Vancouver, Canada, 2024

WelQrate

Defining the Gold Standard in Small Molecule Drug Discovery Benchmarking

Yunchao (Lance) Liu*[1]

Ha Dong*[2]

Xin Wang*[1]

Rocco Moretti[1]

Yu Wang[3]

Zhaoqian Su[1]

Jiawei Gu[4]

Bobby Bodenheimer[1]

Charles David Weaver[1]

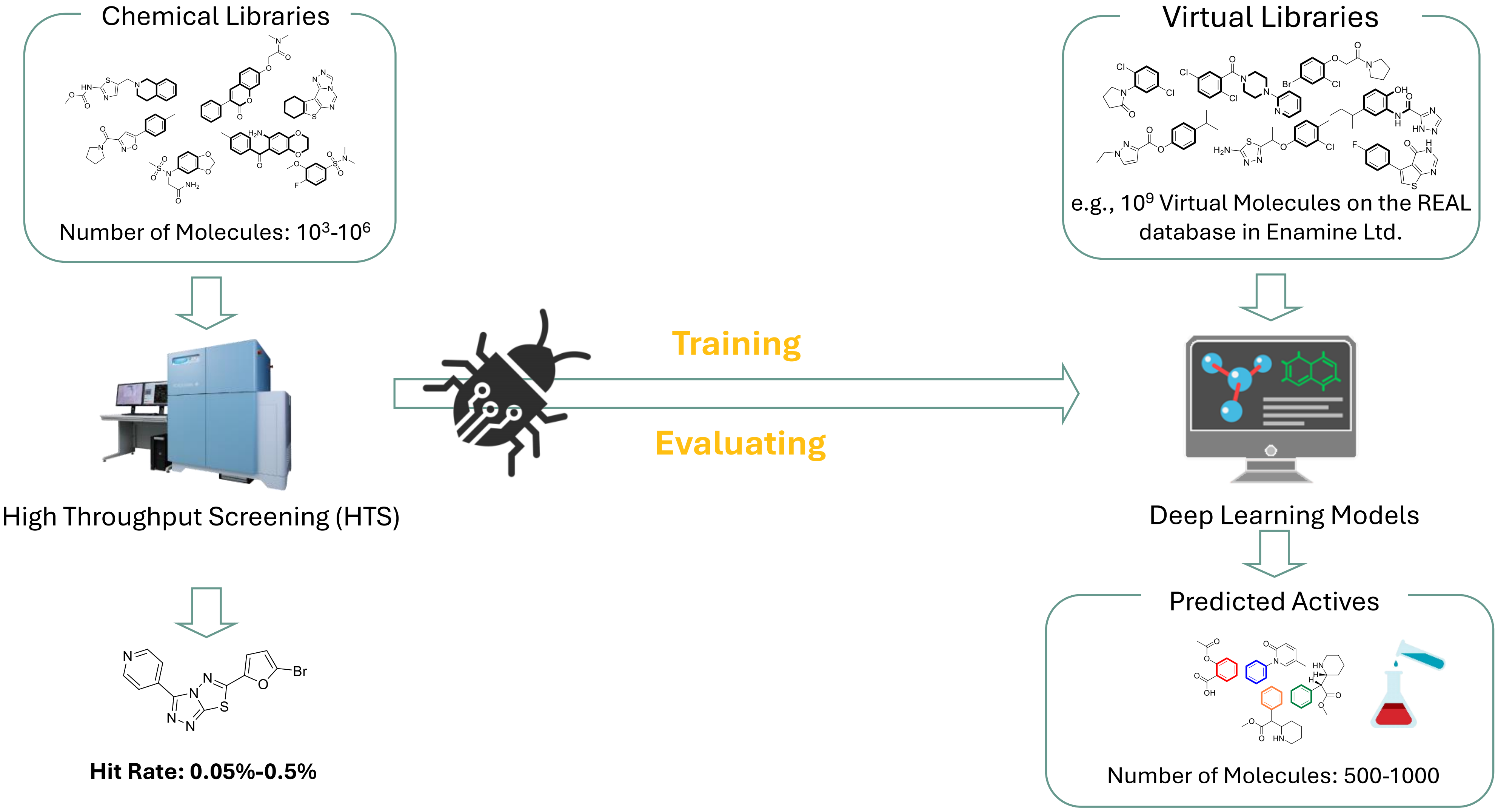
Jens Meiler[1,5]

