PREFER : a new PREdictive modeling FramEwoRk for molecular discovery

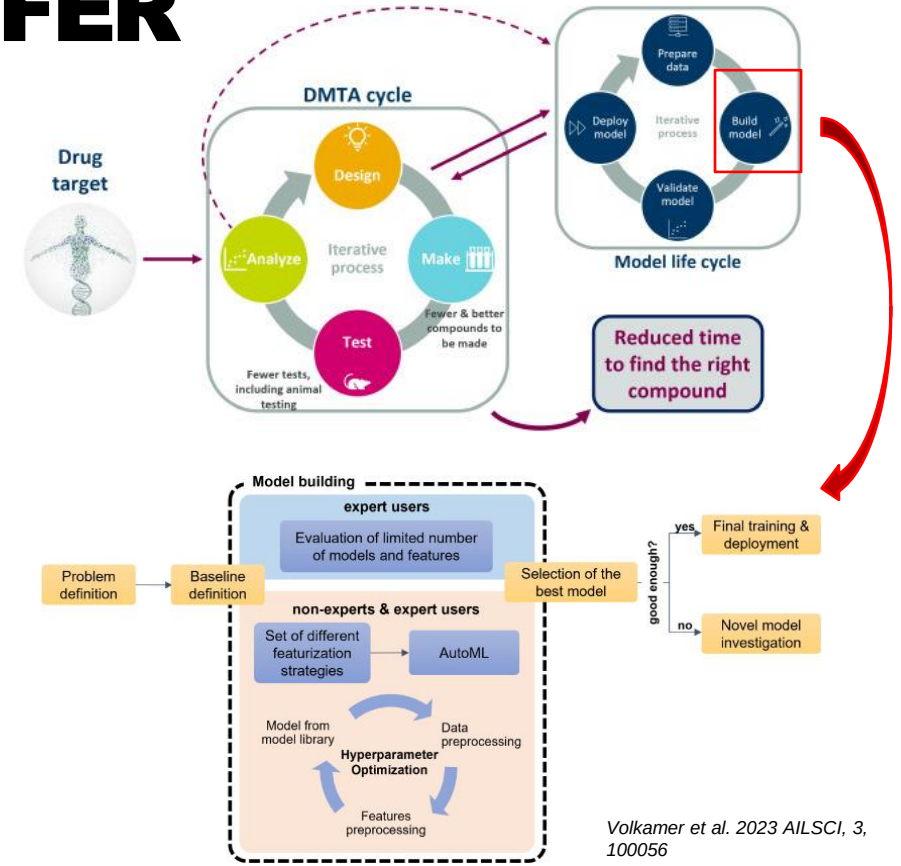
RDKit UGM – September 2023

Jessica Lanini

NOVARTIS | Reimagining Medicine

Introduction to PREFER

- Increased usage machine learning in drug discovery to predict molecular properties
- The goal is to support and speedup the DMTA cycle
- Model life cycle automates the requirements and processes for operationalizing a model
- Model building implies many steps and design decisions



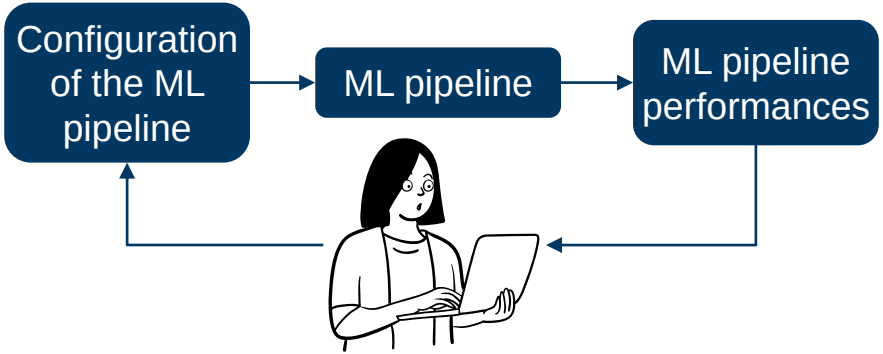
Volkamer et al. 2023 AILSCI, 3, 100056

NOVARTIS | Reimagining Medicine

Introduction to PREFER

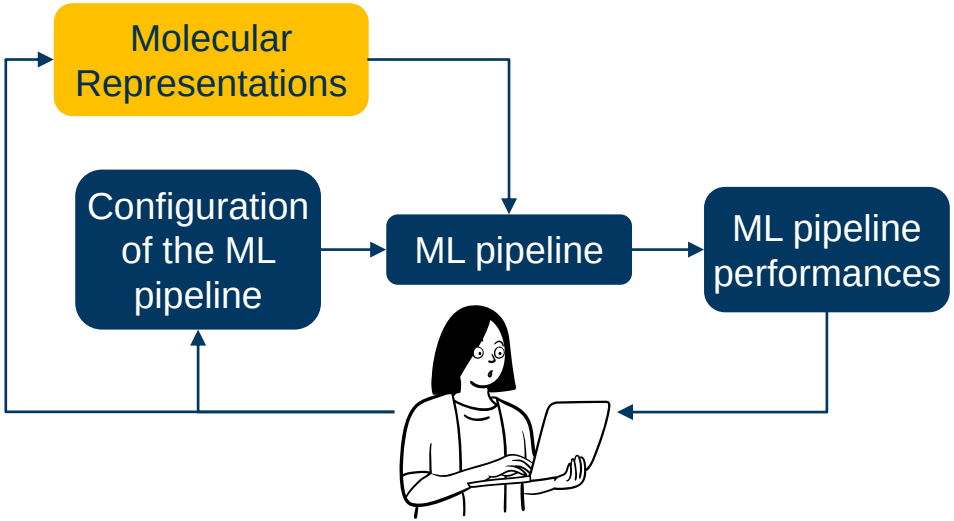
- Design decisions can include
  - Data preprocessors
  - ML algorithms
  - Hyperparameters values of the selected ML algorithm

