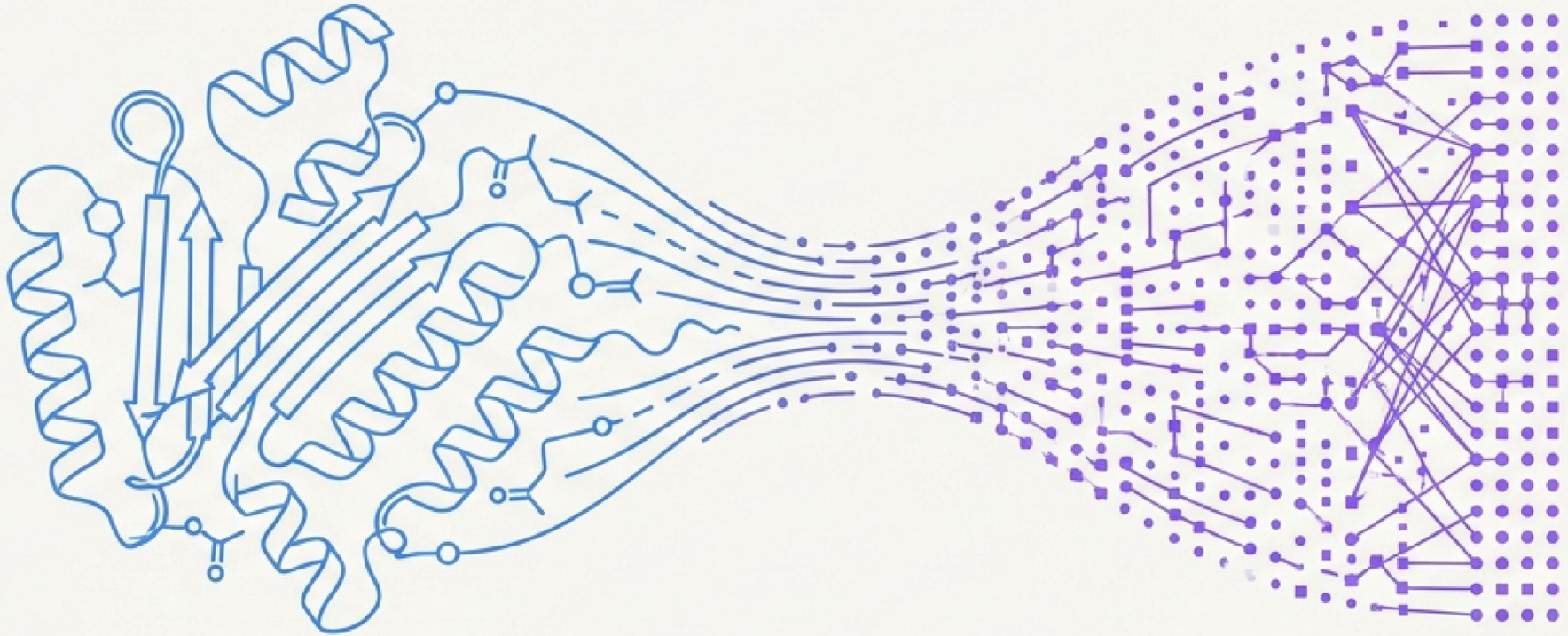
# From Biology to Bits: Predicting Alzheimer's Drugs with Machine Learning

A Computational Drug Discovery Project for Beta-amyloid A4 Protein Inhibitors



The process of building a Quantitative Structure-Activity Relationship (QSAR) model to accelerate the search for a cure.

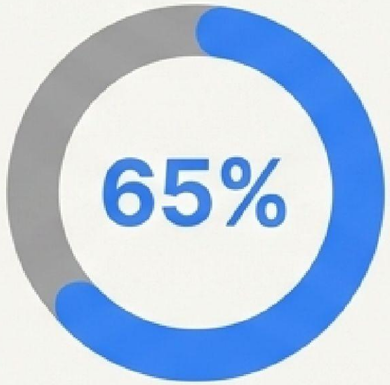
# The Mission: Confronting Alzheimer's Disease

**~50**

million people globally

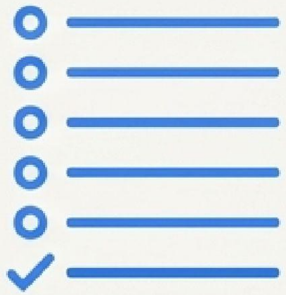
**60-70%**

of all dementia cases worldwide




**#6**

leading cause of death in the US



Despite its prevalence, no cure currently exists. Computational methods offer a new path to accelerate drug discovery, reducing the typical $2.6 billion cost and years of lab testing.

