

Decoding Alzheimer's: A Computational Approach to Drug Discovery

Building a Quantitative Structure-Activity Relationship (QSAR)
Model to Predict Inhibitors for Beta-amyloid A4 Protein.



The Alzheimer's Challenge: A Global Health Crisis



Affects ~50 million people globally.

Causes 60-70% of all dementia cases.

The 6th leading cause of death in the US.

Currently, no cure exists.

The Economic Barrier to a Cure

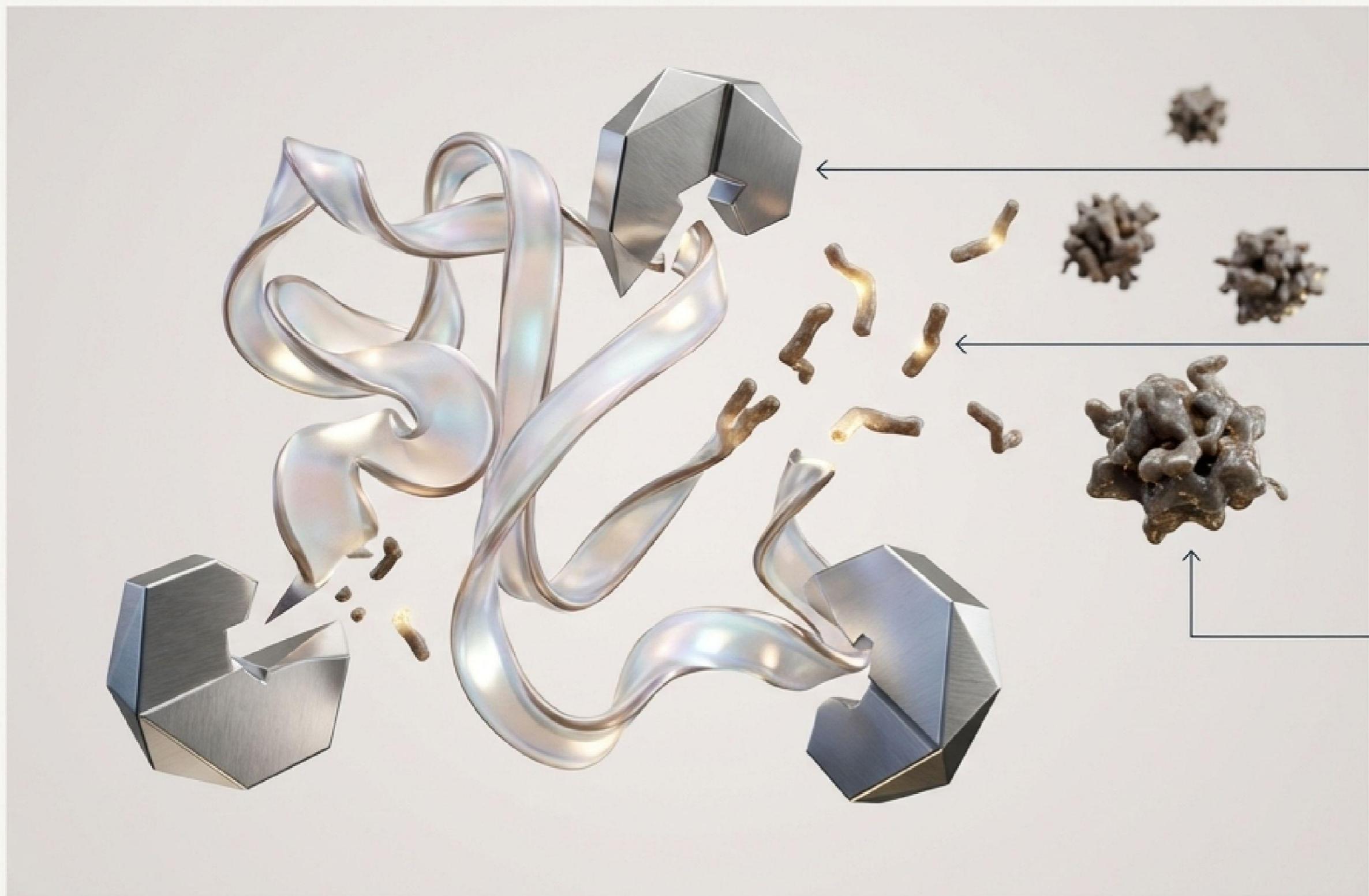
Traditional drug discovery is a slow, expensive, and inefficient process.



\$2.6 billion

The average cost to bring a new drug to market, a process that can take over a decade with a high failure rate in lab testing.