- Automated selection and evaluation of the different possibilities
  - Human
  - Random search
  - Grid Search
  - Bayesian Optimization

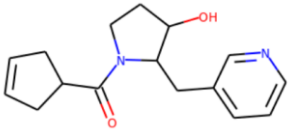


Introduction to PREFER

- Design decisions can include
  - Data preprocessors
  - ML algorithms
  - Hyperparameters values of the selected ML algorithm
  - Molecular representations
- Automated selection and evaluation of the different possibilities
  - Human
  - Random search
  - Grid Search
  - Bayesian Optimization



Molecular Property Predictions



CHEMBL3470905

Classical ML models + Traditional descriptors

Property:
- Solubility
- Lipophilicity
- Permeability
- ...

Molecular Property Predictions

CHEMBL3470905

Classical ML models + Traditional descriptors

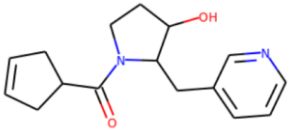
2D Descriptors

Fingerprints

Property:
- Solubility
- Lipophilicity
- Permeability
- ...

RDKit
Open-Source Cheminformatics and Machine Learning

Molecular Property Predictions



CHEMBL3470905

Classical ML models + Traditional descriptors

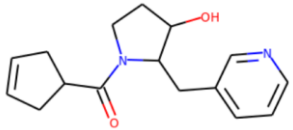
Classical ML models + Improved expert-engineered descriptors[1]

Property:
- Solubility
- Lipophilicity
- Permeability
- ...

NOVARTIS | Reimagining Medicine

[1] Finkelmann, Arndt R. et al. "Site of metabolism prediction based on ab initio derived atom representations."

Molecular Property Predictions



CHEMBL3470905

Classical ML models + Traditional descriptors

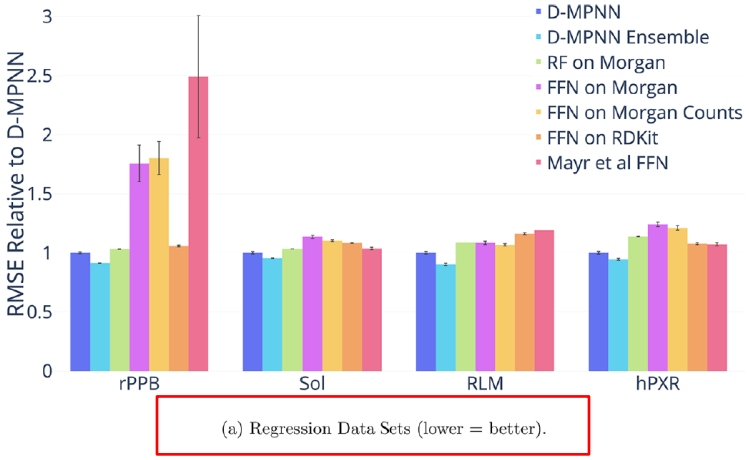
Classical ML models + Improved expert-engineered descriptors

DNN & representation learning

Property:
- Solubility
- Lipophilicity
- Permeability
- ...

Comparison between fixed descriptors and learned molecular representations

- New hybrid model that combines convolution and descriptors (D-MPNN) [2]
- The D-MPNN model matches or outperforms the baseline models[2]
- Performances drop for complex tasks under data scarcity[3]

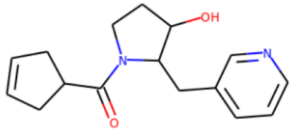


(a) Regression Data Sets (lower = better).

NOVARTIS | Reimagining Medicine

[2] Yang, Kevin, et al. "Analyzing learned molecular representations for property prediction." Journal of chemical information and modeling 59.8 (2019): 3370-3388.
[3] Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.;Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. Chem. Sci. 2018, 9, 513−530.

Molecular Property Predictions



CHEMBL3470905

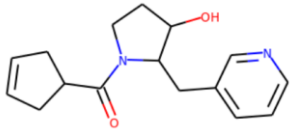
Classical ML models + Traditional descriptors

Classical ML models + Improved expert-engineered descriptors

DNN & representation learning

Property:
- Solubility
- LogD
- Permeability
- ...

Molecular Property Predictions



CHEMBL3470905

Classical ML models + Traditional descriptors

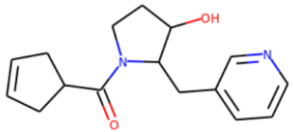
Classical ML models + Improved expert-engineered descriptors

DNN & representation learning

DNN / Classical ML models + Unsupervised data-driven representation

Property:
- Solubility
- LogD
- Permeability
- ...

Autoencoders for data-driven molecular representation

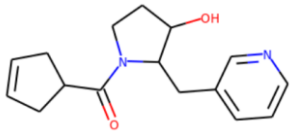


CHEMBL3470905

Encoder

Molecular Descriptors

Decoder

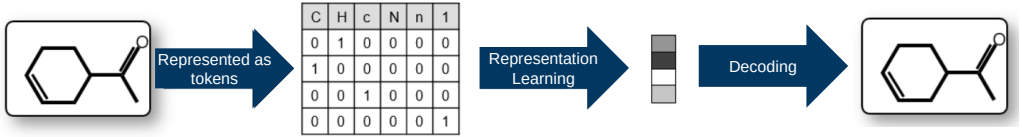


CHEMBL3470905

[4] Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276.
[5] Winter, Robin, et al. "Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations." Chemical science 10.6 (2019): 1692-1701.