# The Biological Target: Pinpointing the Cause

## The Key Player: Beta-amyloid A4 Protein (Aβ)

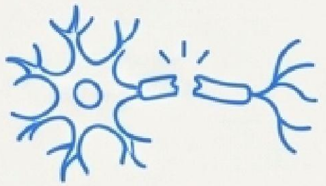
A small protein fragment produced when a larger protein (APP) is cut. In Alzheimer's, it accumulates and forms toxic clumps, or plaques.

ChEMBL Target ID: CHEMBL2487

## The Disease Process: A Cascade of Damage

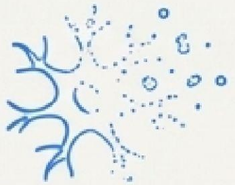
**1 Aβ Aggregation**
Proteins clump together.

**2 Synaptic Dysfunction**
Clumps disrupt neuron signals, affecting memory.



**3 Inflammation & Stress**
The brain's immune response causes collateral damage.

**4 Neuronal Death**
Neurons die, the brain shrinks, and cognitive function declines.

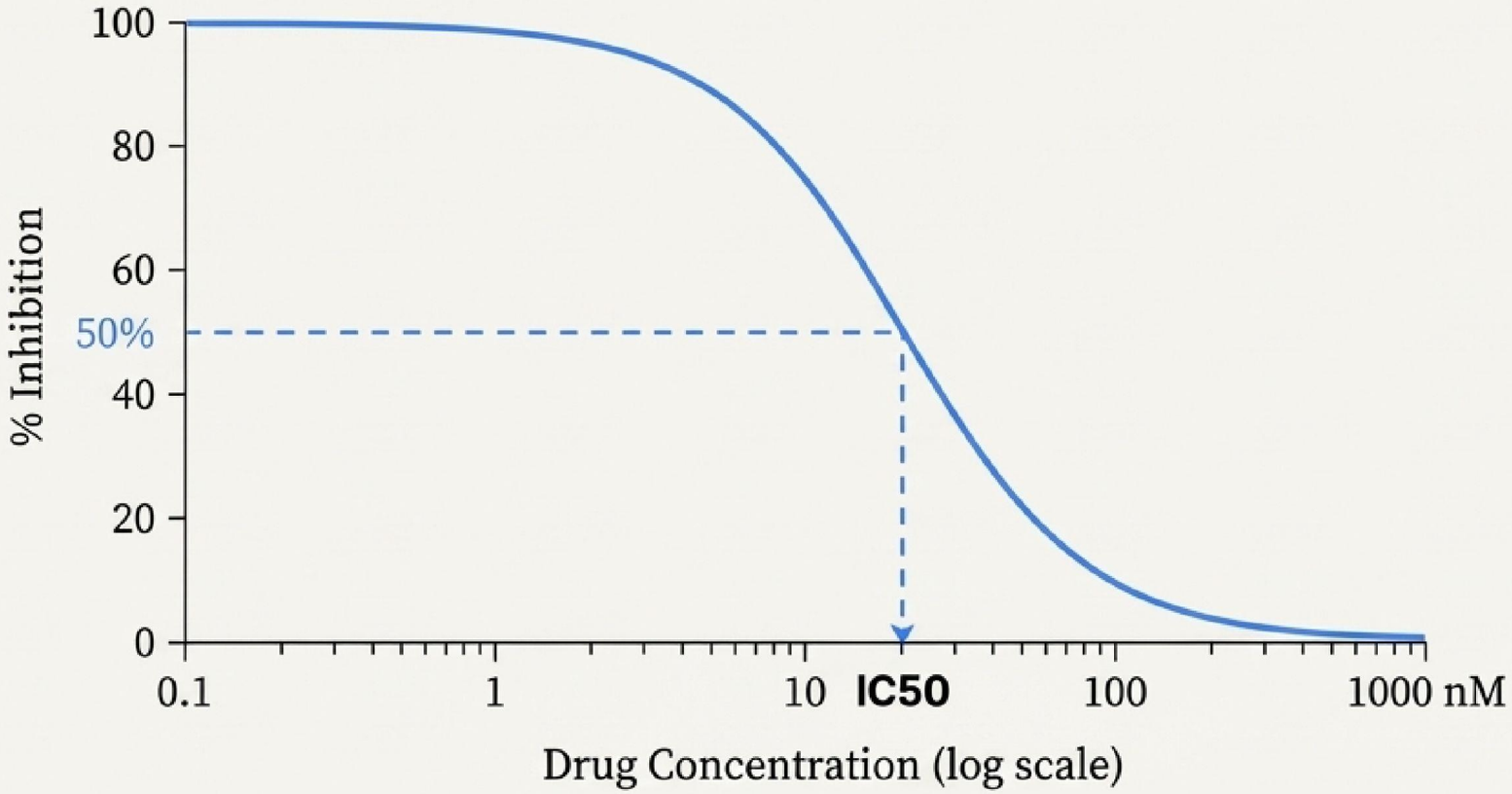


Our Goal: Find chemical compounds (inhibitors) that stop this process at its source.