The Target: Beta-amyloid A4 Protein

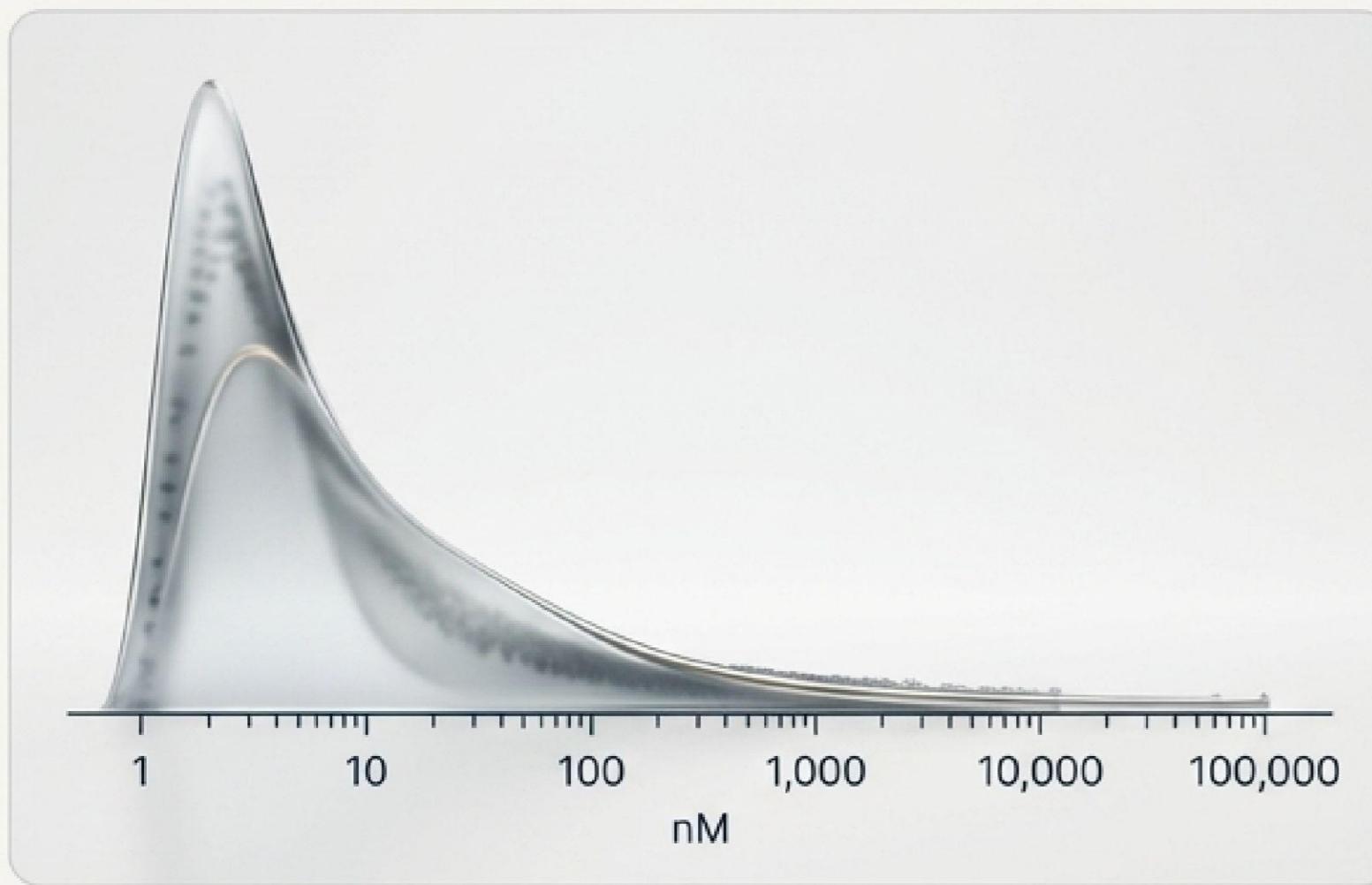


The Amyloid Hypothesis

1. A **larger protein (APP)** is cut by **enzymes**.
2. This produces small protein fragments called **Beta-amyloid (A β)**.
3. In Alzheimer's, these fragments accumulate, forming **toxic clumps (plaques)** that **disrupt synapses and lead to neuron death**.

Quantifying Potency: From IC50 to pIC50

IC50 (Lower is Better)