Tyler Derr[1]

* Equal Contribution

[1]Vanderbilt University, [2]Amherst College, [3]University of Oregon, [4]MD Anderson Cancer Center, [5]Leipzig University

Introduction



Chemical Libraries

Number of Molecules: 10^3-10^6

High Throughput Screening (HTS)

Hit Rate: 0.05%-0.5%

Training

Evaluating

Virtual Libraries

e.g., 10^9 Virtual Molecules on the REAL database in Enamine Ltd.

Deep Learning Models

Predicted Actives

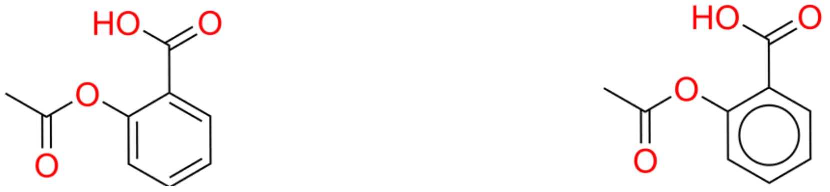
Number of Molecules: 500-1000

Current Issues

Data

1. Inconsistent Representations

e.g. from MoleculeNet’s BBBP dataset



2. Noisy Data Labels

ID	aspirin	acetylsalicylate
SMILES	C1=CC=CC(=C1C(O)=O)OC(C)=O	CC(=O)Oc1ccccc1C(O)=O
Label	0	1

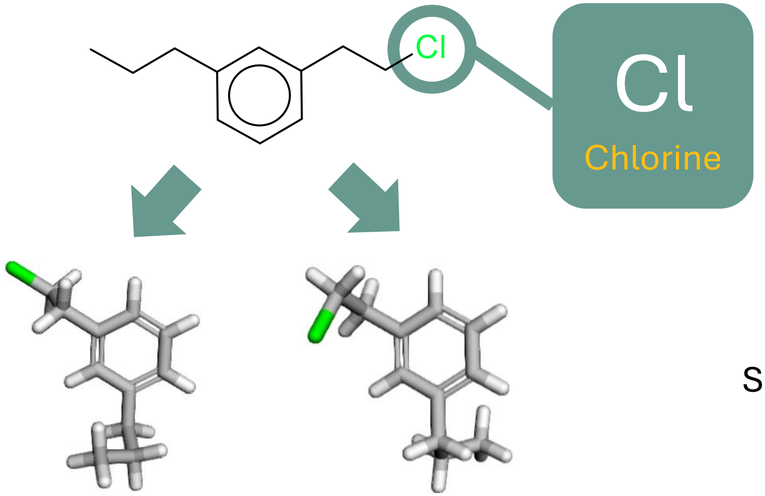
Evaluation

1. Variation in Featurization

e.g.

2. Variation in 3D Conformations

e.g.