Unsupervised and Data-Driven Representations

String-based methods (e.g. CDDD[6])



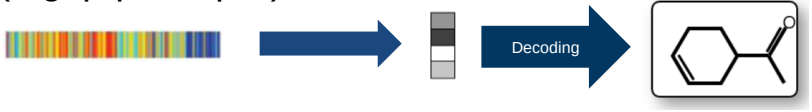
[6] Winter, R. et al. Chem. Sci. 10, 1692–1701 (2019)
[7] Jin, W. et al. arXiv (2019).
https://arxiv.org/pdf/1802.04364.pdf
[8] Maziarz, K. et al. arXiv (2021)
https://arxiv.org/pdf/2103.03864.pdf
[9] Pikusa M, et al. bioarXiv (2022)
https://biorxiv.org/content/10.1101/2021.12.10.472084v1

Graph-based methods (e.g. CGVAE[7], **MoLeR[8]**)

Microsoft

https://github.com/microsoft/molecule-generation

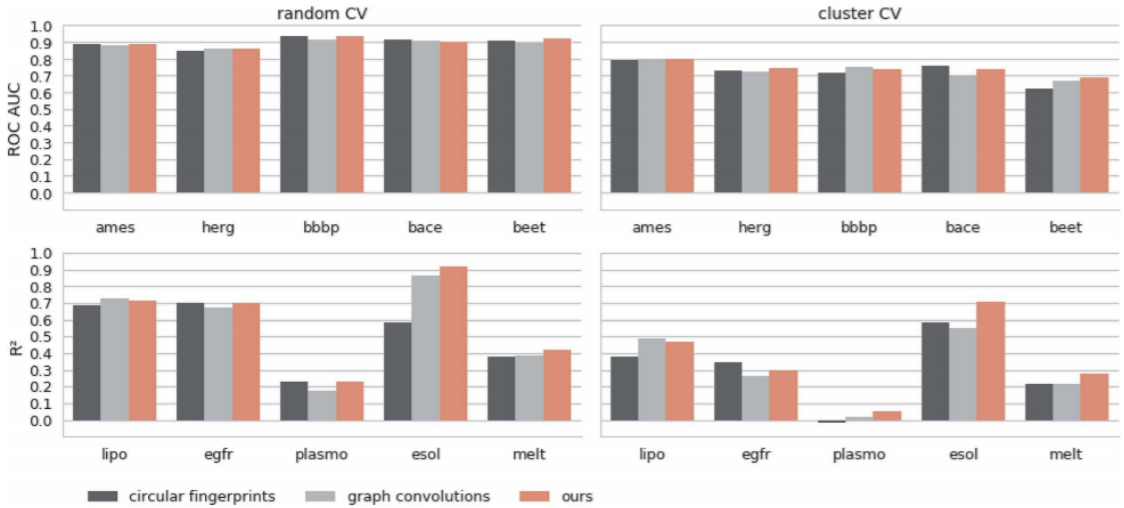
Conditional generation using [signatures, profiles, sequences] (e.g. pqsar2cpd[9])



NOVARTIS | Reimagining Medicine

Unsupervised and Data-Driven Representations: CDDD performances

- Models* based on cddd matches or outperforms models* based on circular fingerprints and GCNN



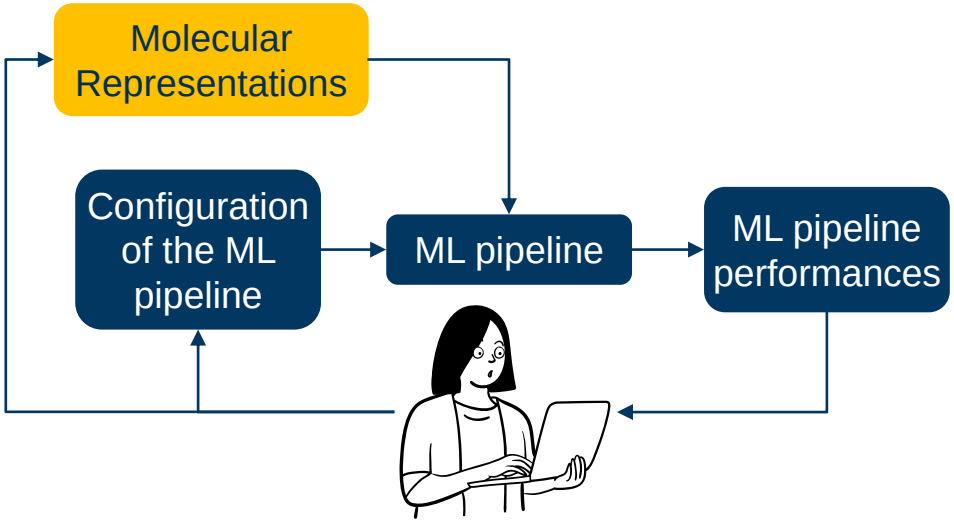
*models: best among SVM, RF and GB

NOVARTIS | Reimagining Medicine

[10] Winter, Robin, et al. "Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations." Chemical science 10.6 (2019): 1692-1701.

