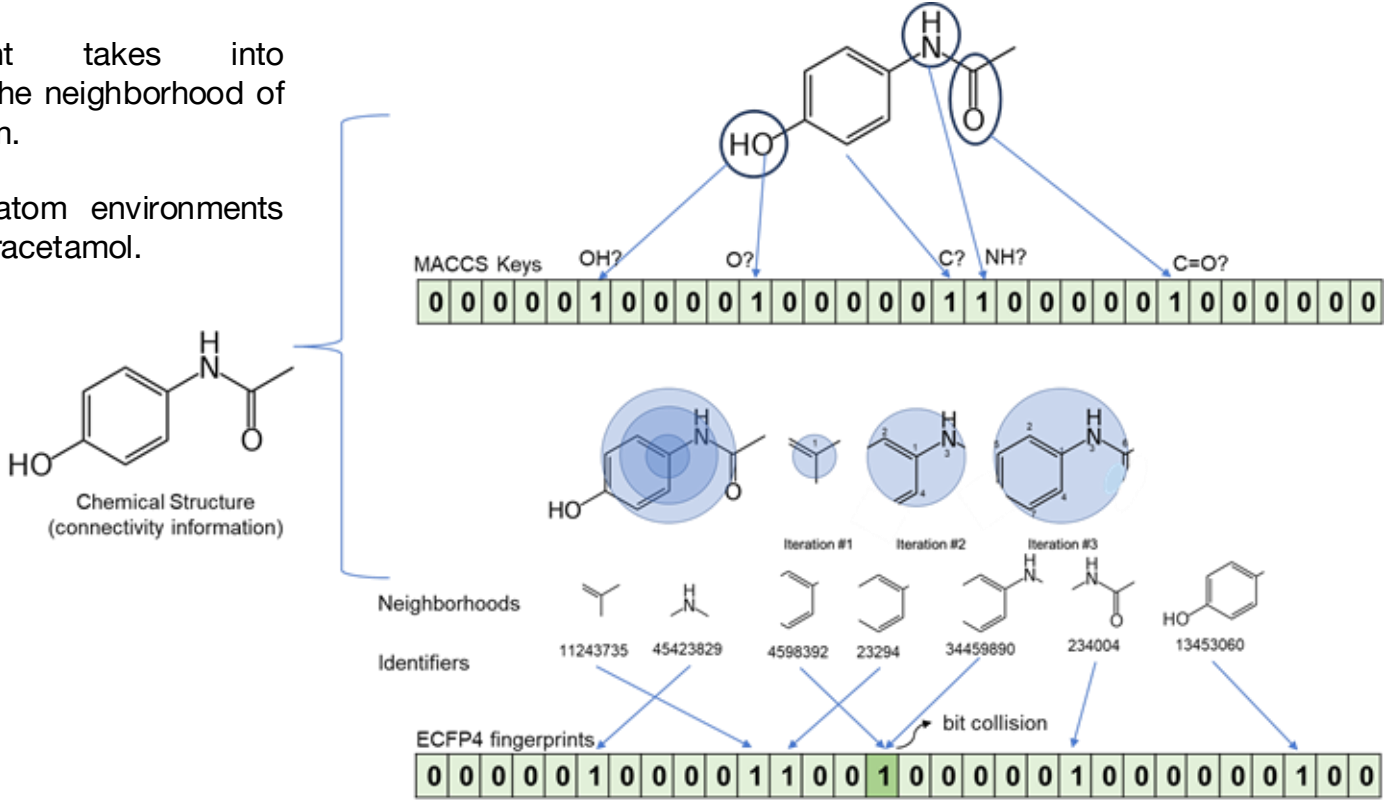
SMILES: C1=CC(O)=CC=C1NC(=O)C

chemical structure, molecular formula, SMILES identifier of the common anti-inflammatory drug paracetamol

