ChemDiv
The chemistry of cures℠

# Premium Dataset for AI-Driven Drug Discovery

Extensively curated dataset made of experimental data from validated assays, designed to maximize AI/ML model performance.

**10,000,000+**
Data points

**615,000+**
unique molecules

**98%**
NOT in ChemBL

**1,300,000+**
ligand-target pairs

**3,000+**
specific MoA related programs

**+30**
years of success in life sciences

# Public Datasets Challenges

## Data Quality and Verification

Literature-derived datasets **lack standardized verification,** with missing experimental details and quality controls.

AI models built on this u**nverified data** may perform well in cross-validation but fail dramatically when applied to novel compounds in real experimental settings.

## Negative Data Samples

Public databases overwhelmingly contain active compounds while **unsuccessful experiments** remain unpublished.

This imbalance leads AI models to overpredict activity and produce high false-positive rates, as they've rarely been trained on truly inactive compounds.

## Consistent Experimental Conditions

Bioactivity and ADMET data comes from experiments with varying protocols, reagents, and conditions.

What appears as a single property (like "solubility" or "IC50") can represent **fundamentally different** measurements depending on experimental setup, causing models to learn artifacts rather than intrinsic molecular properties.

## Chemical Space Coverage

Public datasets have **significant gaps**, focusing heavily on popular scaffolds and "drug-like" compounds.

Most molecules are tested against only a fraction of possible targets, creating an **incomplete** activity matrix that limits model performance when encountering novel chemical structures or target families.

# ChemDiv’s Experimental Data

## Verified Experimental Data

Generated in a single laboratory using a uniform experimental protocols → Ensures reliability and reduces batch variability.

## Balanced Positive & Negative Data

Unlike ChEMBL, which has mostly active compounds, ChemDiv offers both positive and negative data essential for robust and unbiased ML models.

## Comprehensive Chemical Space

Covers a distinct chemical space compared to ChEMBL, enriching diversity in ML model training.

Enhanced predictive power of ML models

## Enhanced predictive power of ML models

Improves generalizability and predictive power of ML models for drug discovery and virtual screening.

## Better ML Performance & Decision-Making

More reliable predictions, reduced false positives/negatives, and better lead identification.

# ChemDiv's Experimental Dataset Value

The database consists of two primary categories of data: **Bioactivity Data** and **ADMET Data**, each providing significant value for research and analysis in drug discovery and pharmacology.

The total estimated data generation cost is **$26,611,705** demonstrating its critical role in advancing drug development, optimizing lead compounds, and reducing the risk of late-stage failures in pharmaceutical research.

## Bioactivity Data

- Contains **1,305,157 unique ligand-target pairs**, which are essential for understanding interactions between chemical compounds and biological targets.
- Data generation cost: **$18,607,505**, reflecting its importance in drug-target interaction research and potential therapeutic applications.

## ADMET Data

- Comprises **6,166 unique compounds** evaluated across multiple absorption, distribution, metabolism, excretion, and toxicity (ADMET) assays.
- Data generation cost: **$8,004,200**, highlighting its significance in predicting the pharmacokinetics and safety profiles of drug candidates.

# Enhancing hERG Inhibition Prediction with ChemDiv Experimental Data

**Significant performance boost** after integrating ChemDiv data

**Better prediction of inactive compounds**, solving data imbalance issue

**More reliable model** for hERG inhibition activity

## Model Performance Comparison

| | Before (TDC Data Only) | After (TDC + ChemDiv Data) |
|---|---|---|
| Accuracy | 0.35 | **0.8** |
| Cohen Kappa | 0.044 | **0.565** |

### Before (TDC Only)

| | Predicted Active | Predicted Inactive |
|---|---|---|
| Actual Active | 6 | 0 |
| Actual Inactive | 13 | 1 |

### After (TDC + ChemDiv)

| | Predicted Active | Predicted Inactive |
|---|---|---|
| Actual Active | **5** | **1** |
| Actual Inactive | **3** | **1** |

### Initial Dataset (TDC Data)

1. Collected from TDC
2. Features: ECFP4 fingerprints (Length: 1024)
3. Developed an XGBoost model using training and validation sets
4. Evaluated performance on the ChemDiv test set

### Data Augmentation (ChemDiv Data)

1. ChemDiv data partitioned into training and test sets
2. ChemDiv training data added to TDC training set
3. Model re-trained and evaluated on ChemDiv test set

# Comparison of Data from ChEMBL 34 and ChemDiv Data

## Unique Data

98% of ChemDiv is not in ChEMBL

## Competitive Advantage

The uniqueness of ChemDiv's proprietary dataset provides an edge for machine learning (ML) model development and customization for customers.

## Molecular Differences

ChemDiv molecules exhibit distinct chemical properties compared to public datasets, as demonstrated in the visualization below.

**560,440**

Unique molecules in **ChemDiv**

**615,719**

ChemDiv unique molecules

**2,409,270**

ChEMBL 34 unique molecules

**55,279**

Union (Overlap)

# Data Coverage

## Bioactivity

Bioactivity datasets were aligned with ligand-receptor interactions, inhibition constants, and potency metrics.

## Cytotoxicity

This included toxicity assays conducted across various cell lines and species to identify potential cytotoxic effects of compounds.

## ADME Properties

Comprehensive ADME profiling was performed using standard assays for solubility, stability, permeability, plasma protein binding, blood-plasma distribution, CYP inhibition, and hERG binding. The specific methodologies for these assays are outlined below.

**12730 High Bluff Dr, Suite 100**
**San Diego CA 92130 USA**

Tel: +1 858-794-4860
Fax: +1 858-794-4931

chemdiv@chemdiv.com