# Measuring Success: How We Quantify a Drug's Potency

## IC50 (Half Maximal Inhibitory Concentration)

"A quantitative measure that indicates how much of a drug is needed to inhibit a biological process by 50%."




**Lower is Better.**

- A low IC50 (e.g., 5 nM) means a tiny amount is needed = a very potent drug.
- A high IC50 (e.g., 50,000 nM) means a huge amount is needed = a weak drug.

All data is standardized to Nanomolar (nM) units for consistency.

# The Strategic Framework: QSAR

$$\text{Biological Activity} = f(\text{Chemical Structure})$$

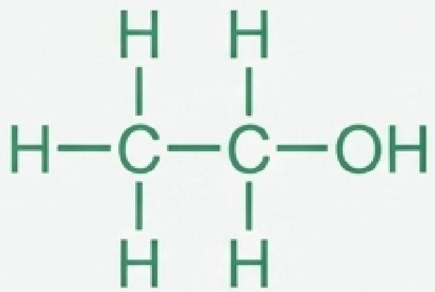
**Quantitative Structure-Activity Relationship** (QSAR) is a hypothesis that allows us to build mathematical models that predict the efficacy of *untested* chemicals. By understanding the relationship between a molecule's structure and its activity, we can screen for new drugs computationally.

This framework is the foundation for our entire project. It allows us to move from slow, expensive lab experiments to rapid, cost-effective virtual screening.

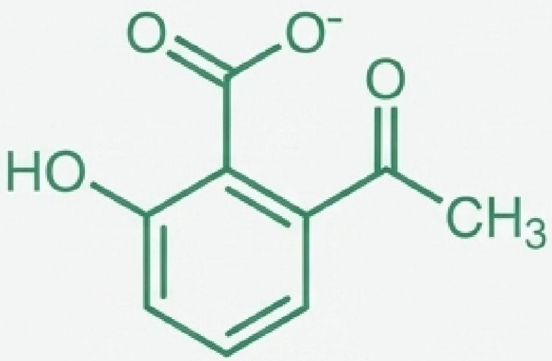
# The Language of Molecules: SMILES Notation

How we represent a 3D chemical structure as text for computer processing.

> **SMILES:** Simplified Molecular Input Line Entry System. A specification for describing molecular structures using short ASCII strings.




CCO




CC(=O)OC1=CC=CC=C1C(=O)O

Ethanol

Aspirin

We use **Canonical SMILES** to ensure every unique molecule has exactly one unique text identifier, preventing ambiguity in the dataset.

# Is It a Viable Drug? Applying Lipinski's Rule of 5

A rule of thumb to evaluate if a compound has properties that would make it a likely orally active drug. It helps filter for "drug-like" candidates.

## The Rules

- **Molecular Weight (MW) < 500 Daltons:** Not too big.
- **LogP < 5:** Not too greasy (balances solubility).
- **Hydrogen Bond Donors < 5:** Controls interactions with water.
- **Hydrogen Bond Acceptors < 10:** Controls interactions with water.

Good Candidate

✓ ✓ ✓ ✓

Poor Candidate

✗ ✗ ✗ ✗

# The End-to-End Data Pipeline

## 1. Data Collection

Fetching experimental data for our target protein from the ChEMBL database.

## 2. Data Curation

Cleaning the data, removing missing values, and standardizing IC50 measurements.

## 3. Feature Engineering

Translating molecules (SMILES) into numerical features (Fingerprints) and normalizing the target variable (pIC50).

## 4. Model Building

Splitting the data and training dozens of ML algorithms to find the best performer.

## 5. Model Evaluation

Assessing the final model's predictive accuracy on unseen data using $R^2$ and RMSE.