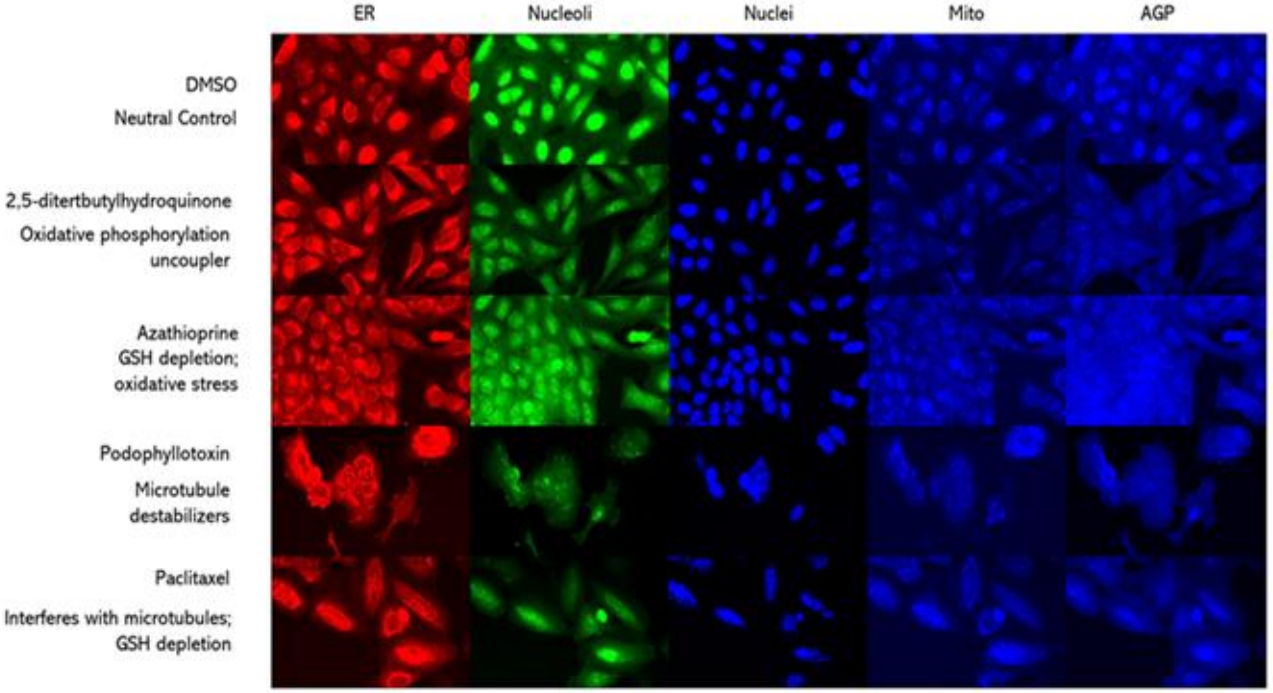
Morgan Fingerprints

Fingerprint takes into account the neighborhood of each atom.

Circular atom environments within Paracetamol.

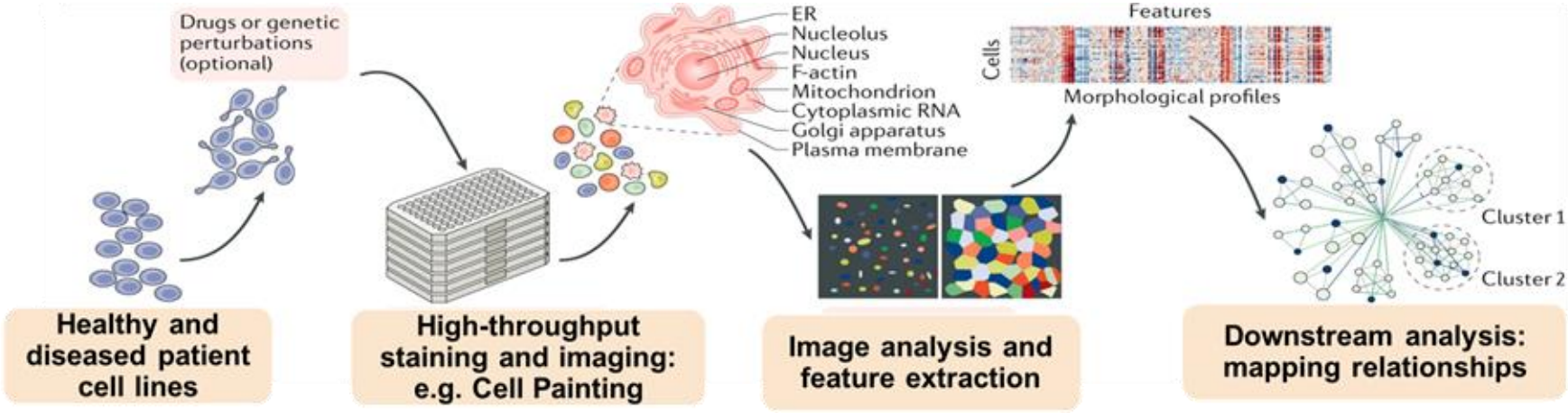


Imaging Assays to “Cell Painting”



Cell Imaging: Diverse phenotypes across compounds as captured in U2OS cells in the Cell Painting assay compared to neutral control DMSO.

-Omics descritors: Cell Painting



The Cell Painting assay stains eight cellular components using six dyes and is imaged in five channels. Thousands of hypothesis-free features extracted from these images from the Cell Painting dataset.

Understanding Cell Painting features: Mitotoxicity

Biological significance of Cell Painting features with respect to Mitochondrial Toxicity :

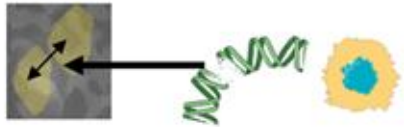
MITOCHONDRIAL FEATURES

Cells Intensity MaxIntensityEdge Mito
(PPV 0.83)

Edge of segmented object potentially indicates loss of membrane integrity

FEATURES FROM OTHER IMAGE CHANNELS

Cells Correlation Costes DNA AGP
(PPV 0.52)