0	0	0	0	0	1	0	0	0

0.5	0.8	2.5	2.6	1.2	0.1	7.2	2.5	3.3

See **more** issues listed in our paper

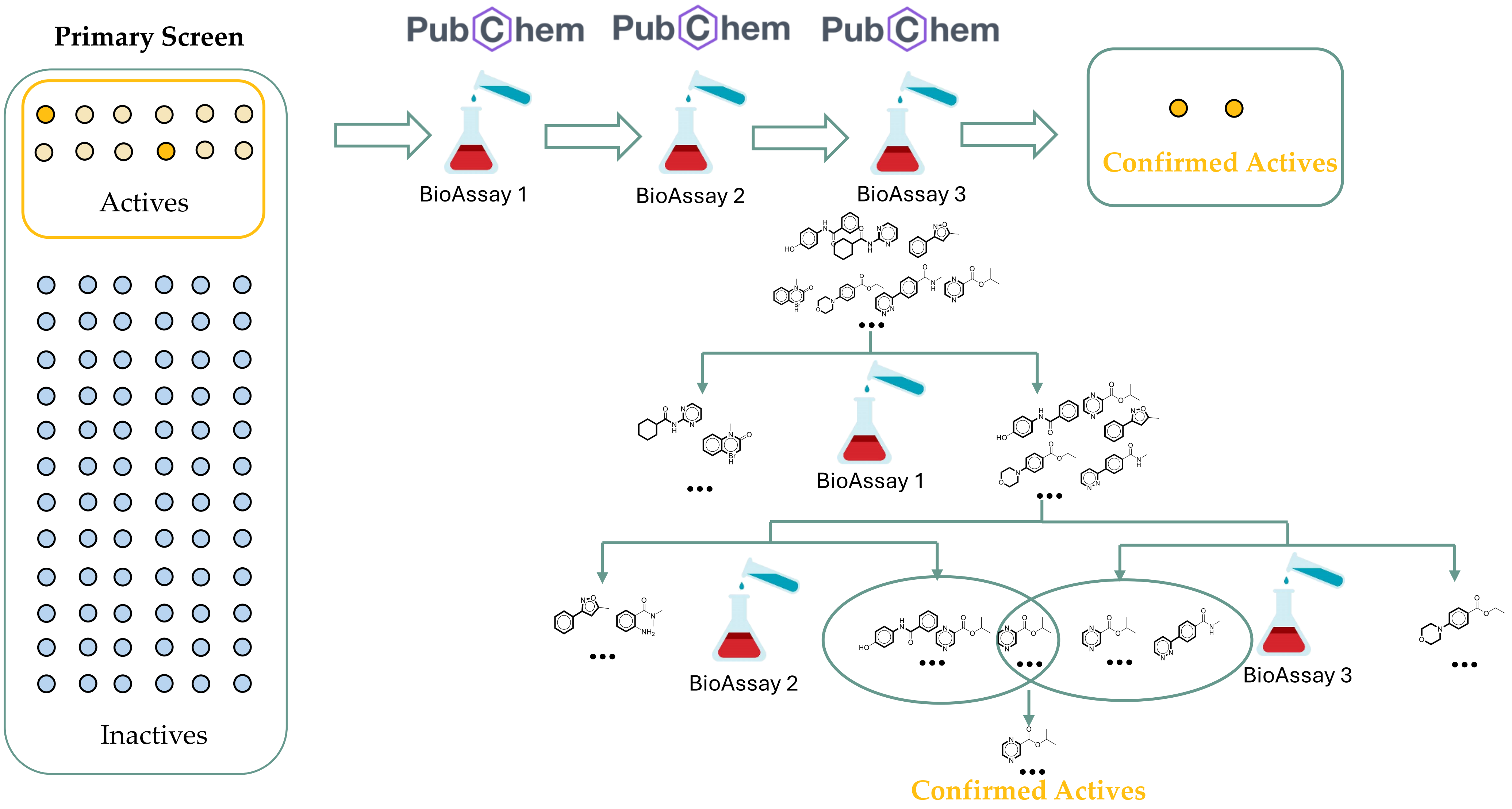
WelQrate-Evaluation Framework

Datasets
High Quality

Evaluation Metrics
Realistic

Data Splits
Robust

WelQrate Data Curation



Primary Screen

Actives

Inactives

PubChem PubChem PubChem

BioAssay 1 BioAssay 2 BioAssay 3