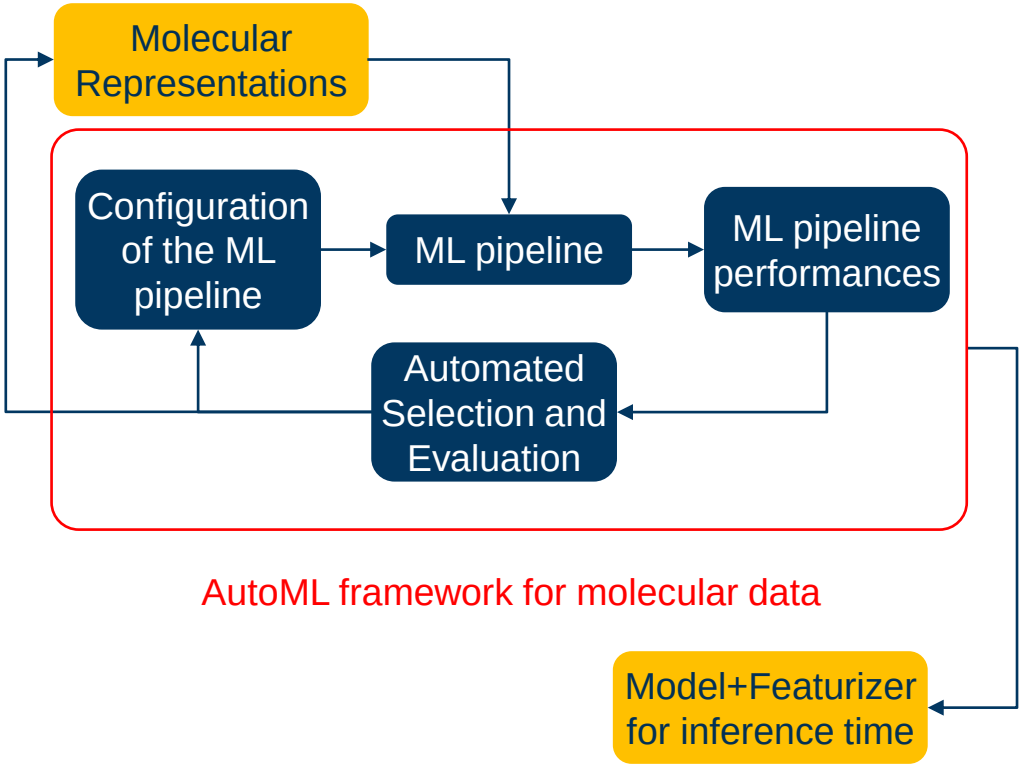
Introduction to PREFER

- Design decisions can include
  - Data preprocessors
  - ML algorithms
  - Hyperparameters values of the selected ML algorithm
  - Molecular representations
- Automated selection and evaluation of the different possibilities
  - Human
  - Random search
  - Grid Search
  - Bayesian Optimization



Introduction to PREFER

- Design decisions can include
  - Data preprocessors
  - ML algorithms
  - Hyperparameters values of the selected ML algorithm
  - Molecular representations
- Automated selection and evaluation of the different possibilities
  - Human
  - Random search
  - Grid Search
  - Bayesian Optimization



State of the art

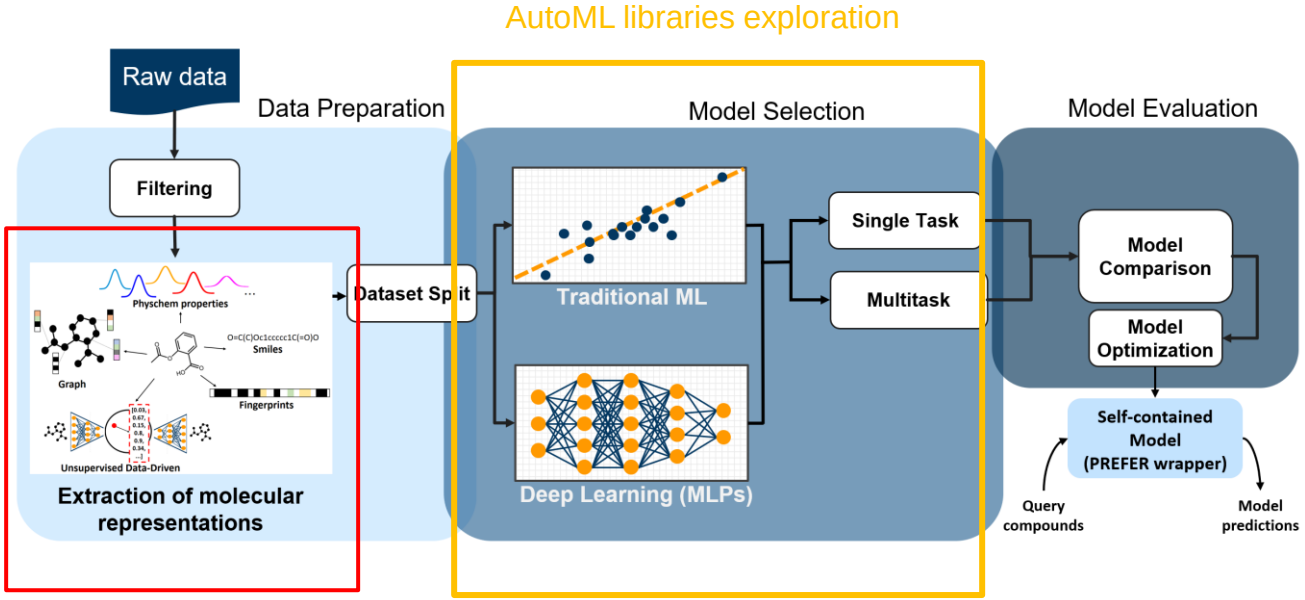
- *AutoML*: process of automating the tasks of applying machine learning to real-world problems. AutoML potentially includes every stage from beginning with a raw dataset to building a machine learning model ready for deployment.
- In the context of ADME and QSAR combining such automation with different molecular representations have been just partially explored:

Works in [11,12,13]	AMPL [14]	OpenChem [15]	Transcreen [16]
• Only traditional molecular representations • Only few classical ML models • Only one type of data split (random or cluster)	• Based on DeepChem library • limitation in modular design	• Pytorch based DL toolkit • No integration of traditional ML technique	• Transfer Learning setup based on GCNN; • No integration of traditional ML technique • No integration of traditional molecular representations

[11] Kausar S, et al. An automated framework for QSAR model building. Journal of cheminformatics. 2018 Dec;10(1):1-23.
[12] Obrezanova O, et al. Automatic QSAR modeling of ADME properties: blood–brain barrier penetration and aqueous solubility. Journal of computer-aided molecular design..
[13] Dixon SL, et al. AutoQSAR: an automated machine learning tool for best-practice quantitative structure–activity relationship modeling. Future medicinal chemistry. 2016
[14] Minnich AJ, et al. AMPL: a data-driven modeling pipeline for drug discovery. Journal of chemical information and modeling. 2020
[15] Korshunova M, et al. OpenChem: A deep learning toolkit for computational chemistry and drug design. Journal of Chemical Information and Modeling. 2021
[16] Salem, Milad, et al. "Transcreen: transfer learning on graph-based anti-cancer virtual screening model." Big Data and Cognitive Computing 4.3 (2020)

PREFER overview

- Goals:
  - build a broad models benchmarking framework for molecular properties
  - Deliver self-contained best model
- How:
  - Adapt well established AutoML libraries to handle molecular data
  - Wrap traditional and data-driven molecular representations

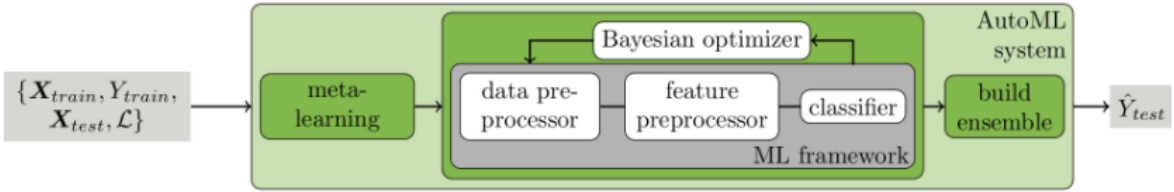


Data-Driven molecular representations are integrated as github submodules

	Hyperopt-Sklearn	TPOT	Auto-Sklearn	Azure AutoML	Optuna
Documentation			Very well done		
Installation	One can have problems with the version of hyperopt (pip install hyperopt==0.2.5)				
Usage	Example in the documentations show errors		Can take a while if you do not set time limits per run	Need some practice	
Model Customization	Need to implement interface for NN-based models		They reported easy example to follow + derived library auto-pythorch		Eventually provided by PREFER
Hyperparams Customization			They reported easy example to follow		
Optimization metric customization		Coupled with the main class	They reported easy example to follow		
Scikit-learn integration					
Dependences		NumPy – SciPy - scikit-learn – DEAP - update_checker - tqdm - pandas – joblib - xgboost		Strong dependency to Azure	
Multitasking + Sparsity	No multitasking supported	No multitasking	Handle multitasking without sparcicity	Only multi-class, no multi-task	Eventually provided by PREFER
Open Source					
Code maintenance		(last issue/PR some days ago)	(last issue/PR some days ago)		
GPU usage	uses scikit-optimize and scikit-learn under the hood (see here)	released version 0.11.6 is now accelerated with RAPIDS cuML and DMLC XGBoost	Not supported		Optuna doesn't have computations that can be speeded up by GPU

Auto-Sklearn: an overview

- Based on Scikit-Learn
- Optimization techniques implemented :
  - Bayesian Optimizer
- Meta-learning* step to start Bayesian optimization procedure
- Automated ensemble procedure step
  - This can help with the integration of the *uncertainty estimation*



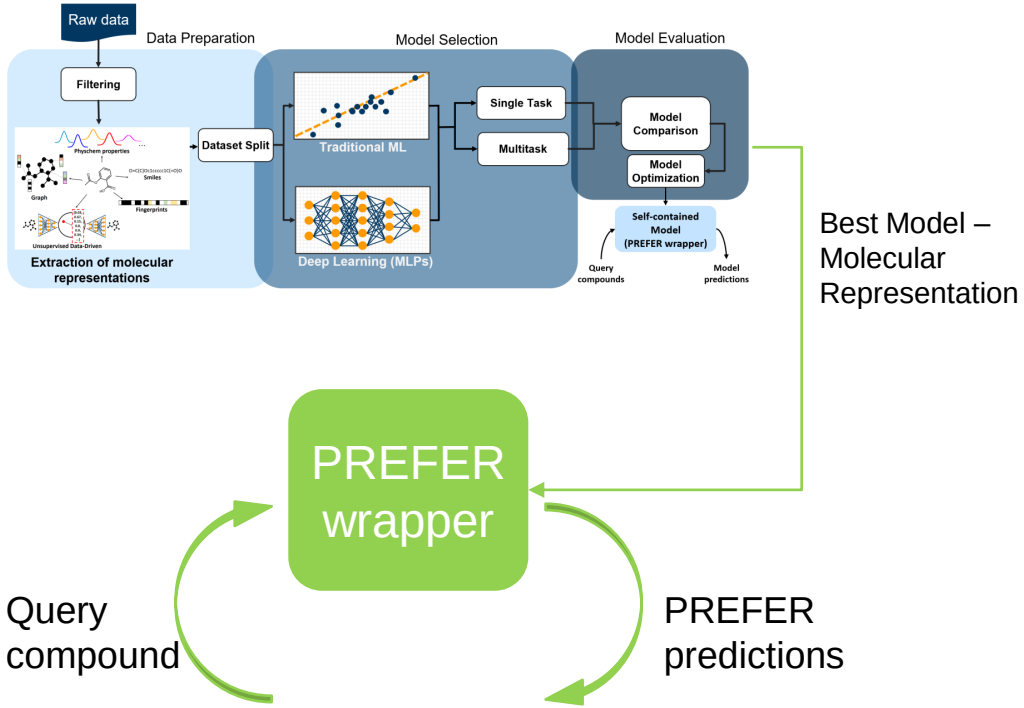

```
import autosklearn.classification
cls = autosklearn.classification.AutoSklearnClassifier()
cls.fit(X_train, y_train)
predictions = cls.predict(X_test)
```

*Given a large number of datasets, we collect both performance data and a set of meta-features, i.e., characteristics of the dataset that can be computed efficiently and that help to determine which algorithm to use on a new dataset.