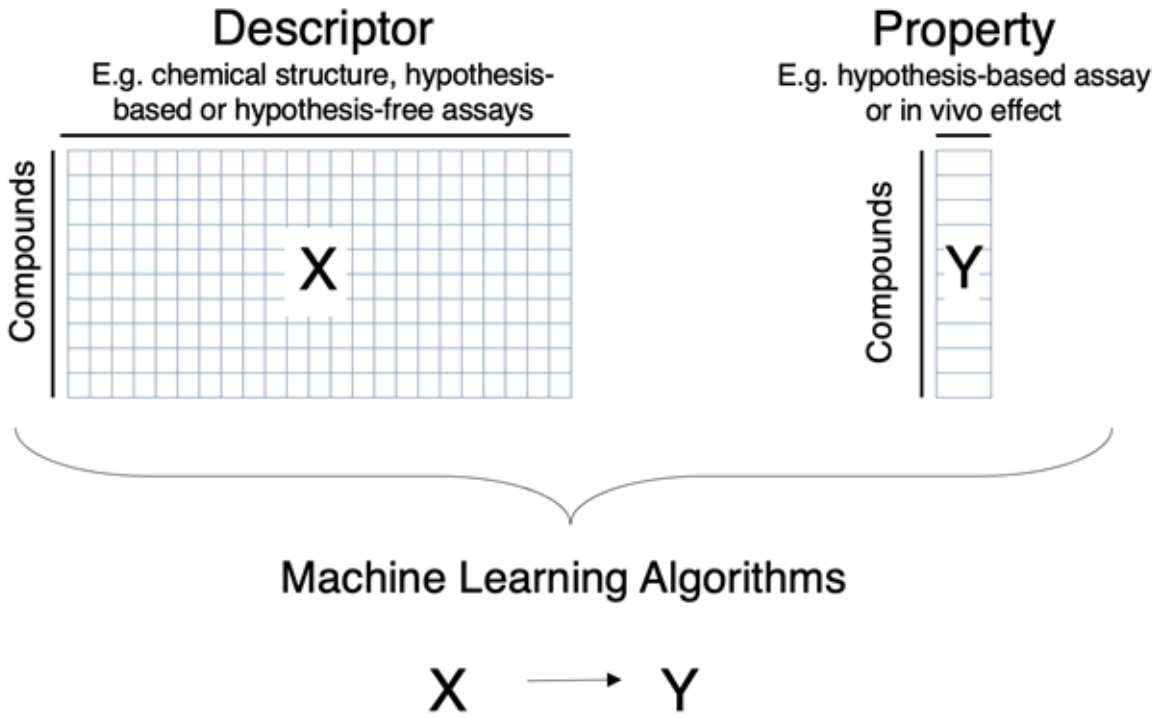
Potentially indicates DNA fragmentation and entering apoptosis or cell death

Two types of ‘modelling’

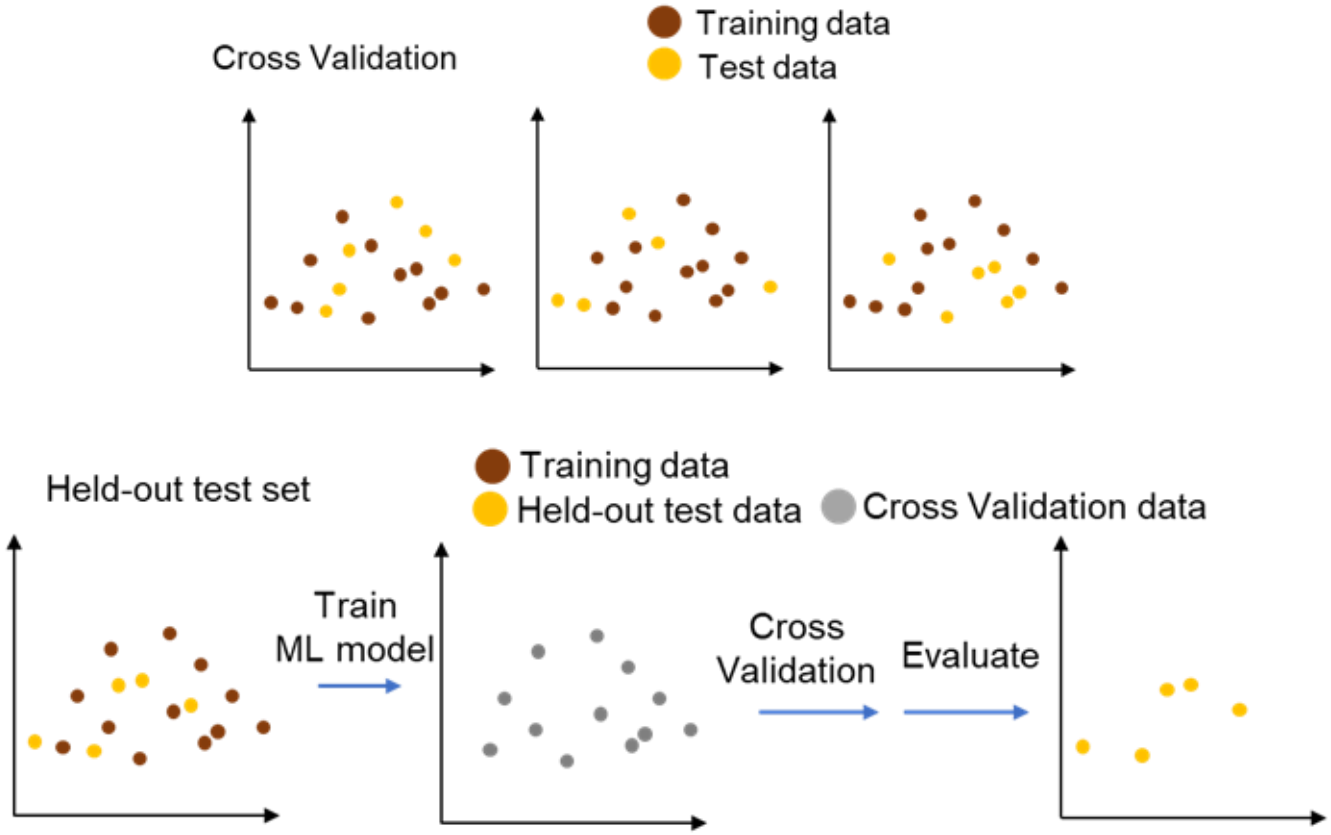
- *Model*-driven
  - We have an underlying, often mechanistic, model of the system we attempt to approach computationally
  - Requires understanding
  - Able to model novel systems (molecules etc.)