Confirmed Actives

BioAssay 1

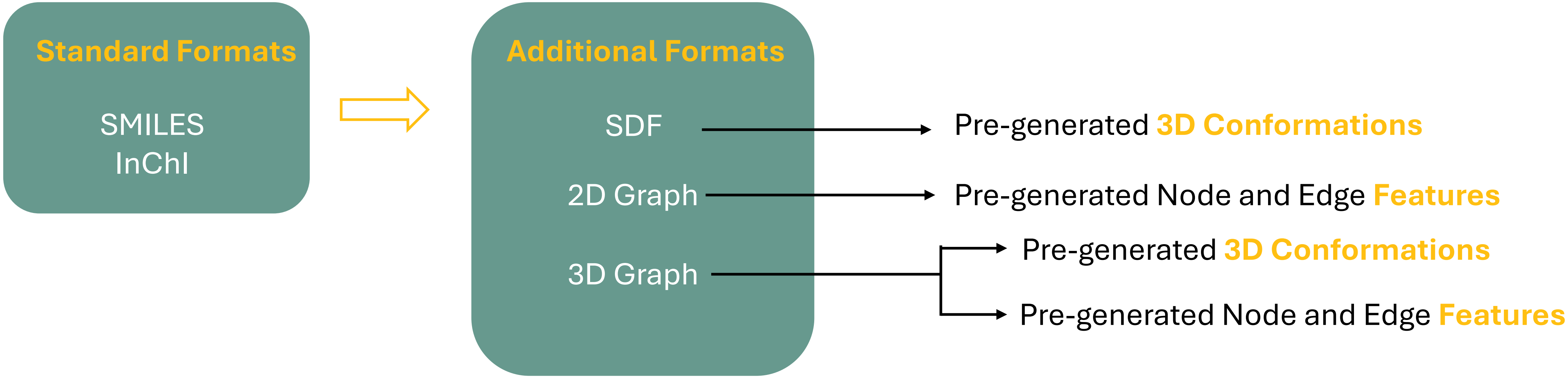
BioAssay 2

BioAssay 3

Confirmed Actives

Multiple filters (e.g., duplicate) are in the curation pipeline as well

Data Format



Standard Formats

SMILES
InChI

Additional Formats

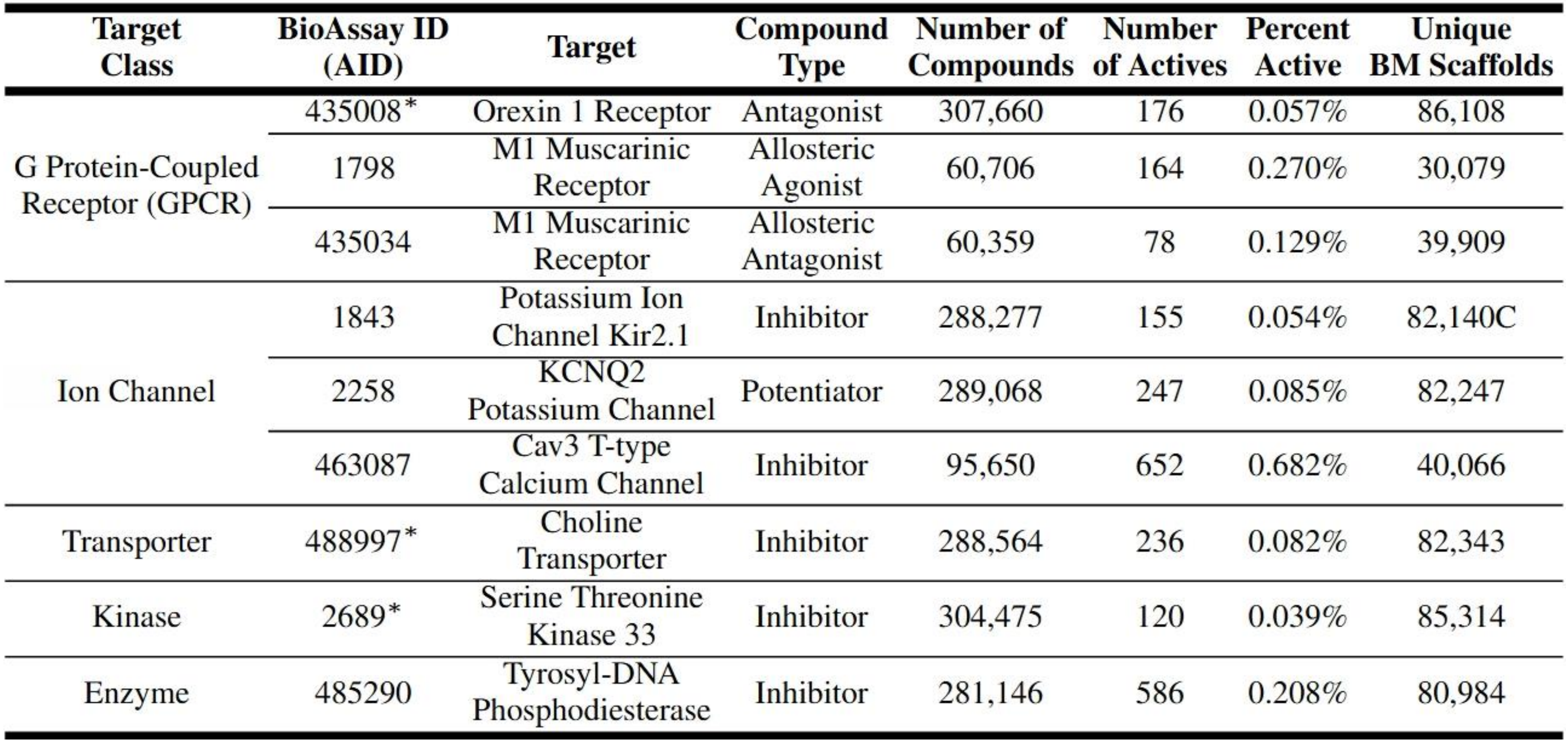
SDF → Pre-generated **3D Conformations**

