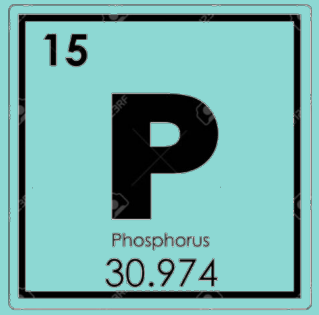


PurifAI

Machine Learning for SPE & LCMS Method Prediction

Created by Jen Amis, Luke Perrin, Teresa Tran, Yingying Cheung

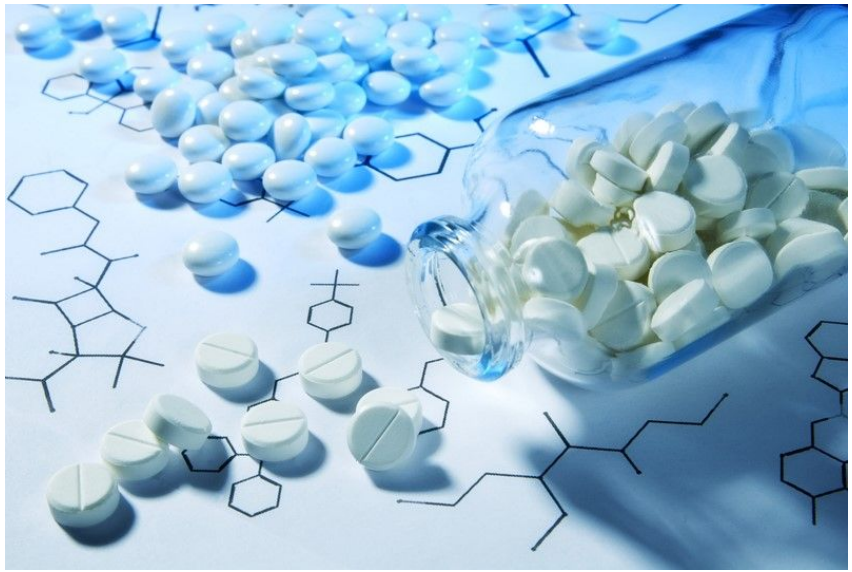
Sections

1. Introduction & Context/Problem (1.5 mins) – Luke
2. Data, Data Source, Database (1mins) – Luke
3. ML Model (2.5 mins) - Jen
4. Dashboard (visualization) and interactive webpage(upload a file for prediction) (2.5 mins) –Teresa
5. Application package (1 mins) –Yingying
6. Conclusion and future industrial application (1.5 mins) - Yingying

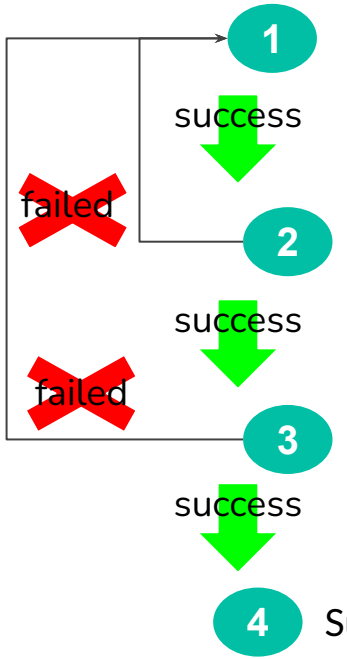
Context & Problem

Context & Problem

- The development of a completely automated chemistry platform is of high importance for the drug discovery process
    - By automating chemical synthesis and purification, many **varieties of novel drug candidates (or chemical structures)** can be tested for pharmacological activity
- Chemistry as a scientific field/practice is inherently **unpredictable and (seemingly) inconsistent,** and thus **very difficult to predict and automate**