- *Data*-driven
  - We do *not* have an understanding, mechanistic model of the system
  - We measure data, model *empirically*
  - Does not require understanding of system
  - Able to model only space *captured by the data used for model generation*

Supervised Machine Learning



Supervised learning of properties (Y) from compound descriptors (X).


Cross Validation
Training data
Test data
Held-out test set
Training data
Held-out test data
Cross Validation data
Train
ML model
Cross
Validation
Evaluate

OCED Guidelines

1. Specific and well-defined **endpoint**.

2. Unambiguous **algorithm** clearly described for reproducibility.

3. Defined **applicability domain** for reliable predictions.

4. Appropriate goodness-of-fit, robustness, and **predictive power measures** on training and new data.

5. Mechanistic **interpretation**, where feasible, should provide insights into the mechanism of action.

Predicting Biological Activity from Cell Painting (CellProfiler-analysed) Features

This interactive tutorial shows how to use **machine learning** to predict biological assay outcomes from **Cell Painting imaging features**. You'll:

1. Explore the dataset
2. Choose a biological endpoint
3. Train a Random Forest model (manual or optimized)
4. Evaluate its performance with various metrics and plots

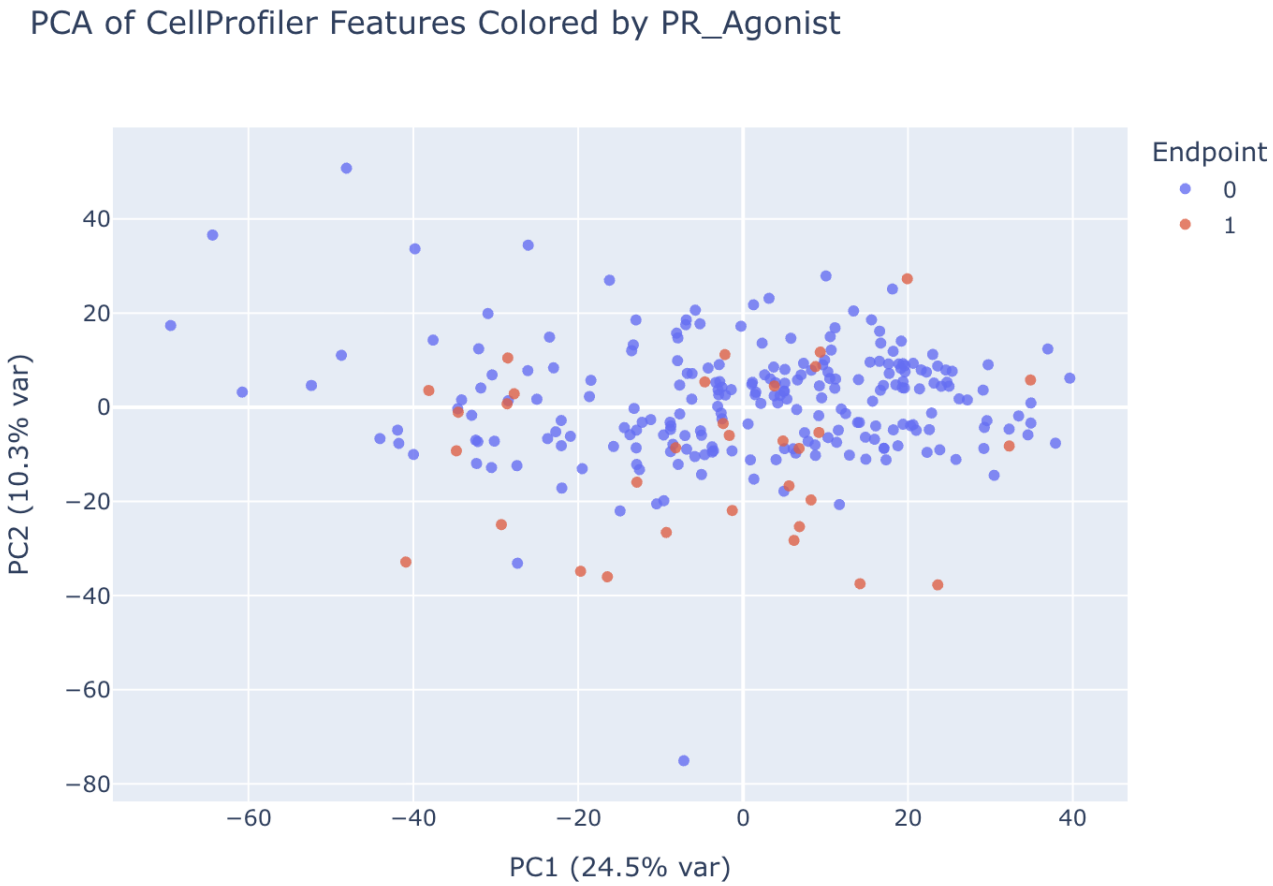
Step 1: Choose a Biological Endpoint

Here, you can select which assay you want to predict. Each assay represents a specific biological response, and our goal is to predict it from the Cell Painting features.