# From Raw Data to a Curated Dataset

## Data Collection & Cleaning

**Source*: ChEMBL Database

**Query**:

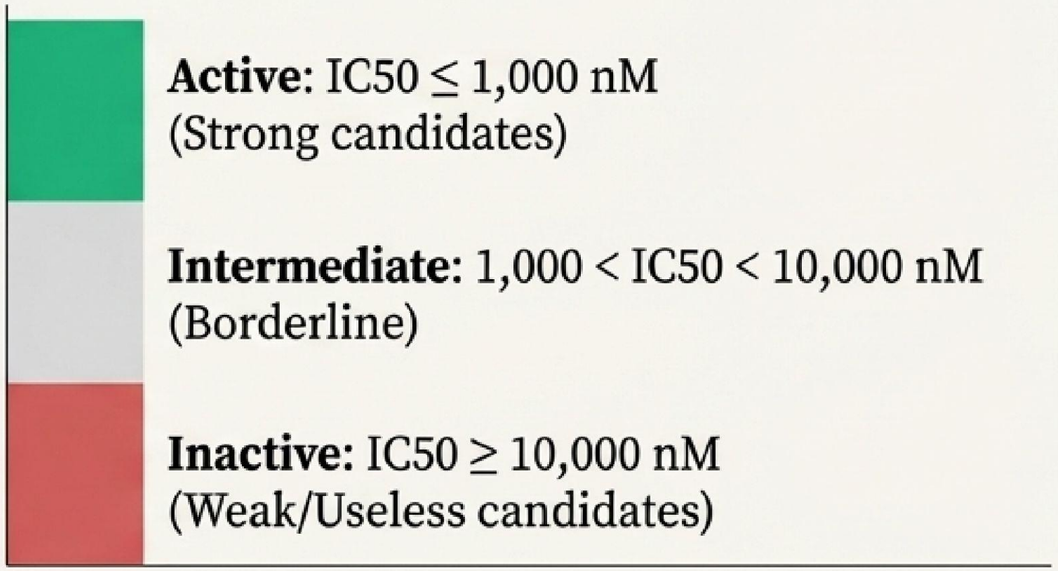
```
`new_client.target.filter(target_chem-
bl_id='CHEMBL2487')`
```

**Filtering**:

- Kept only experiments measuring `standard_type == 'IC50'`.
- Dropped rows with missing `standard_value` or `canonical_smiles`.

## Bioactivity Classification

We created three distinct classes based on IC50 values to enable analysis.



**Active**: IC50 ≤ 1,000 nM
(Strong candidates)

**Intermediate**: 1,000 < IC50 < 10,000 nM
(Borderline)

**Inactive**: IC50 ≥ 10,000 nM
(Weak/Useless candidates)

# Normalizing Potency: The pIC50 Transformation

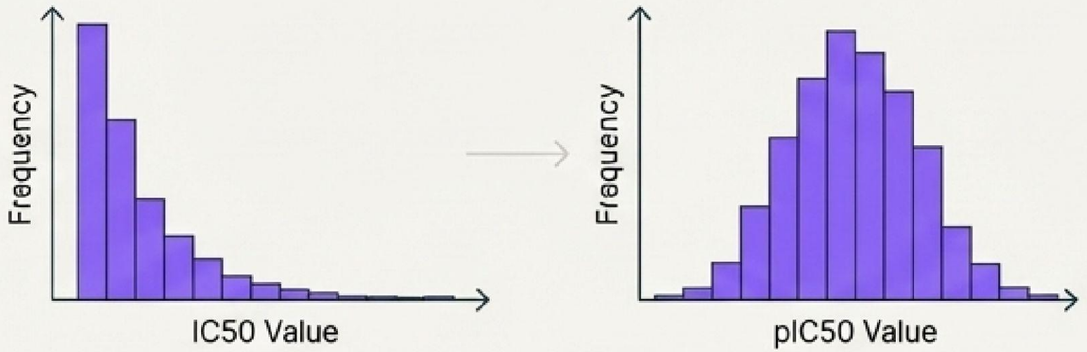
**The Problem:** Raw IC50 data is highly skewed and non-linear, spanning many orders of magnitude. This is problematic for regression models.

**The Solution:** Convert IC50 to pIC50 using the formula: **$pIC50 = -\log_{10}(IC50)$**

## Justification

### Statistical

Transforms the skewed data distribution into a **more uniform**, linear scale suitable for machine learning.




**Before:** IC50 Values

**After:** pIC50 Values

### Intuitive

Converts the “Lower is Better” scale of IC50 into a standard “Higher is Better” scale. A more potent drug now has a higher pIC50 value.