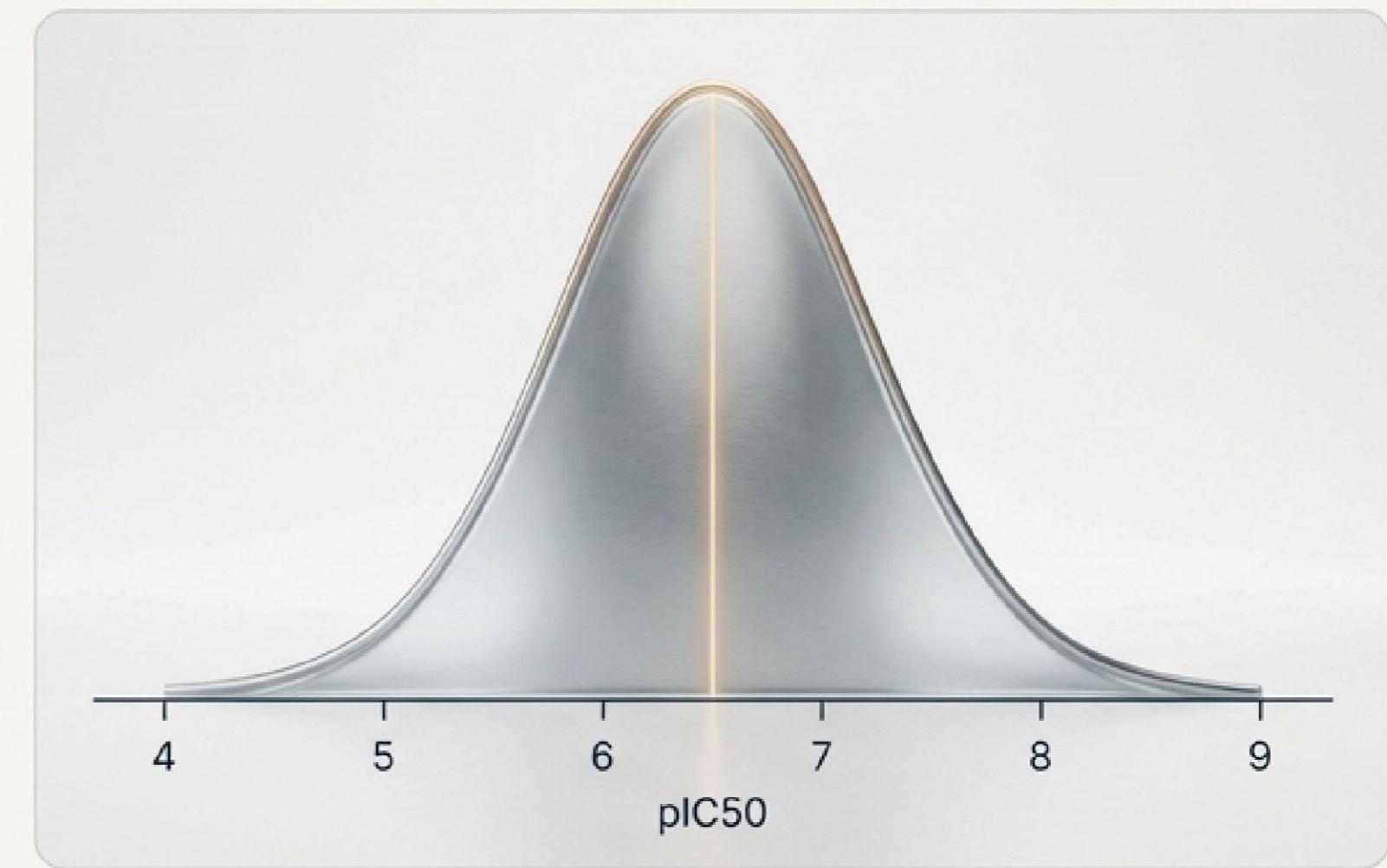


IC50 (Half Maximal Inhibitory Concentration)

Definition: The amount of a drug needed to inhibit a biological process by 50%.

Scale: **Lower is Better** (e.g., 5nM is more potent than 50,000 nM).

pIC50 (Higher is Better)