Select a biological endpoint: PR_Agonist

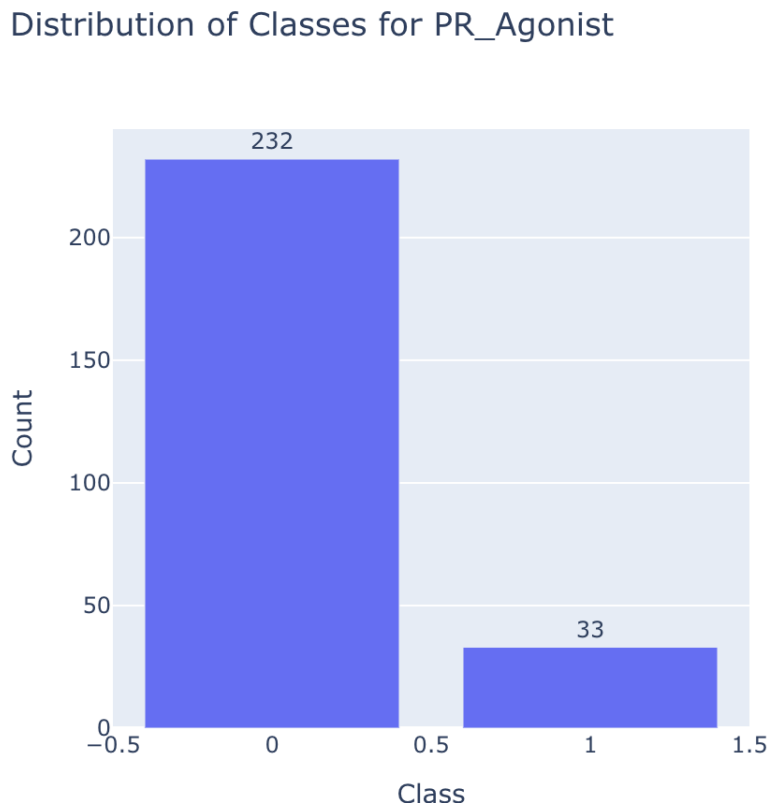
You selected the endpoint: PR_Agonist

Dataset Overview: 269 compounds × 1783 Cell Painting features



Step 2: Explore Class Distribution

It's important to check the balance of your classes (active vs inactive compounds). A very imbalanced dataset may require special considerations for modeling.



Step 3: Configure Random Forest Classifier

You can either:

- Let the app **automatically optimize** hyperparameters
- **Manually select** hyperparameters using sliders below

Random Forest is robust to overfitting and can handle many correlated features, making it ideal for Cell Painting data.

RandomForest Mode Optimize Automatically

Step 4: Train the Model

The model will be trained using stratified k-fold cross-validation to ensure that both classes are represented in each fold. For each fold, we will:

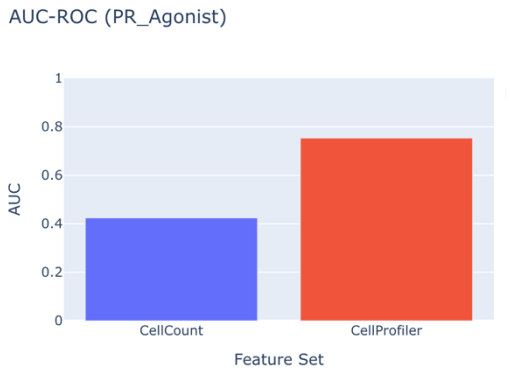
1. Train the model using a 5-fold Cross Validation
2. Train on the 80% data and evaluate on 20% data
3. Repeat this in a 5-fold Cross Validation
4. Calculate metrics like AUC, Balanced Accuracy, AUCPR, and Cohen's Kappa

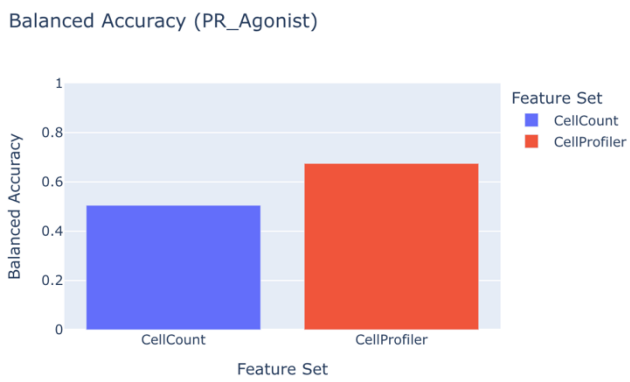
Step 5: Review Model Performance

Here we summarize the mean model performance across feature sets and 5 folds of the CV.

- **AUC-ROC** measures overall ability to discriminate between classes.
- **Balanced Accuracy** accounts for class imbalance.
- **AUCPR** is useful when classes are imbalanced.