[17] Feurer, Matthias, et al. "Auto-sklearn 2.0: The next generation." arXiv preprint arXiv:2007.04074 24 (2020).
[18] Feurer, Matthias, Jost Springenberg, and Frank Hutter. "Initializing bayesian hyperparameter optimization via meta-learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 29. No. 1. 2015.

PREFER overview

- At the end of the PREFER pipeline: a PREFER model wrapper will be created
- Inference Time: given a query compound given as SMILES input to the PREFER wrapper, it will
  - Featurize the SMILES according to the molecular representation used during training
  - Scale the feature vector if needed
  - Predict the corresponding label



PREFER details

Molecular Representations

- Morgan Fingerprints
- 2D Descriptors
- Continuous and Data Driven Descriptors (CDDD)
- Representation based on the MoLeR model

ML algorithms types

- Regression Single Task
- Binary Classification (best decision threshold with GHOST [19])
- Regression Multitask
- Binary Classification Multitask (best decision threshold with GHOST [19])

ML Algorithms

- Adaboost,
- Decision tree,
- Extra trees,
- Gaussian process,
- Gradient boosting,
- KNN,
- Linear svr,
- Mlp,
- RF,
- SGD

Evaluation metrics

- Regression: RMSE, R2, RMSE normalized, Mean test error, Max test error, Min test error, error 25th percentile, error 50th percentile, error 75th percentile
- Classification: Balanced Accuracy, F1 score, Precision, Recall, AUC, kappa score

[19] Esposito C, Landrum GA, Schneider N, Stiefl N, Riniker S. GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. Journal of Chemical Information and Modeling 2021

Experiments

Data used for the experiments

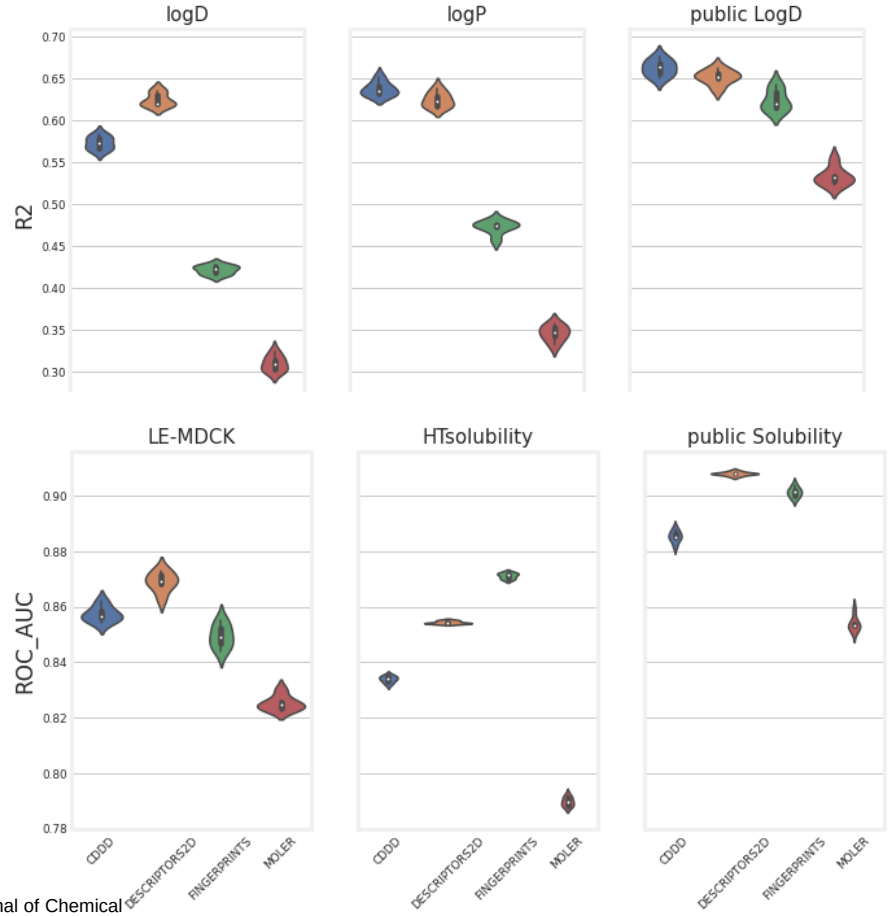
- All datasets comprise assay readouts important in early drug discovery
- According to the dynamic range of the data and the fraction of censored data, each task was modeled as classification/regression model
- Time-split and random split were used for the evaluation of internal and public data respectively

Name	Measure of	Class Balance	Problem Type	Number of observations	Source
LE-MDCK	Permeability	Permeable = 59%, Impermeable = 41%	Classification	14K	Novartis
HT Solubility	High-throughput solubility	Soluble = 66%, Insoluble = 34%	Classification	200K	Novartis
Public Solubility	High-throughput solubility	Soluble = 70%, Insoluble = 30%	Classification	56K	PubChem [37]
Direct logD	Lipophilicity of a compound at pH 7.4	-	Regression	20K	Novartis
Direct logP	Lipophilicity of a compound at a pH where the compound is uncharged	-	Regression	13K	Novartis
Public logD	Lipophilicity of a compound at pH 7.4	-	Regression	4K	ChEMBL [38]

[20] Lanini, Jessica, et al. "PREFER: A New Predictive Modeling Framework for Molecular Discovery." Journal of Chemical Information and Modeling (2023).