**Potency:** Weak → Strong

**IC50 Scale:** Higher → Lower

**pIC50 Scale:** Lower → **Higher**

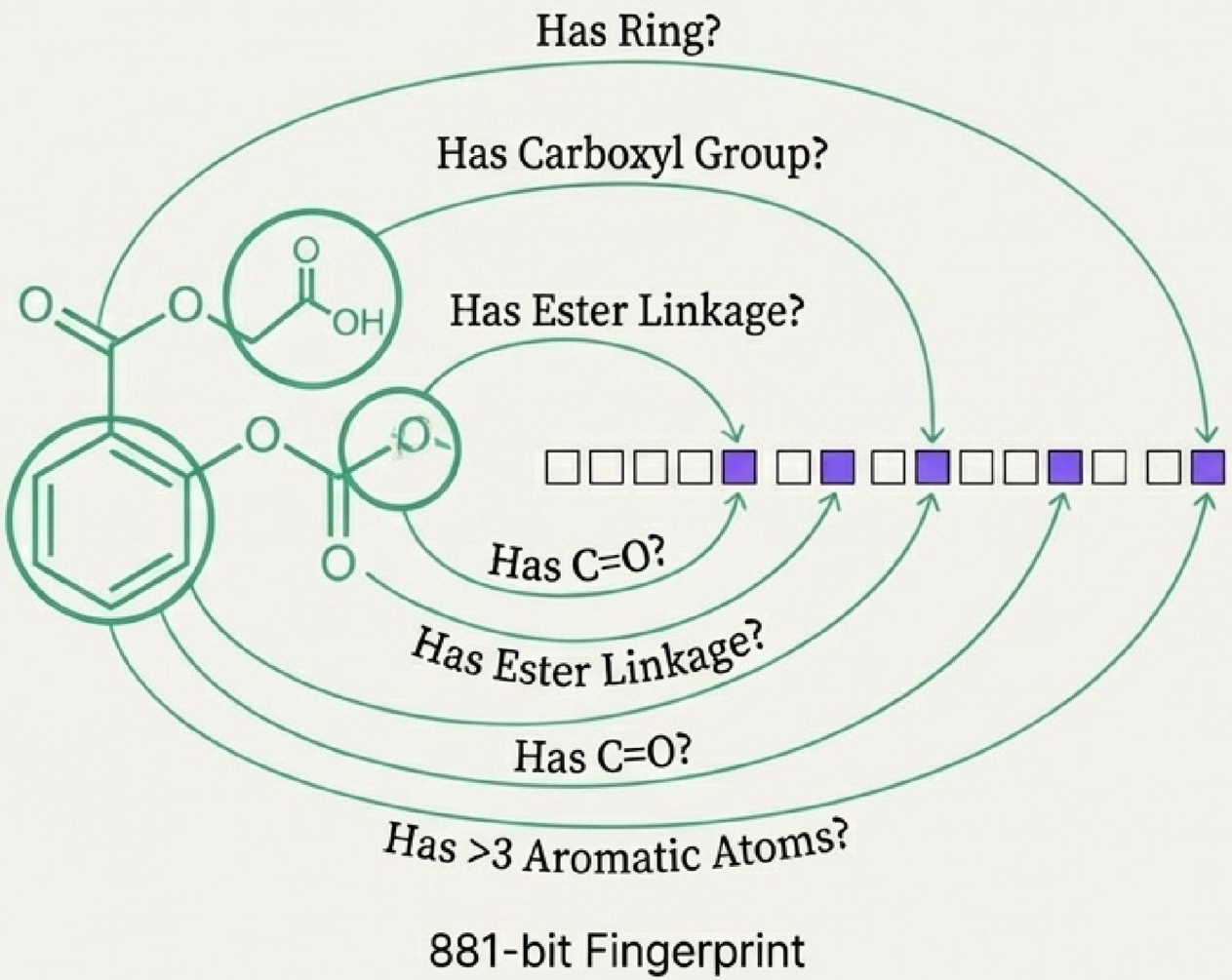
# Creating a “Barcode” for Every Molecule