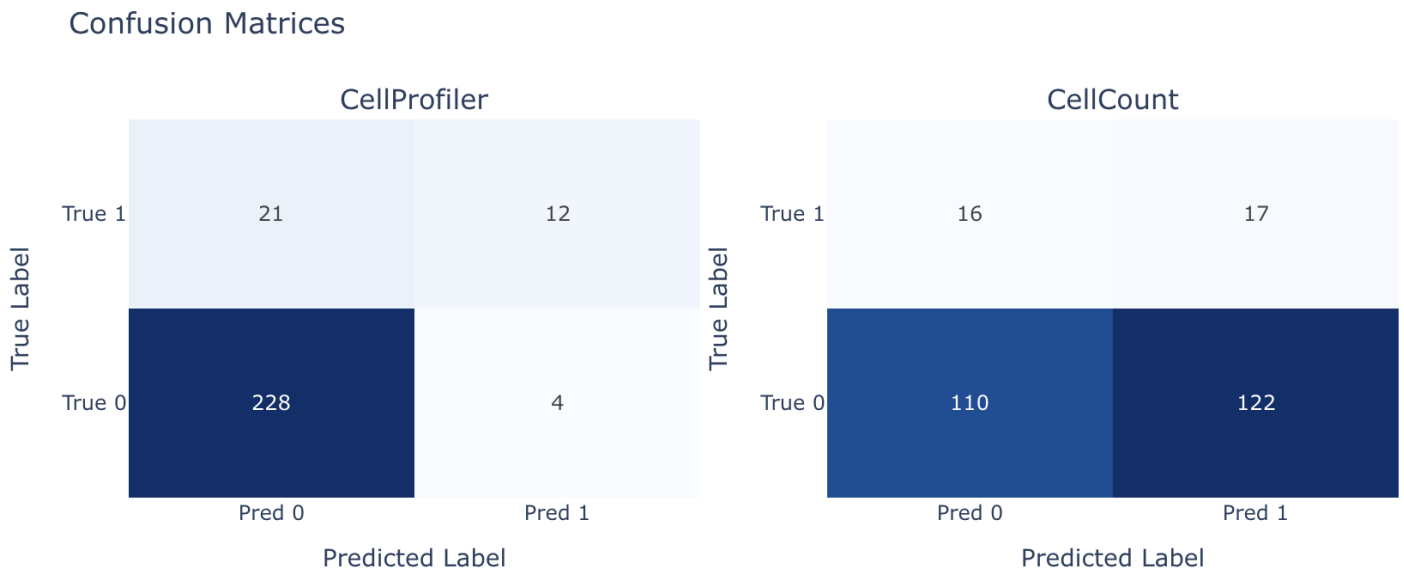
Task object	Feature Set object	AUC float64	Balanced Accuracy float64	AUCPR float64	Sensitivity float64	Specificity float64	MCC float64
PR_Agonist	**CellCount**	0.42	0.51	0.18	0.53	0.48	0.01
PR_Agonist	**CellProfiler**	0.75	0.67	0.56	0.37	0.98	0.49





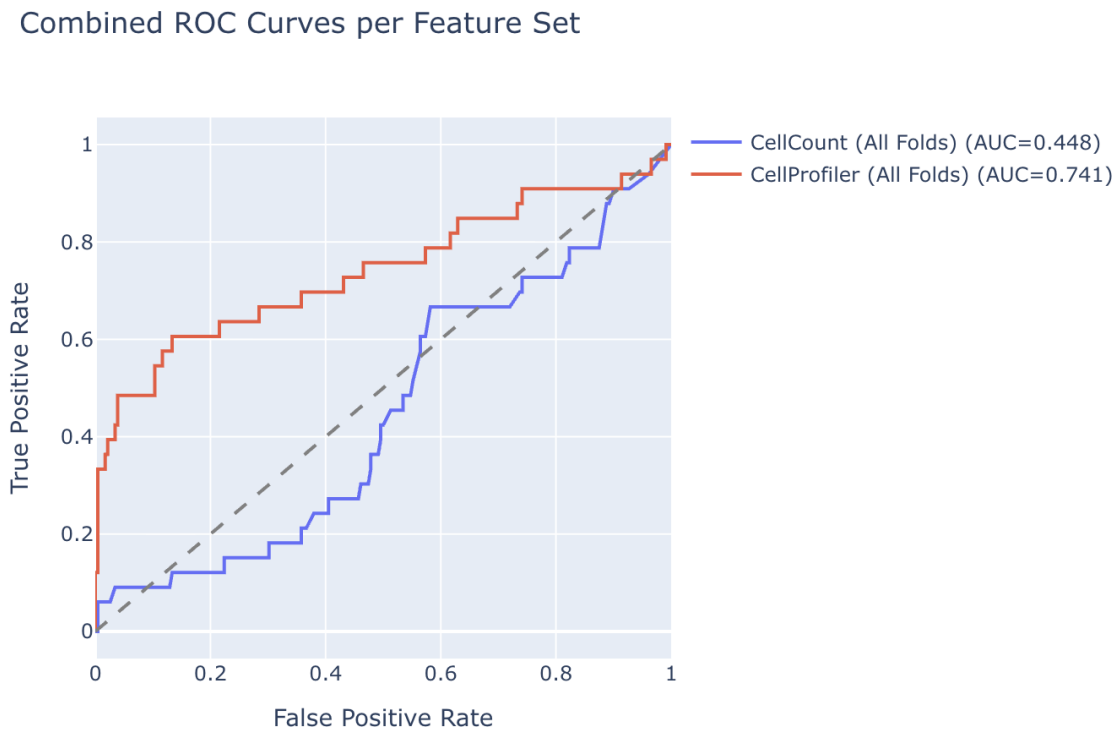
Step 6: Confusion Matrices

Confusion matrices allow you to see the distribution of true positives, false positives, true negatives, and false negatives for each feature set.