Chemistry Purification Workflow Overview

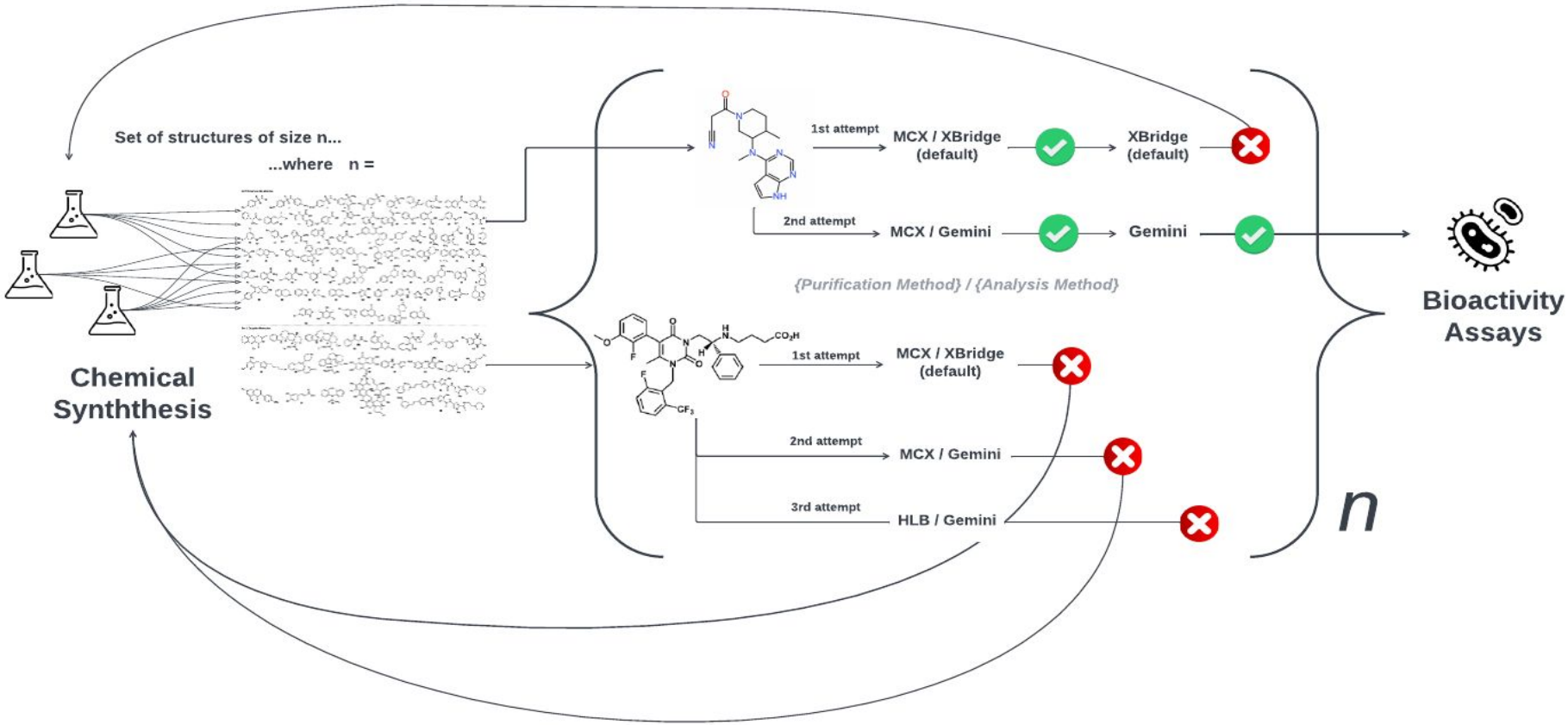


Library (set) of molecules synthesized (made)

Each molecule in the library is **purified** by one of two ***Solid Phase Extraction (SPE)*** methods

Each molecule in the library is **analyzed** by one of two ***Liquid-Chromatography Mass-Spectroscopy (LCMS)*** methods

Successfully purified molecules proceed to bioassays (testing for pharmacological activity)