The pIC50 Transformation

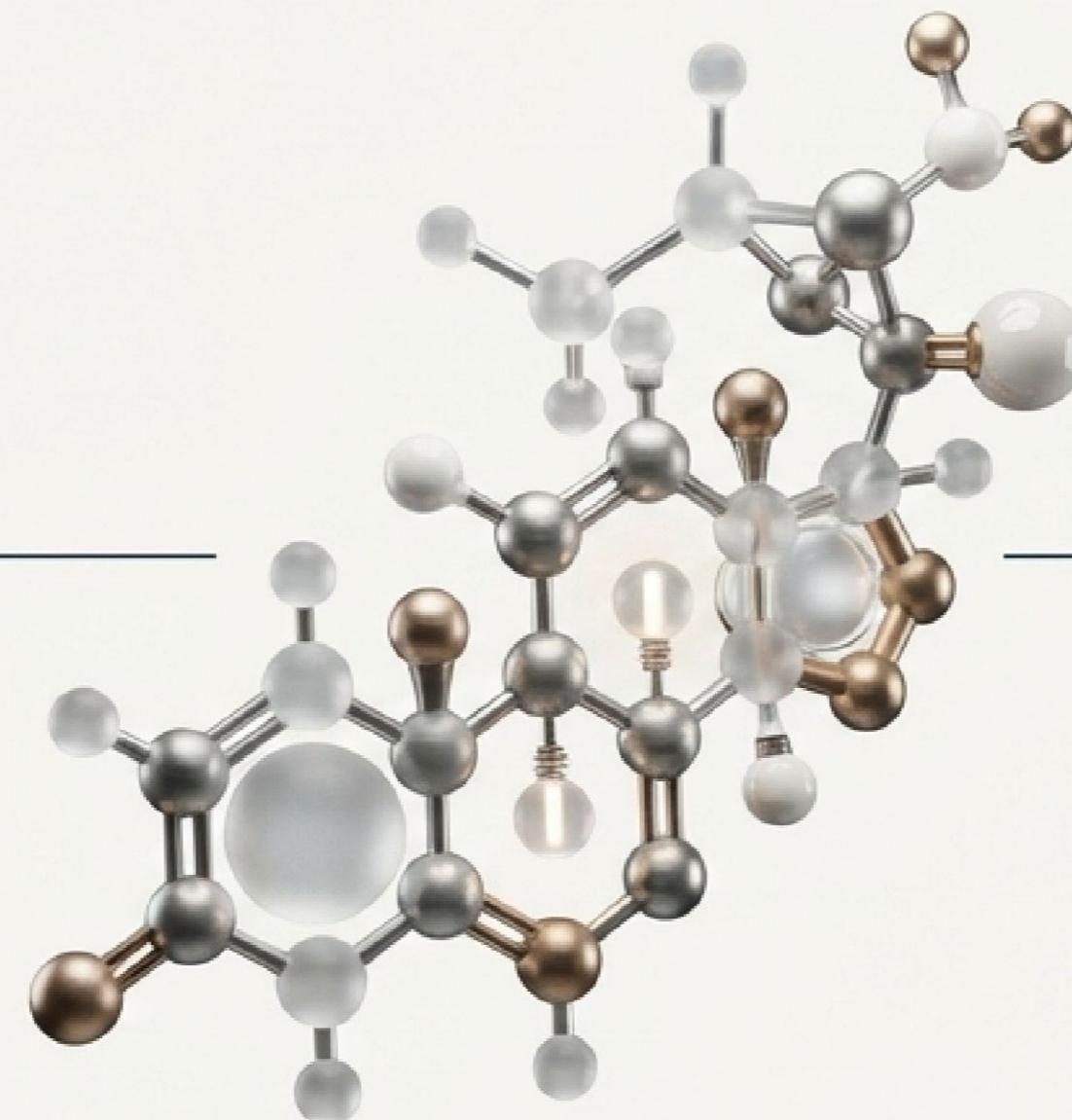
Formula: $\text{pIC50} = -\log_{10}(\text{IC50})$

Why? It converts the skewed, exponential IC50 scale into a linear one suitable for regression models and makes the scale more intuitive: **Higher is Better**.

Translating Molecules into Machine-Readable Language

The Language (SMILES)

- **SMILES:** Simplified Molecular Input Line Entry System.
- **Function:** Represents a 3D chemical structure as a simple text string. (e.g., CCO for Ethanol).



The Features (Molecular Fingerprints)

- **Analogy:** A 'molecular barcode.'
- **Method:** We use PubChem Fingerprints, which generate an 881-bit binary vector (an array of 0s and 1s) representing the presence or absence of 881 specific chemical substructures.

The Guiding Principle: Quantitative Structure-Activity Relationship (QSAR)

Core Hypothesis: Biological Activity can be predicted as a function of Chemical Structure.

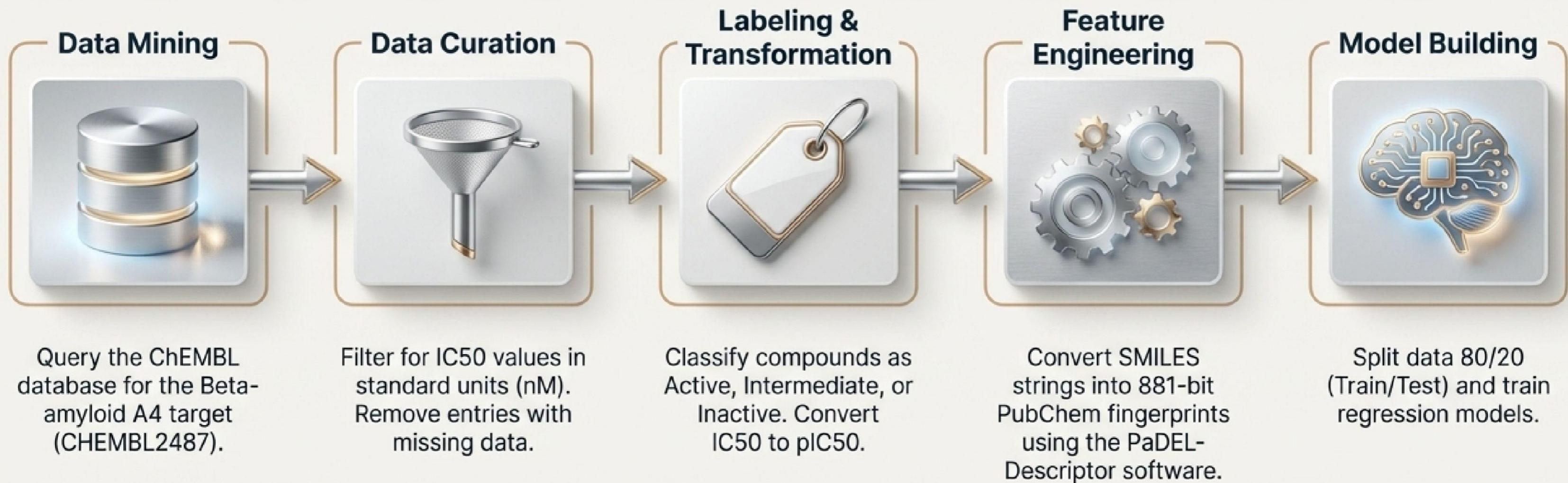
Formal Expression:

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

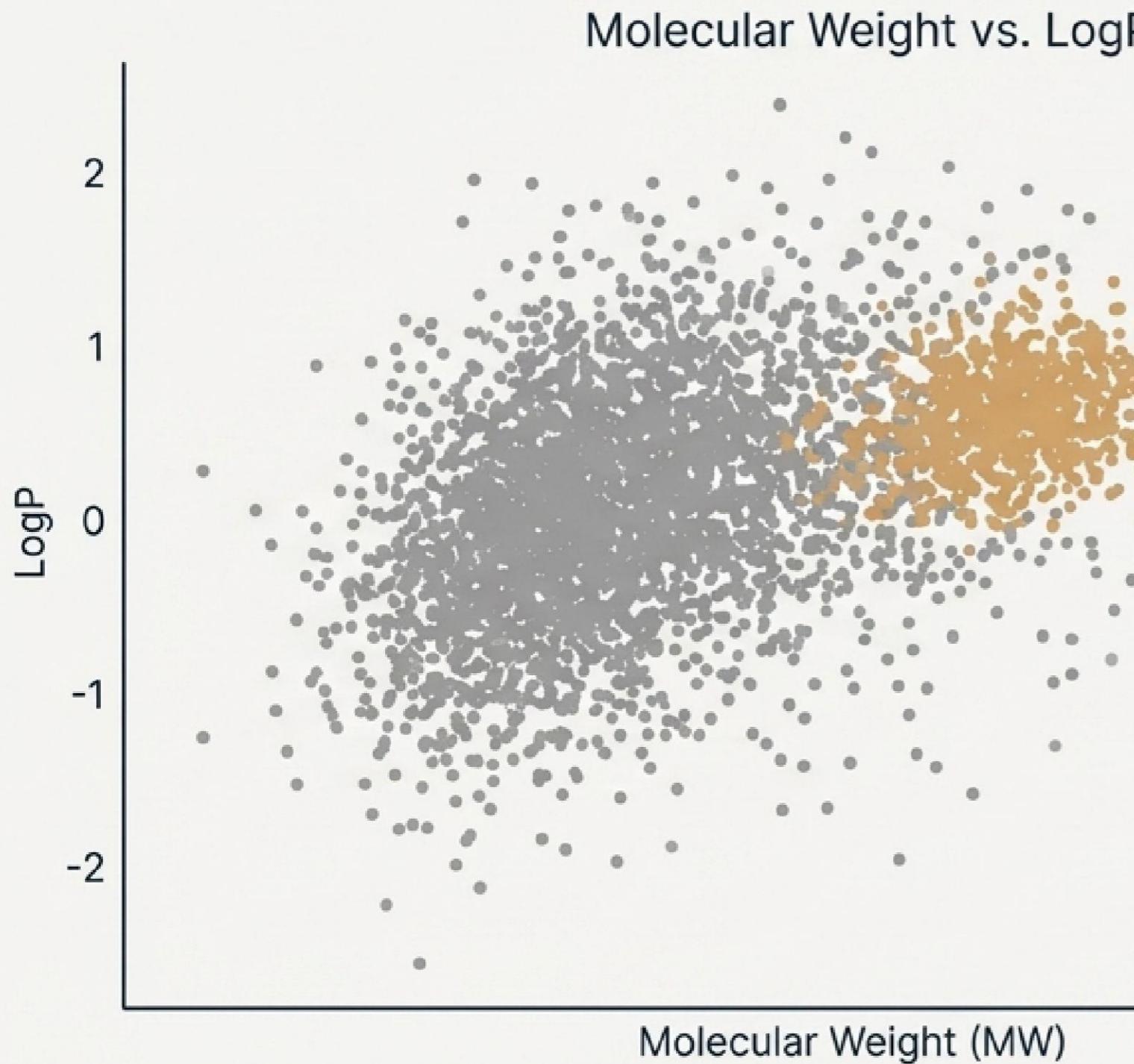


Our Goal: Use machine learning to learn this function, allowing us to predict the pIC_{50} of any new molecule from its structure alone.