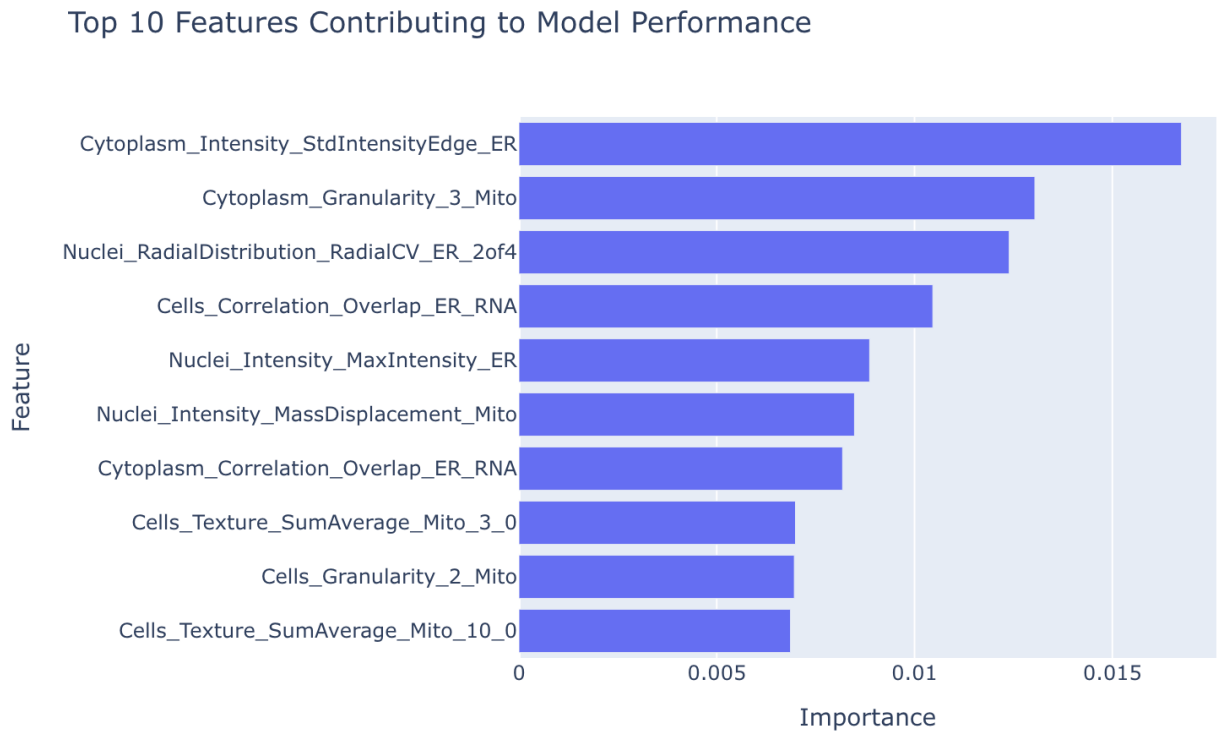
Step 7: ROC Curves

ROC curves show the tradeoff between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate). A model with an AUC close to 1 performs very well; close to 0.5 is no better than random guessing.