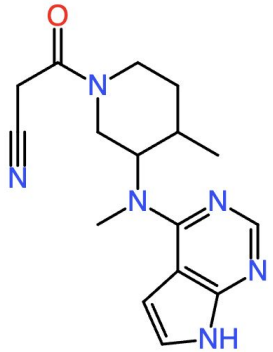

Set of structures of size n...
...where n =
Chemical Synthesis
1st attempt
MCX / XBridge (default)
XBridge (default)
2nd attempt
MCX / Gemini
Gemini
{Purification Method} / {Analysis Method}
1st attempt
MCX / XBridge (default)
2nd attempt
MCX / Gemini
3rd attempt
HLB / Gemini
Bioactivity Assays
n

About the Data

Data Sources

1. **Purification Outcomes**
    a. Successfully purified via either SPE method:
        i. MCX
        ii. HLB
    b. Successfully analyzed via either LCMS method:
        i. Xbridge
        ii. Gemini
2. **Chemical Structure Data**
    a. Data/metrics calculated from chemical structures (purified on the automated platform)
    b. These describe chemical properties of each molecule

Purification “Outcomes”



sample_id	structure_id	preferred_lcms_method	spe_method	method	spe_successful	crashed_out	sample_status
00YLL22-042-014	00YLL22-042-014	Xbridge HpH	MCX	MCX/Xbridge HpH	true	NULL	Complete
00YLL22-042-015	00YLL22-042-015	Gemini LpH	HLB	HLB/Gemini LpH	NULL	NULL	Failed
00YLL22-042-016	00YLL22-042-016	Gemini LpH	MCX	MCX/Gemini LpH	true	NULL	Complete
00YLL22-042-017	00YLL22-042-017	Gemini LpH	MCX	MCX/Gemini LpH	true	NULL	Complete
00YLL22-042-018	00YLL22-042-018	Gemini LpH	MCX	MCX/Gemini LpH	true	NULL	Complete

Example **successfully purified** molecule
(structure_id = 00YLL-042-016)

LCMS Method = **Gemini**

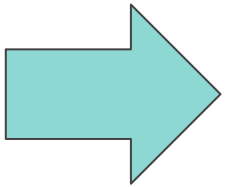
SPE Method = **MCX**

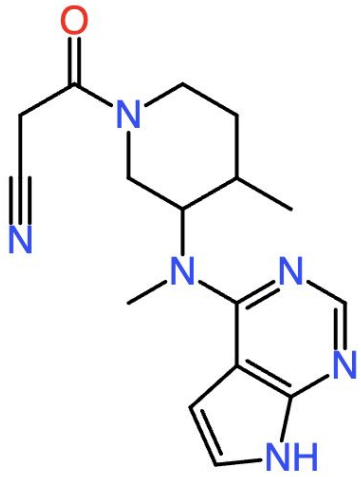
Purification Successful = **true**

Chemical Structure Data

`CC1CCN(CC1N(C)C2=NC=NC3=C2C=CN3)C(=O)CC#N`

SMILES representation (text string representation of a 3D molecular structure)