The Computational Workflow: An End-to-End Pipeline



Exploring the Chemical Space: Finding the Signal in the Noise



Insight: **Active** compounds tend to cluster in a specific, 'drug-like' region, consistent with Lipinski's Rule of 5.

Statistical Proof: A **Mann-Whitney U Test** was performed, confirming that the difference in descriptor distributions between Active and Inactive groups is statistically significant ($p < 0.05$), not due to random chance.

Mann-Whitney U Test: $p < 0.05$

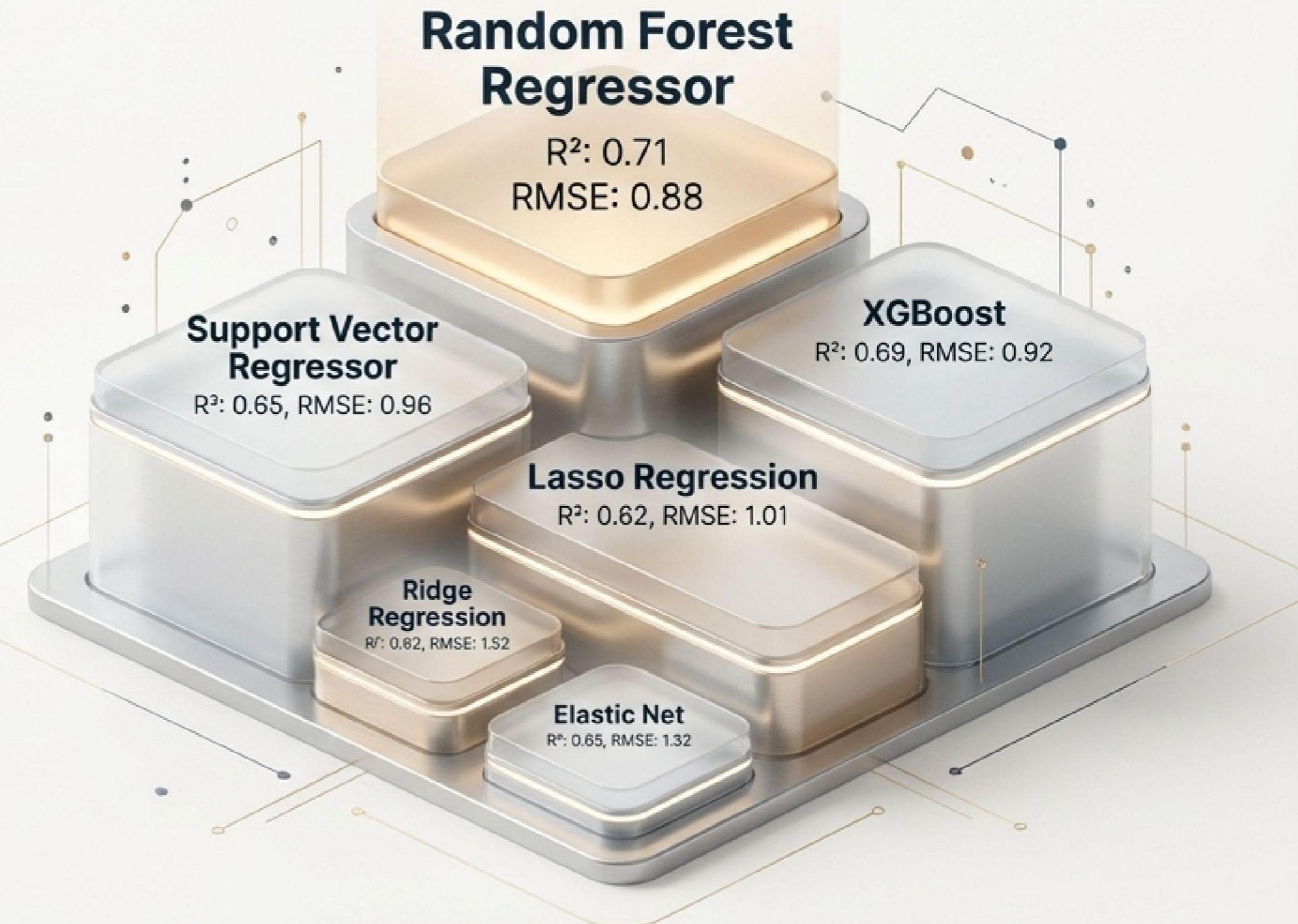
An Algorithm Tournament to Identify the Top Performer

Methodology: We used the **LazyPredict** library to rapidly build and evaluate over 30 regression models with default parameters.

Purpose: Instead of guessing, this approach provides an empirical baseline, ranking algorithms based on their performance on our specific chemical dataset.



Result: The **Random Forest Regressor** consistently emerged as a top-performing model.