2D Graph → Pre-generated Node and Edge **Features**

3D Graph → Pre-generated **3D Conformations**
3D Graph → Pre-generated Node and Edge **Features**

We encourage the researchers to innovate on featurization and generate conformations.

WelQrate Dataset Collection Statistics



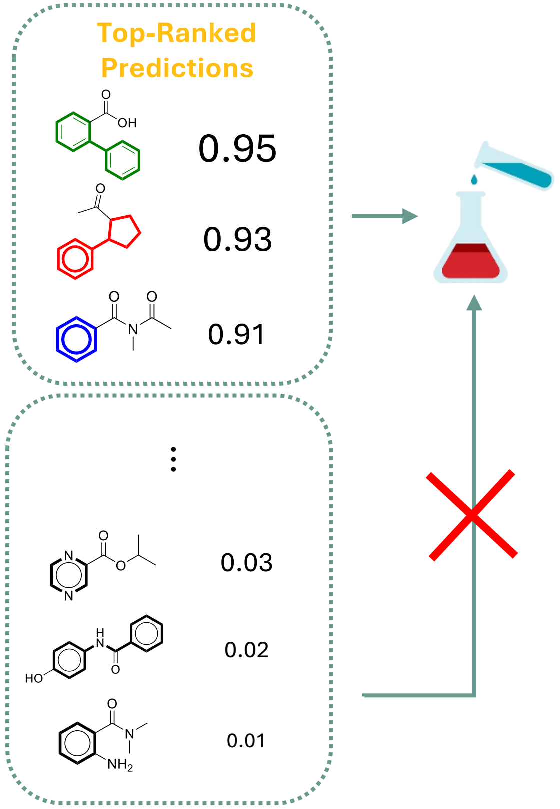
Target Class	BioAssay ID (AID)	Target	Compound Type	Number of Compounds	Number of Actives	Percent Active	Unique BM Scaffolds
G Protein-Coupled Receptor (GPCR)	435008*	Orexin 1 Receptor	Antagonist	307,660	176	0.057%	86,108
	1798	M1 Muscarinic Receptor	Allosteric Agonist	60,706	164	0.270%	30,079
	435034	M1 Muscarinic Receptor	Allosteric Antagonist	60,359	78	0.129%	39,909
Ion Channel	1843	Potassium Ion Channel Kir2.1	Inhibitor	288,277	155	0.054%	82,140C
	2258	KCNQ2 Potassium Channel	Potentiator	289,068	247	0.085%	82,247
	463087	Cav3 T-type Calcium Channel	Inhibitor	95,650	652	0.682%	40,066
Transporter	488997*	Choline Transporter	Inhibitor	288,564	236	0.082%	82,343
Kinase	2689*	Serine Threonine Kinase 33	Inhibitor	304,475	120	0.039%	85,314
Enzyme	485290	Tyrosyl-DNA Phosphodiesterase	Inhibitor	281,146	586	0.208%	80,984

* Indicates additional experimental measurements are available.