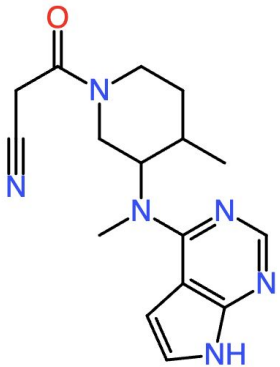




Molecular depiction

Generation of Chemical Structure Data

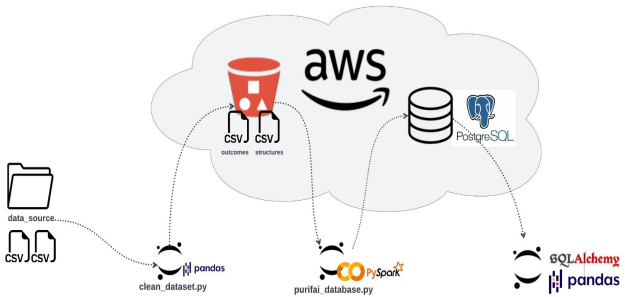
structure_id	MolWt	exactMolWt	qed	TPSA	HeavyAtomMolWt	MolLogP	MolMR	FractionCSP3
00YLL22-042-011	475.336	474.0973939	0.589825726	94.54	455.176	2.9629	119.2907	0.333333333
00YLL22-042-012	446.338	445.1072303	0.667154151	74.23	425.17	3.6913	115.3657	0.380952381
00YLL22-042-013	462.337	461.1021449	0.607254623	83.46	441.169	2.9293	116.9727	0.380952381
00YLL22-042-014	446.338	445.1072303	0.667154151	74.23	425.17	3.6913	115.3657	0.380952381
00YLL22-042-015	435.311	434.0912459	0.666314391	72.38	415.151	3.4859	110.1327	0.4
00YLL22-042-016	450.326	449.1021449	0.625596379	92.25	429.158	2.8315	114.4374	0.35
00YLL22-042-017	450.326	449.1021449	0.648315106	83.4				
-042-018	464.334	463.0272657	0.642066735	91				
-042-019	446.338	445.1072303	0.669017878	74.				

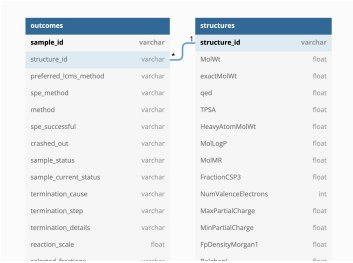


Example molecule
(structure_id = 00YLL-042-016)

- Every molecule has a set of calculated attributes/features called ***molecular descriptors*** that describe the **chemical properties** of a structure
- This was accomplished using the open-source python library RdKit

Data Pipeline, Storage, & Retrieval





Pipeline

1. Data is pulled from company database:
   a. RdKit used to calculate chemical descriptors from SMILES
   b. Pandas used to clean the purification outcomes dataset
2. Cleaned data is stored in AWS S3 buckets
3. Cleaned data is sourced from buckets and written to AWS RDS (postgres) using Pyspark (*see partial schema to the left*)

Storage

- Data is stored in AWS RDS instance

Retrieval

- Data is retrieved from AWS RDS using SQLAlchemy
  - ‘outcomes’ and ‘structures’ are merged into one table, joining on ‘structure_id’
- Data is further processed for ML modeling using Pandas

Machine Learning

Goal

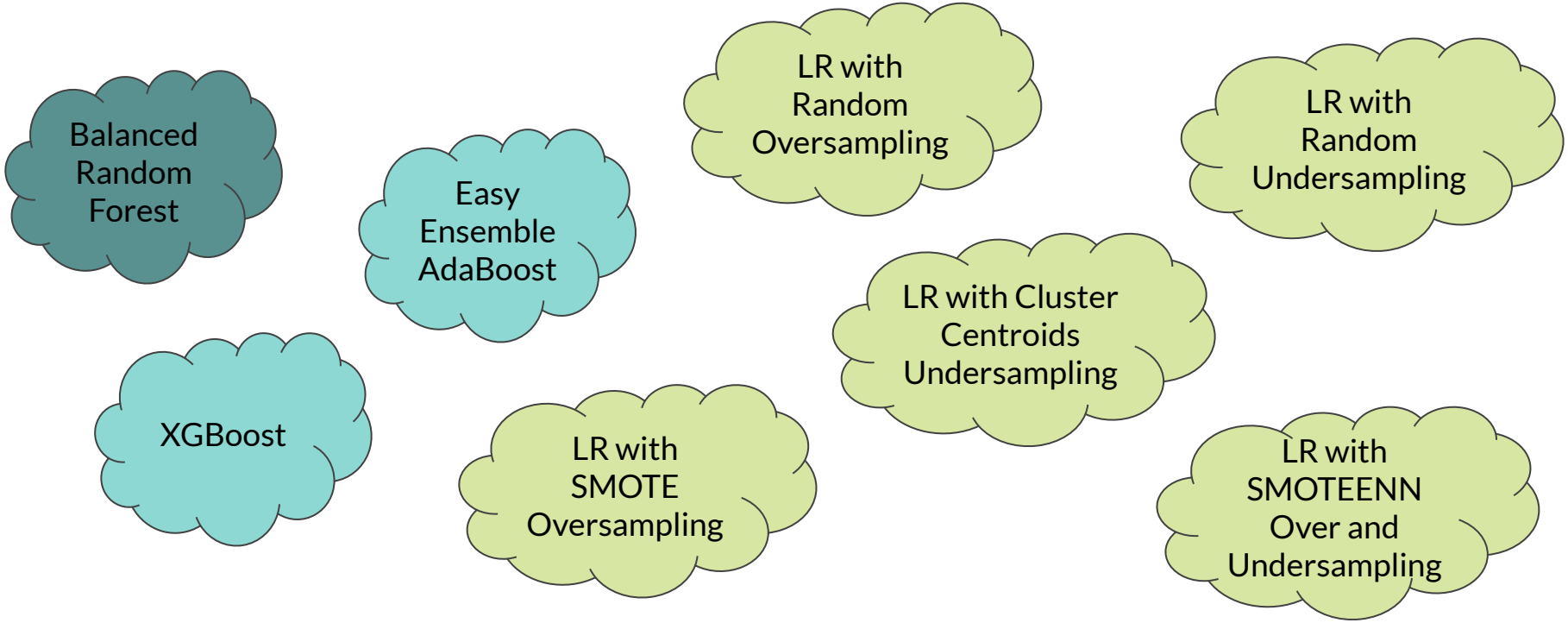
Develop two supervised ML models:

1. Predict optimal **SPE method** for compound purification
2. Predict optimal **LCMS method** for compound analysis

Variables

Features 45 molecular descriptors	
SPE Method	**LCMS Method**
Target MCX: 873 HLB: 185	**Target** Xbridge: 729 Gemini: 319

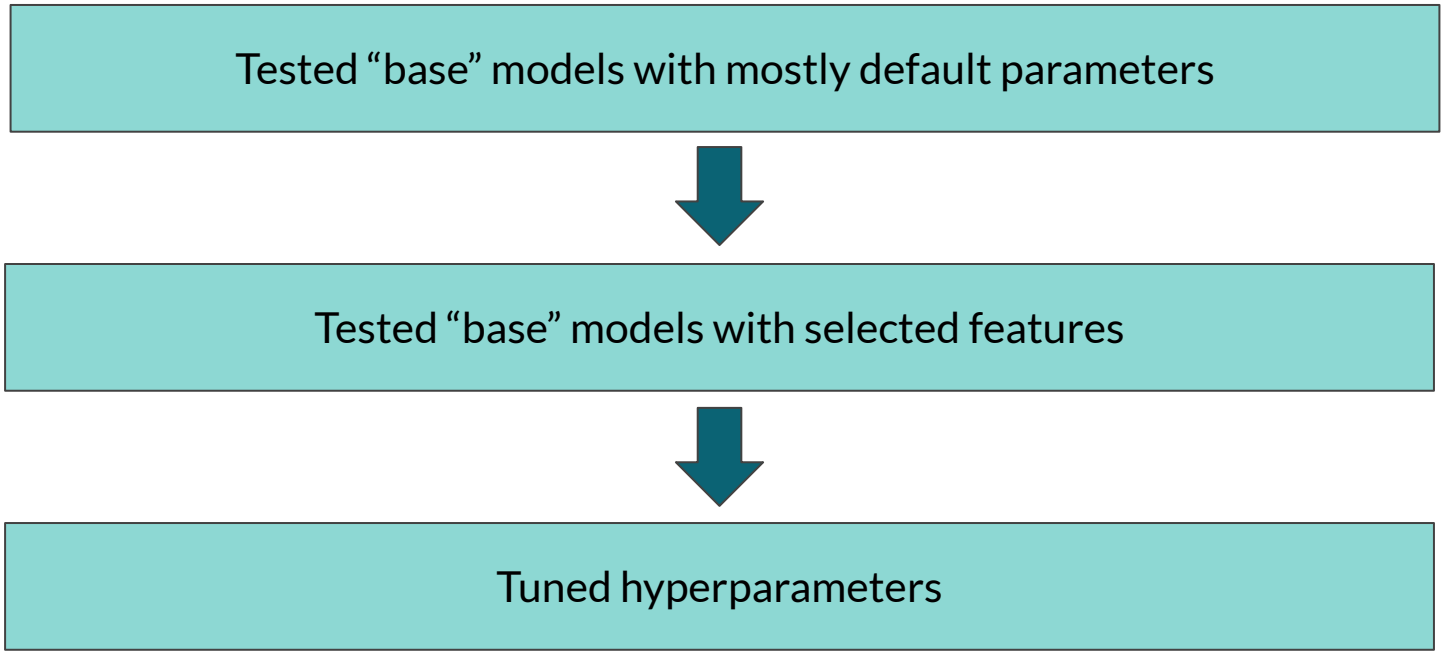
Candidate Models



Data Preprocessing

1. Only kept rows where compound successfully completed purification stage of testing
2. For LCMS model, dropped rows with null or very rare LCMS method
3. Dropped duplicate rows
4. Scaled features

Model Testing



SPE Model Selection - XGBoost

Balanced Accuracy Score: 0.8938679245283019

Confusion Matrix:

	Predicted HLB	Predicted MCX
Actual HLB	43	10
Actual MCX	5	207

Imbalanced Classification Report:

	pre	rec	spe	f1	geo	iba	sup
HLB	0.90	0.81	0.98	0.85	0.89	0.78	53