****Concept*: Molecular Fingerprints**

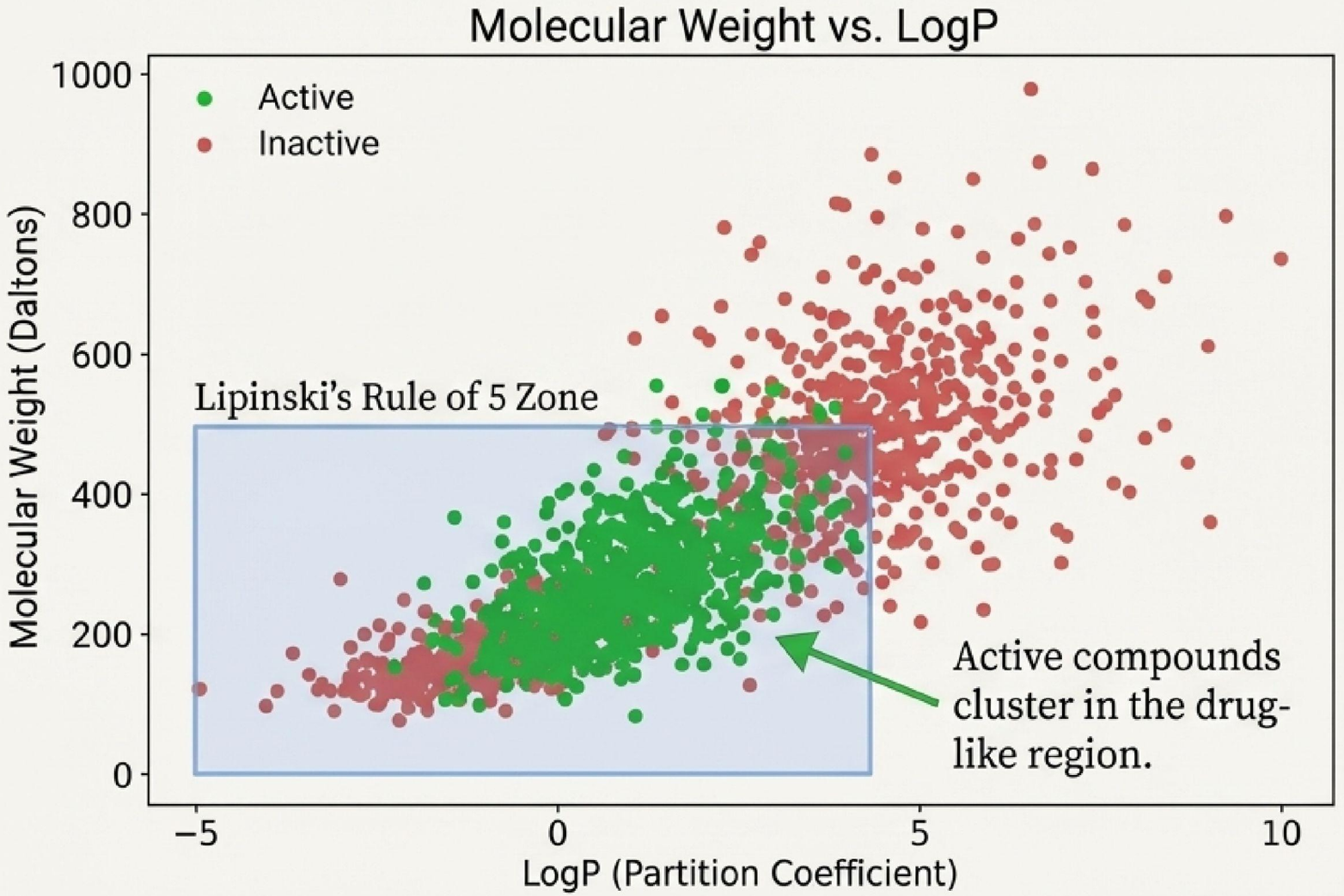
Like a barcode for a molecule, a fingerprint is a binary vector (an array of 0s and 1s) representing the presence or absence of specific chemical substructures.

****Implementation : PubChem Fingerprints**

We used the PaDEL-Descriptor software to calculate fingerprints. This system checks for **881 specific substructures**. The output for each molecule is a binary vector of length 881.



# Exploring the Chemical Space




**How do we know these differences aren't random chance?**

**Mann-Whitney U Test:** We used this statistical test to compare the distributions of descriptors (like pIC50, MW, LogP) between the Active and Inactive groups.