Features of the Dataset Collection

- Diverse important therapeutic target classes
- Large number of molecules
- High quality label
- Realistic label imbalance

Evaluation Metrics



logAUC$_{[0.001,\ 0.1]}$

Logarithmic Receiver-Operating-Characteristic Area Under the Curve with the False Positive Rate in the Range [0.001, 0.1]

BEDROC

Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic

EF$_{100}$

Enrichment Factor with Cutoff 100

DCG$_{100}$

Discounted Cumulative Gain with Cutoff 100

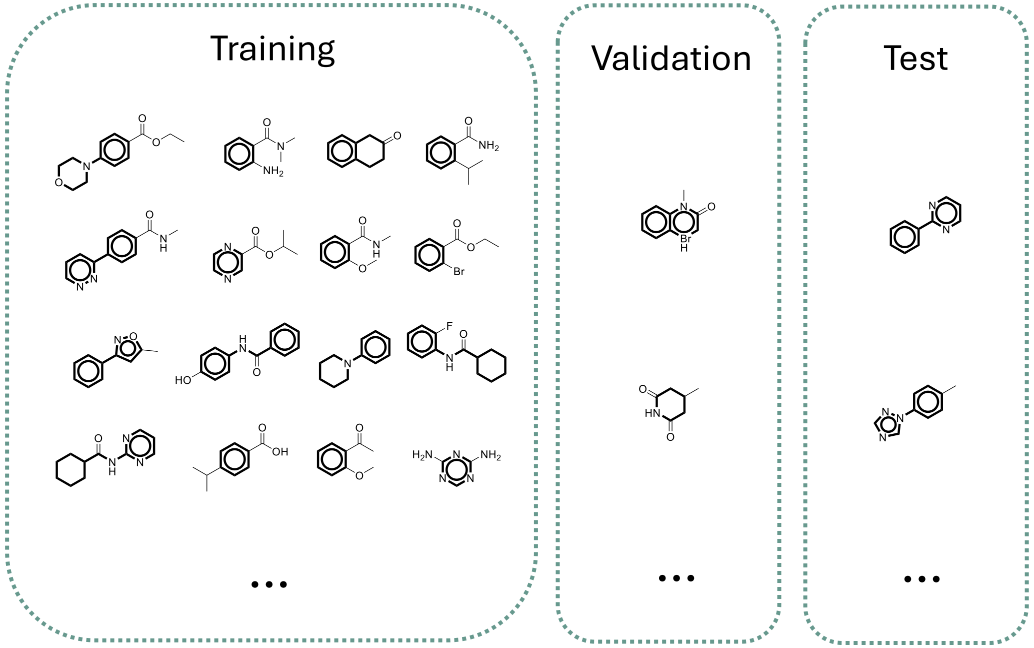
More metrics with real-world considerations are encouraged