MCX	0.95	0.98	0.81	0.97	0.89	0.81	212
avg / total	0.94	0.94	0.84	0.94	0.89	0.80	265

LCMS Model Selection - XGBoost

Balanced Accuracy Score: 0.8857998885172798

Confusion Matrix:

	Predicted Gemini LpH	Predicted Xbridge HpH
Actual Gemini LpH	64	14
Actual Xbridge HpH	9	175

Imbalanced Classification Report:

	pre	rec	spe	f1	geo	iba	sup
Gemini LpH	0.88	0.82	0.95	0.85	0.88	0.77	78
Xbridge HpH	0.93	0.95	0.82	0.94	0.88	0.79	184
avg / total	0.91	0.91	0.86	0.91	0.88	0.78	262

Final Model Selection - XGBoost

SPE Method Prediction Model

Balanced Accuracy Score	0.89
Weighted F1 Score	0.94

	Precision	**Recall**
HLB	0.90	0.81
MCX	0.95	0.98

LCMS Method Prediction Model

Balanced Accuracy Score	0.89
Weighted F1 Score	0.91

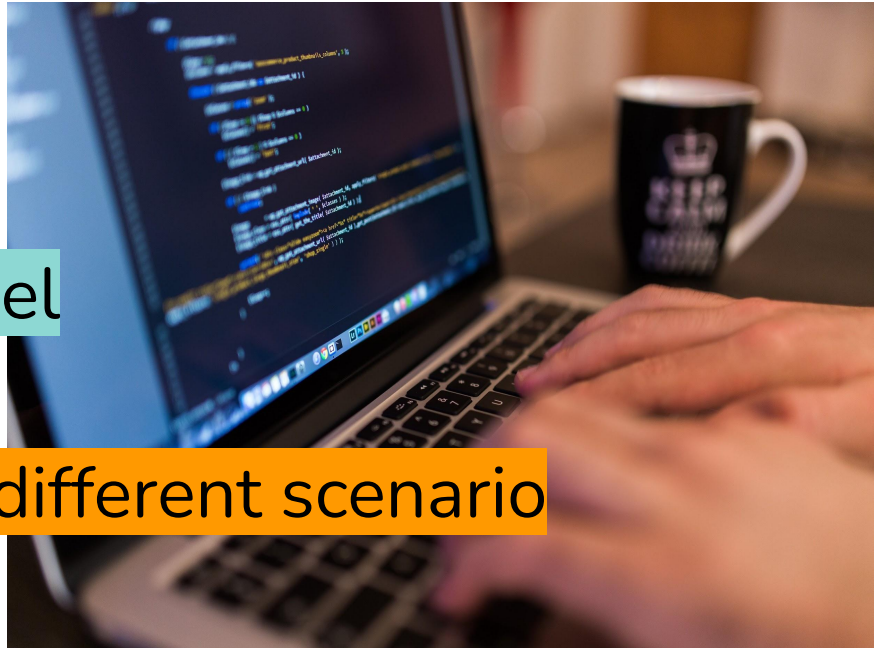
	Precision	**Recall**
Gemini	0.88	0.82
Xbridge	0.93	0.95