PREFER out-of-the box performances

- Without any manual intervention in the entire ML pipeline a reasonable model (ROC-AUC > 0.86 or R2 > 0.6) can be created for each task (ML model + molecular representation), using PREFER with its defaults.
- Performance variability on the test set given different molecular representations

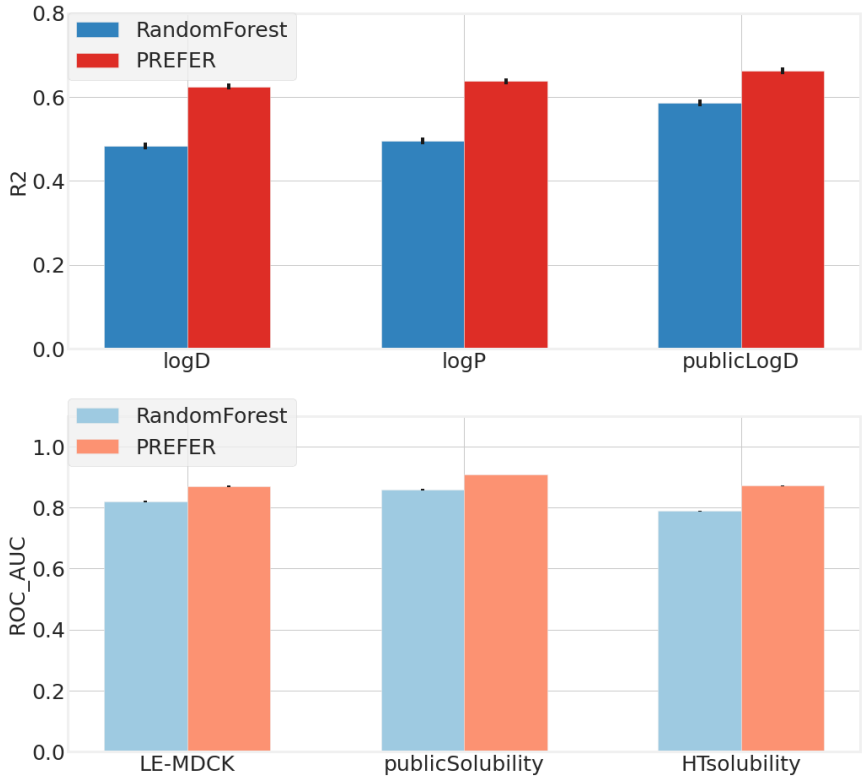


[20] Lanini, Jessica, et al. "PREFER: A New Predictive Modeling Framework for Molecular Discovery." Journal of Chemical Information and Modeling (2023).

PREFER comparison with baseline

- Random Forest model (100 estimators, max depth of 20) with Morgan fingerprints as the baseline. Random Forest model (100 estimators, max depth of 20) and Morgan fingerprints as the baseline

- Overall, the best PREFER model outperforms the corresponding baseline, particularly in the case of regression tasks where the average improvement is more than 10%.

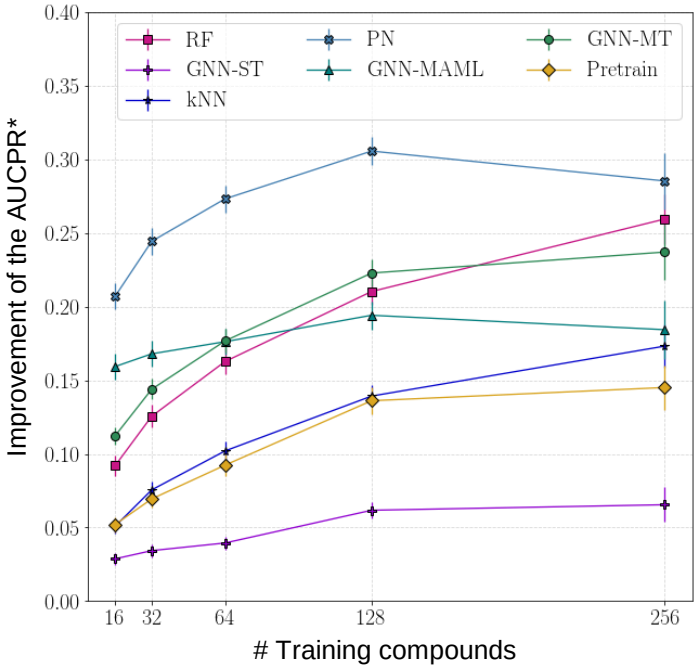


[20] Lanini, Jessica, et al. "PREFER: A New Predictive Modeling Framework for Molecular Discovery." Journal of Chemical Information and Modeling (2023).