**Result:** The p-value was < 0.05, proving the observed differences are statistically significant.

# Preparing for the Algorithm Tournament

**Step 1: Feature Selection (Reducing Noise)

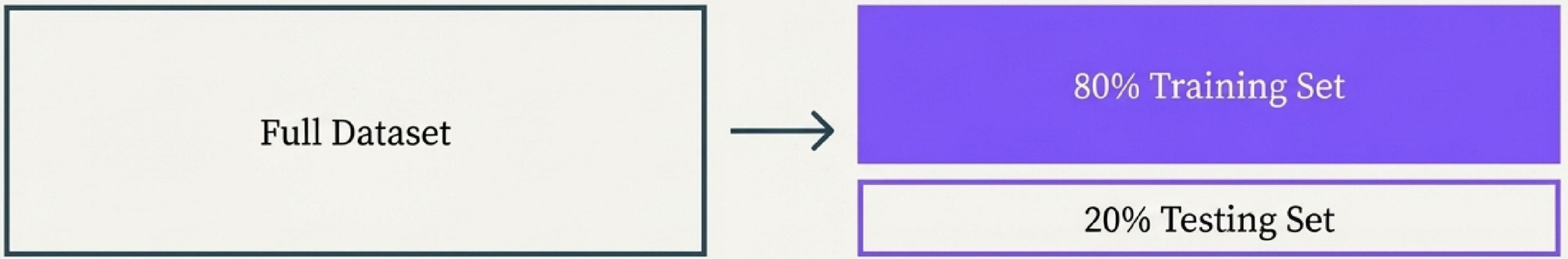
`VarianceThreshold(threshold=0.1)`

We remove fingerprint columns (features) that are the same for nearly all molecules. If a feature doesn't vary, it provides no information for the model to learn from. This speeds up training and improves performance.

**Step 2: Train/Test Split (Preventing Memorization)

`train_test_split(test_size=0.2)`

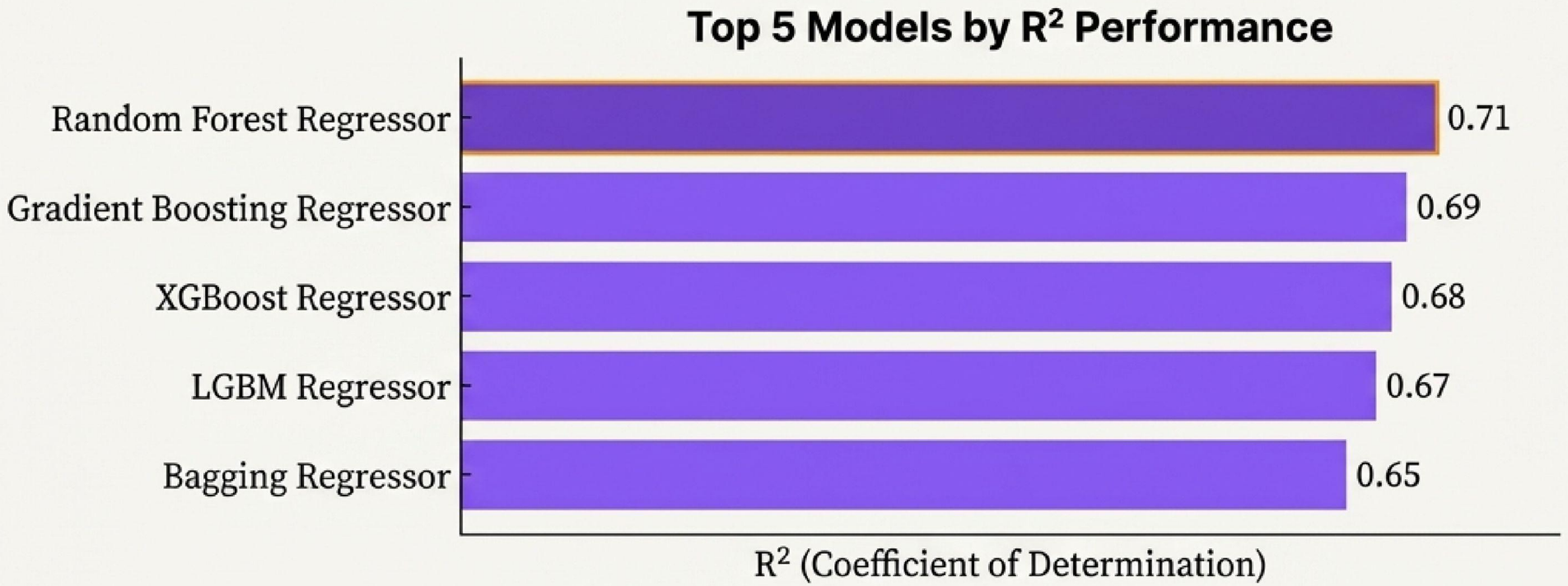
We must evaluate the model on data it has never seen before to ensure it has truly learned the underlying patterns, not just memorized the training data (a problem known as overfitting).



# The Algorithm Tournament: Finding the Best Predictor

## **The Tool*: LazyPredict

Instead of guessing which algorithm would perform best, we used the LazyPredict library to rapidly train and evaluate over 30 regression models on our dataset.




The **Random Forest Regressor** consistently emerged as one of the top-performing models for this QSAR dataset.