Dashboard

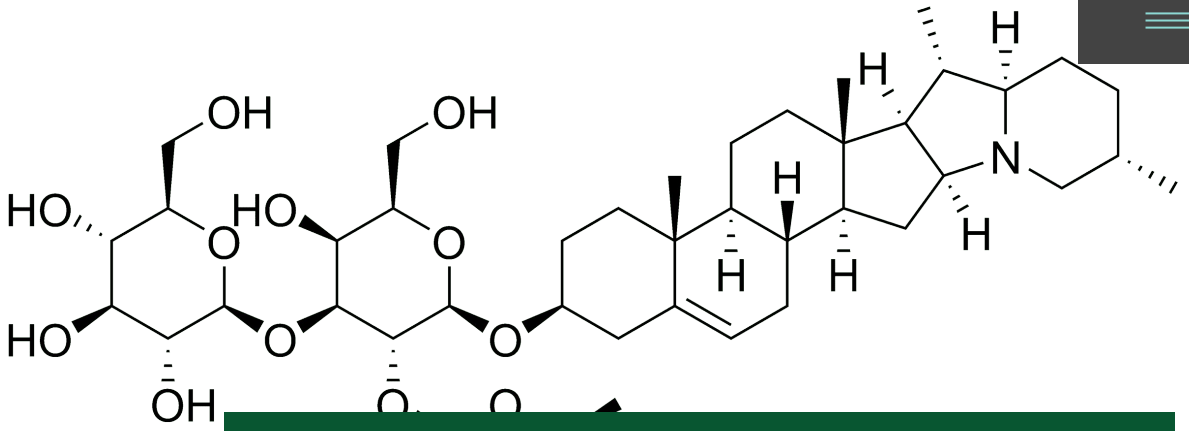
PurifAI Package

Pip install purifAI

- Bulk prediction
- Data scientist friendly
  - Free to change model
- Developer friendly
  - Free to integrate in different scenario



Functions

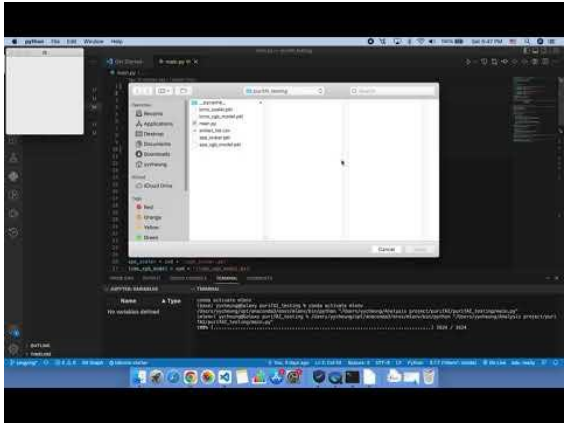


calculate_descriptors() — To convert input SMILES to features

RunSPEPrediction() — To output a SPE method prediction

RunLCMSPrediction() — To output a LCMS method prediction

Demo



Conclusion

- Increasing success quantity metrics
- Minimize failed sample counts
- Limit the amount of time and resources taken
- Identify structures that require new methods
- Open doors for applying ML in other obstacles

Questions?