Dealing with small data: Project-specific data modeling

- **Quantitative structure-activity relationship (QSAR)** models for project-data/assay face a "**small data issue**"
- **Low performance & narrow applicability domain**
- Several methods (like few-shot learning) can be used to address this challenge
- **FS-Mol benchmarking suite**[21] to enable comparison of different models in few-shot learning tasks

FS-Mol results



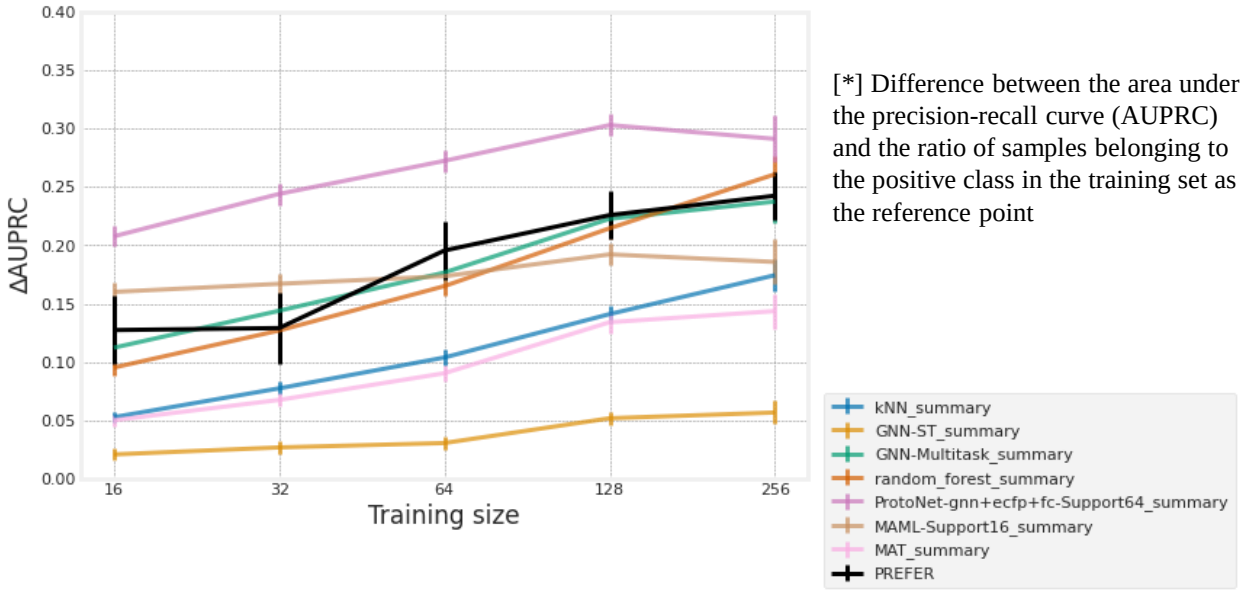
*AUCPR = Area under the Precision-Recall curve

[21] Stanley, M., Bronskill, J. F., Maziarz, K., Misztela, H., Lanini, J., Segler, M., Schneider, N. & Brockschmidt, M. (2021). FS-Mol: A few-shot learning dataset of molecules. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Microsoft NOVARTIS | Reimagining Medicine

PREFER for small data

- PREFER has been applied to FS-Mol few-shot learning benchmark
- ΔAUPRC is used as metric [*]
- Increasing training set size, PREFER performance improves to be almost comparable to the best model in [21]



[*] Difference between the area under the precision-recall curve (AUPRC) and the ratio of samples belonging to the positive class in the training set as the reference point

[21] Stanley M, et al. Fs-mol: A few-shot learning dataset of molecules. InThirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track 2021 Aug 25.

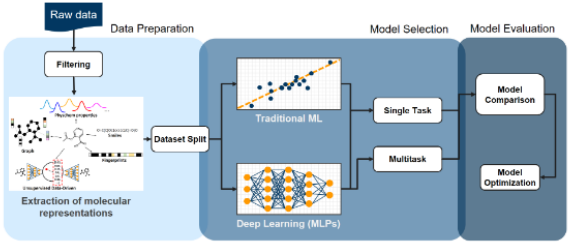
PREFER GitHub repository

prefer	add(submodules): cddd and moler	5 months ago
small_data_experiments	typo	2 months ago
.gitmodules	add(submodules): cddd and moler	5 months ago
LICENSE	Initial commit	5 months ago
README.md	add(files): main code	5 months ago
Run_PREFER.ipynb	add(files): main code	5 months ago
__init__.py	add(files): main code	5 months ago
api_version.txt	add(files): main code	5 months ago
cddd-environment-light.yml	add(files): main code	5 months ago
cddd-environment.yml	add(files): main code	5 months ago
compute_model_based_representatio...	add(files): main code	5 months ago
moler-environment-light.yml	add(files): main code	5 months ago
moler-environment.yml	add(files): main code	5 months ago
prefer-environment.yml	update(prefer-environment.yml): added two dependences	5 months ago
pyproject.toml	add(files): main code	5 months ago
run_prefer_automation.py	add(files): main code	5 months ago
setup.py	add(files): main code	5 months ago

README.md

Benchmarking and Property Prediction Framework (PREFER)

The PREFER framework automatizes the evaluation of different combinations of molecular representations and machine learning models for predicting molecular properties. It covers different molecular representation from classical, e.g. Fingerprints and 2D Descriptors, to data-driven representations, e.g. Continuous and Data Driven representations (CDDD) [1] or MoLeR[2]. PREFER uses AutoSklearn [3] to implement the ML model selection and the hyperparameter tuning.



3 stars
4 watching
1 fork
Report repository

Releases
No releases published
Create a new release

Packages
No packages published
Publish your first package

Languages
Python 80.4% Jupyter Notebook 19.6%

Suggested Workflows
Based on your tech stack

Actions Importer — Set up
Automatically convert CI/CD files to YAML for GitHub Actions.

Python package — Configure
Create and test a Python package on multiple Python versions.

SLSA Generic generator — Configure
Generate SLSA3 provenance for your existing release workflows

More workflows Dismiss suggestions

https://github.com/rdkit/PREFER

NOVARTIS | Reimagining Medicine

Thank you

NOVARTIS | Reimagining Medicine