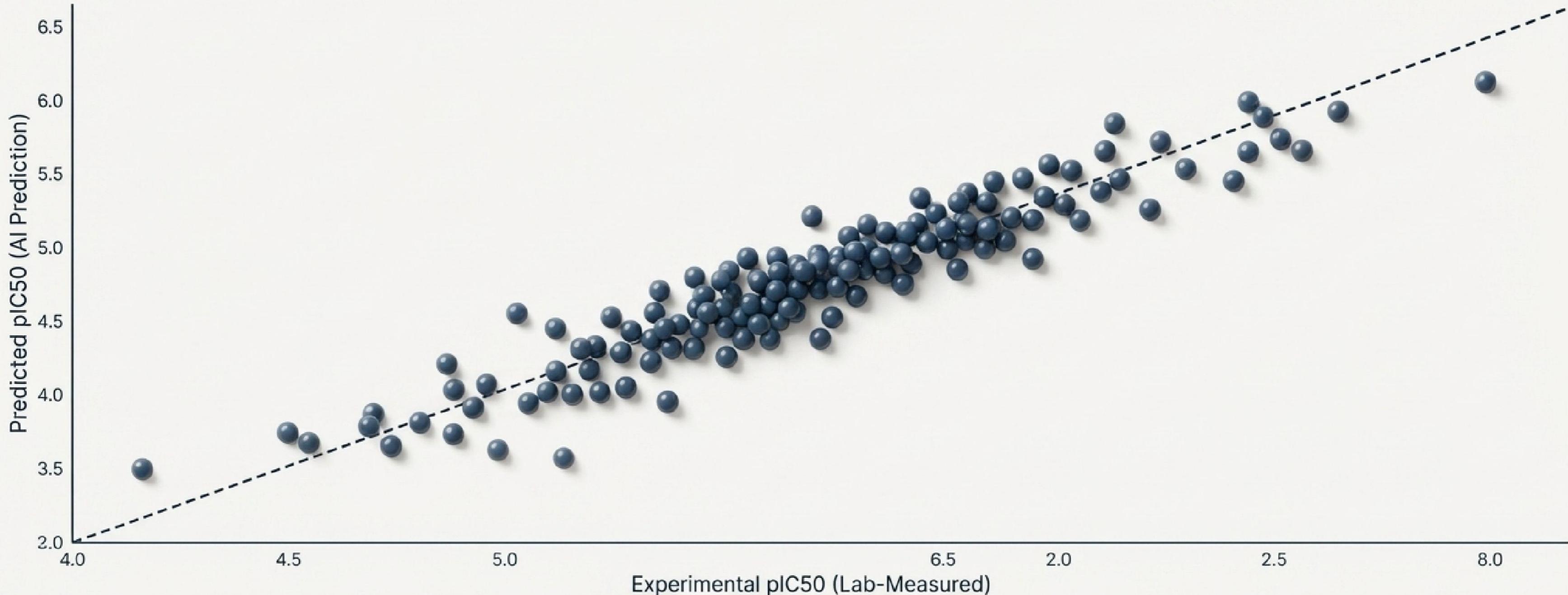


The Verdict: Predicting Biological Activity with High Accuracy

Model: Random Forest Regressor

R² = 0.71

(Note: An R² value > 0.6 is generally considered a strong result for QSAR models).



The data points cluster tightly along the diagonal line, indicating a strong correlation between the lab-measured values and the AI's predictions.

The Outcome: A Validated Tool for Virtual Screening

Capability: Our QSAR model can now predict the pIC₅₀ for *new, untested* chemical compounds based solely on their structure.



Capability: Our QSAR model can now predict the pIC₅₀ for *new, untested* chemical compounds based solely on their structure.

Application (High-Throughput Virtual Screening): We can computationally screen vast digital libraries of millions of molecules to identify the most promising candidates for synthesis and lab testing.

Impact

- Drastically reduces time and cost.
- Active candidate molecules.

Impact

- Drastically reduces time and cost.
- Focuses lab resources on high-potential compounds.
- Accelerates the entire drug discovery pipeline.