*See paper supplement for the details of the metrics

Data Split

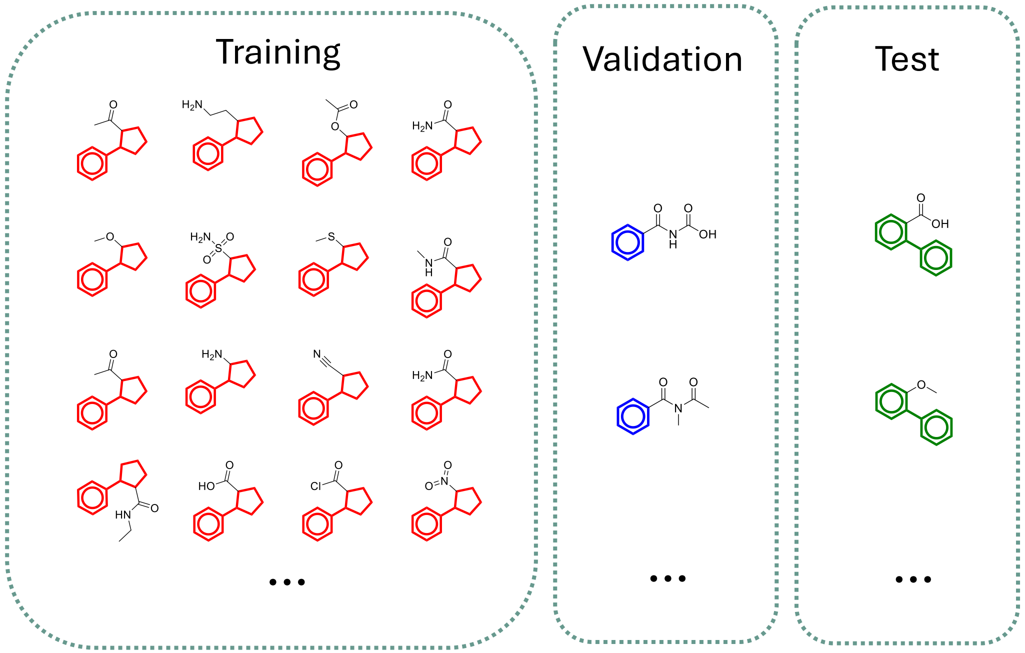
Random

Split Scheme



Scaffold

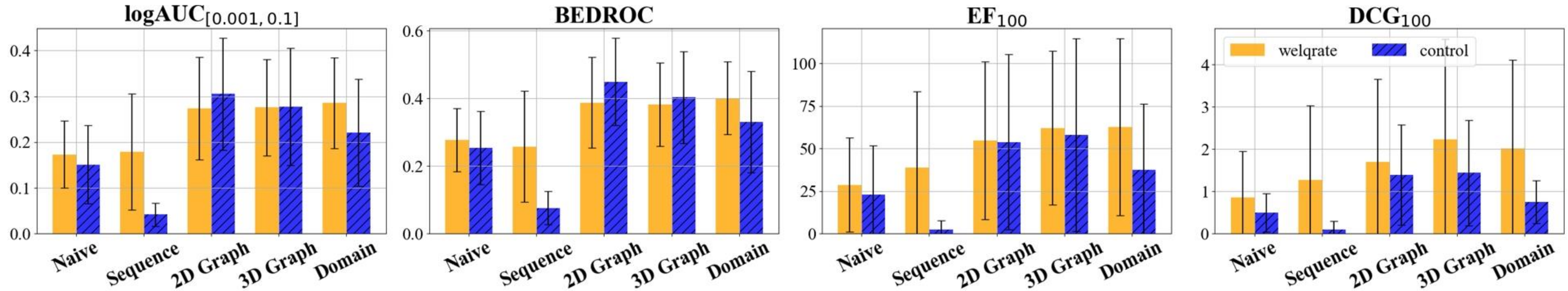
Split Scheme



Each Scheme Provides Five Different Splits Per Dataset
The Reported Performance Should Be Averaged From the Five Splits
to Ensure the Evaluation Robustness

Benchmarking

RQ1: How Do Different **Models** And **Data Representation** Affect Performance?
RQ2: How Does **Datasets Quality** Impact Model Evaluation?
RQ3: How Significant Is **Featurization** in Model Evaluation?
RQ4: How Do Different Models Perform Under **Scaffold Splitting**?



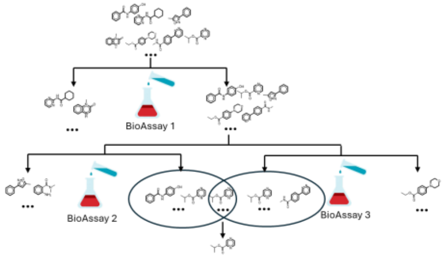
Results for RQ1 & RQ2

Thank You!

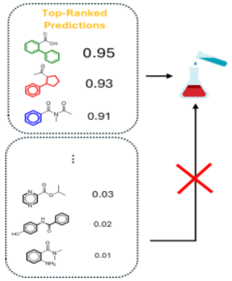
WelQrate

www.WelQrate.org

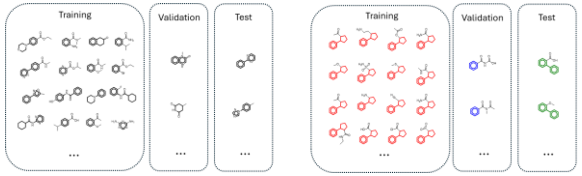
Datasets – High Quality



Evaluation Metrics - Realistic



Data Split – Robust



NIH DFG NSF NVIDIA Alexander von Humboldt Stiftung / Foundation