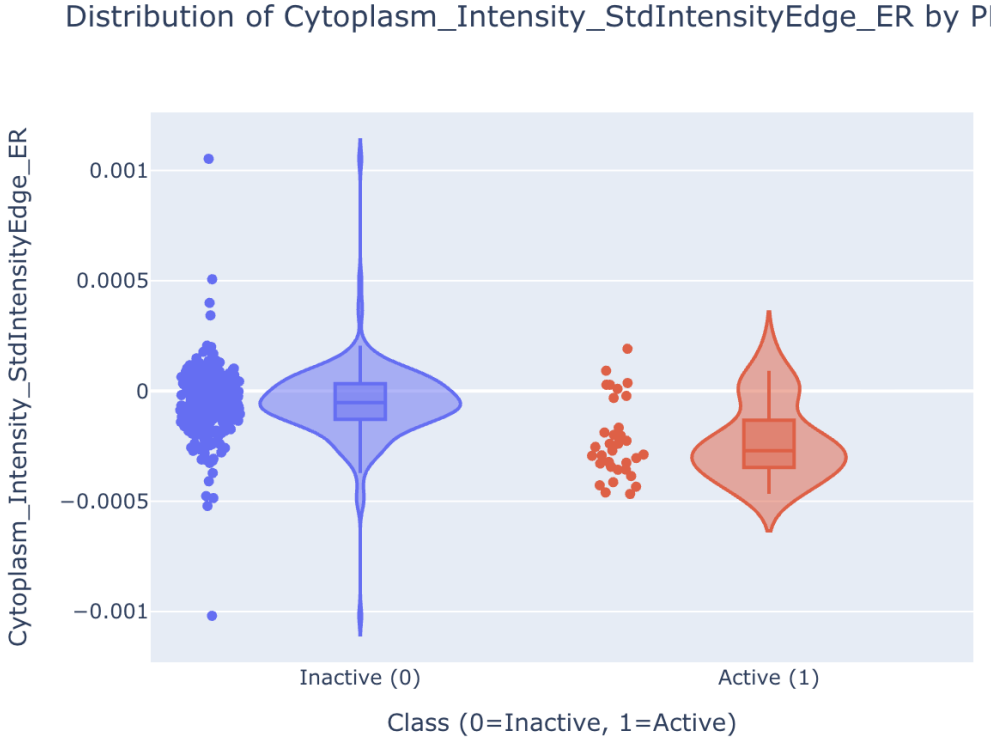
Step 8: Feature Importance

Random Forest models provide feature importance scores. Here we show the **top 10 features** contributing most to the prediction for each feature set. This helps interpret which Cell Painting features are most informative for the selected endpoint.



🔬 Feature Distribution Analysis

Select a feature to analyze: Cytoplasm_Intensity_StdIntensityEdge_ER



Statistical Test Results

Feature: `Cytoplasm_Intensity_StdIntensityEdge_ER`

Endpoint: `PR_Agonist`

Metric	Value
Mann-Whitney U statistic	5948.00
Mann-Whitney p-value	2.6757e-07
T-test statistic	5.52
T-test p-value	8.1670e-08
Significance	✅ Significant (p < 0.05)

Sample Sizes:

- Inactive (0): 232 compounds (mean = -0.000, std = 0.000)
- Active (1): 33 compounds (mean = -0.000, std = 0.000)

✅ Tutorial Complete

You have now:

- Explored Cell Painting data
- Trained and tuned a Random Forest classifier
- Evaluated it using multiple metrics
- Visualized confusion matrices, ROC curves, and feature importances

You can repeat this workflow for other assays or feature sets, and experiment with hyperparameters to improve performance.

Resources

1. http://broad.io/moltox

ACS Publications
Most Trusted. Most Cited. Most Read.

Search text, DOI, authors, etc. Advanced Search

Chemical Research in Toxicology > Vol 38/Issue 5 > Article

Open Access Editors' Choice

Cite Share Jump to Expand

REVIEW | May 2, 2025

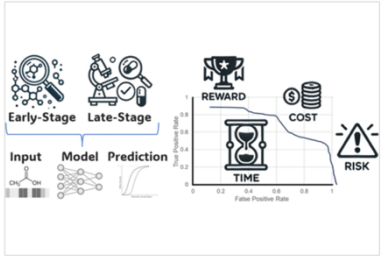
Machine Learning for Toxicity Prediction Using Chemical Structures: Pillars for Success in the Real World

Srijit Seal*, Manas Mahale, Miguel García-Ortegón, Chaitanya K. Joshi, Layla Hosseini-Gerami, Alex Beatson, Matthew Greenig, Mrinal Shekhar, Arijit Patra, Caroline Weis, Arash Mehrjou, Adrien Badré, Brianna Paisley, Rhiannon Lowe, Shantanu Singh, Falgun Shah, Bjarki Johannesson, Dominic Williams, David Rouquie, Djork-Arné Clevert, Patrick Schwab, Nicola Richmond, Christos A. Nicolaou, Raymond J. Gonzalez, Russell Naven, Carolin Schramm, Lewis R Vidler, Kamel Mansouri, W. Patrick Walters, Deidre Dalmas Wilk, Ola Spjuth*, Anne E. Carpenter*, and Andreas Bender*

Open PDF

Abstract

Machine learning (ML) is increasingly valuable for predicting molecular properties and toxicity in drug discovery. However, toxicity-related end points have always been challenging to evaluate experimentally with respect to *in vivo* translation due to the required resources for human and animal studies; this has impacted data availability in the field. ML can augment or even potentially replace traditional experimental processes depending on the project phase and specific goals of the prediction. For instance, models can be used to select promising compounds for on-target effects or to deselect those with undesirable characteristics (e.g., off-target or ineffective due to unfavorable pharmacokinetics). However, reliance on ML is not without risks, due to biases stemming from nonrepresentative training data, incompatible choice of algorithm to represent the underlying data, or poor model



2. http://broad.io/AIDDCourse

Introduction to Cheminformatics and AI in Drug Discovery: Hands-on Modeling of Safety Data

Machine learning (ML) and artificial intelligence (AI) models are becoming increasingly popular in drug discovery. This course aims to explain how to construct and use ML models for a non-expert audience with a background in life science or safety sciences.

Overview

This course focuses on hands-on learning. It begins with a lecture covering the basics of predictive ML models, including data, descriptors, ML algorithms, and model validation. Participants will then move to two practical sessions (target prediction and mitochondrial toxicity classification), where they will learn to prepare input data, train models, and apply them to new datasets. The interactive approach will deepen understanding of the concepts.

3. http://broad.io/BookADMEML

ACADEMIC PRESS

Machine Learning and Artificial Intelligence in Toxicology and Environmental Health

2026, Pages 61-97

Chapter 3 - Application of machine learning and artificial intelligence methods in predictions of absorption, distribution, metabolism, and excretion properties of chemicals *

Wei-Chun Chou [1], Miao Li [2], Srijit Seal [3], Zhoumeng Lin [4 5]

Show more