# The Champion Model: Random Forest Regressor

### $R^2$ (Coefficient of Determination)

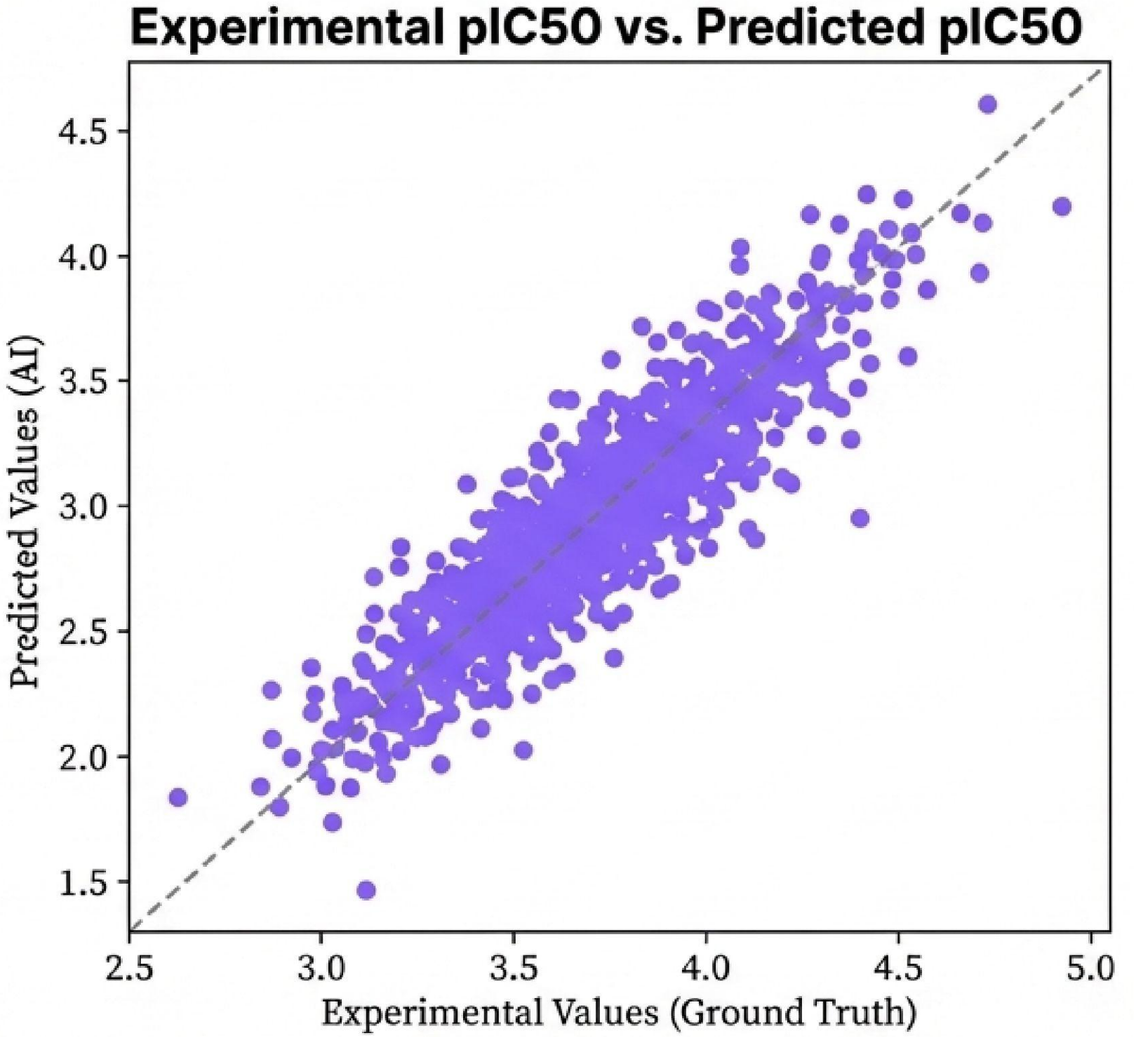
## ~0.71

*Represents the proportion of variance explained by the model. A value > 0.6 is generally considered good for QSAR models.*

### RMSE (Root Mean Squared Error)

## 0.85

*Measures the standard deviation of prediction errors. A lower value indicates a better fit.*



# The Result: A Validated Tool to Accelerate Discovery

**Biology:** Identified the target, Beta-amyloid A4 protein.

**Chemistry:** Represented molecules as numerical fingerprints.

**Data Science:** Built and validated a QSAR model to predict pIC50.

## The Final Product

We have created a robust, predictive model that can accurately estimate the bioactivity of a chemical compound based solely on its structure.

**This model can now be used to screen millions of *new, unmade* chemical structures virtually, drastically accelerating the identification of promising drug candidates for Alzheimer's Disease.**