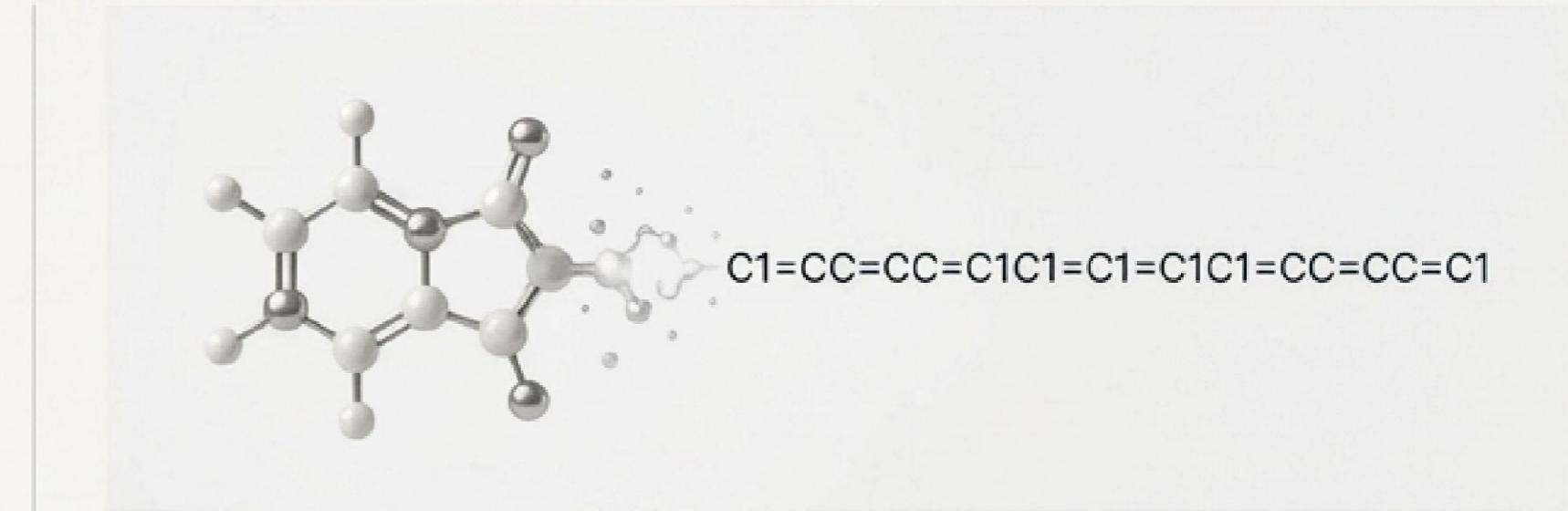
The Journey: From Biological Chaos to Computational Clarity



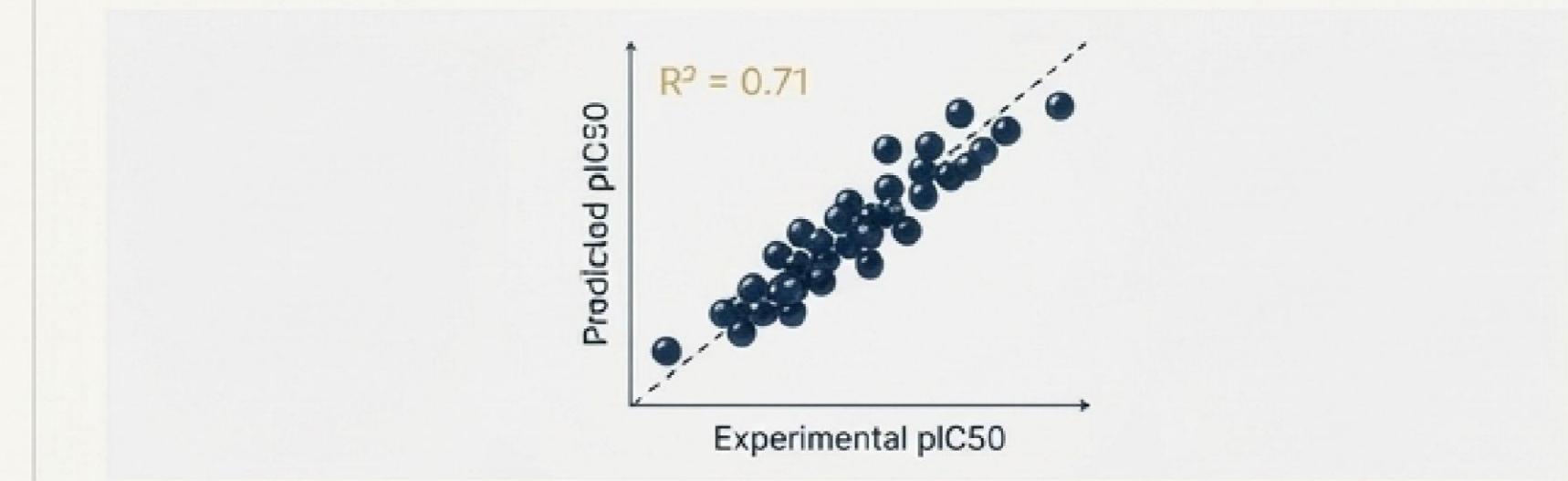
We started with the complex problem of Alzheimer's and identified the **Beta-amyloid protein**.



We converted chemical structures into numerical **fingerprint**s to serve as features.



We translated molecules into a machine-readable language using **SMILES**.



We built a predictive **QSAR model** that learned the rules connecting structure to activity.

Accelerating the Path to a Treatment



By successfully bridging biology, chemistry, and machine learning, we have created more than just a model. We have developed a validated computational engine to systematically navigate the vast chemical space in search of a cure for Alzheimer's disease.

QSAR