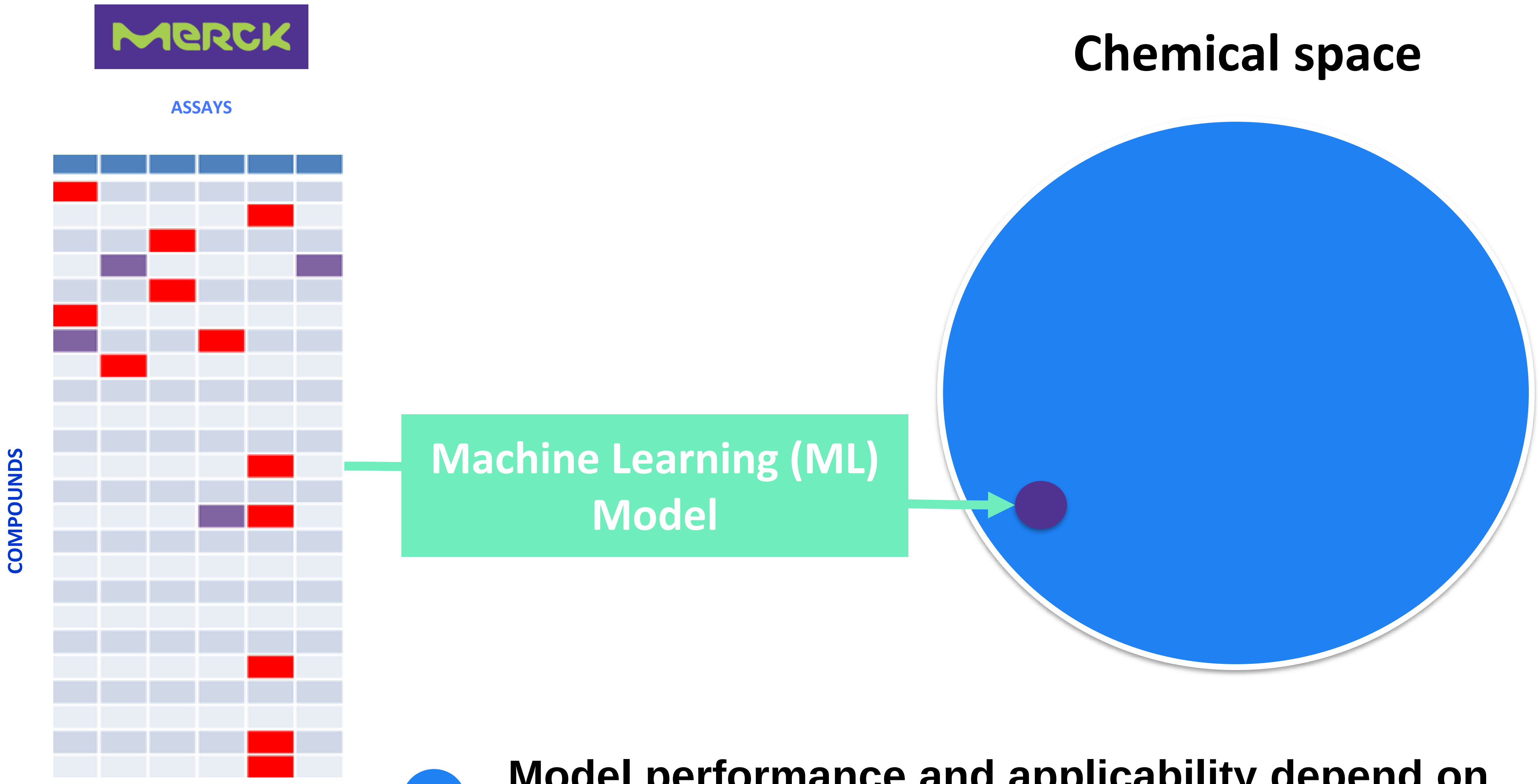
MELLODDY-TUNER:
Data Standardization Framework for Federated Machine Learning

RDKit UGM 2020
Lukas Friedrich (Merck KGaA, Darmstadt)

MELLODDY

MACHINE LEARNING LEDGER ORCHESTRATION FOR DRUG DISCOVERY

PREDICTIVE MODELING IN DRUG DISCOVERY



Model performance and applicability depend on amount of high quality data

MACHINE LEARNING LEDGER
ORCHESTRATION FOR DRUG DISCOVERY

MELLODDY

powered by aws

PHARMA PARTNERS

AMGEN

astellas

AstraZeneca

BAYER

Boehringer Ingelheim

gsk

janssen
PHARMACEUTICAL COMPANIES OF Johnson&Johnson

MERCK

NOVARTIS

SERVIER

PUBLIC PARTNERS

MŰEGYETEM 1782

IKTOS

KU LEUVEN

Kubermatic

NVIDIA.

OWKIN

Substra Foundation

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement N° 831472. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation programme and EFPIA

imi innovative medicines initiative

efpia

THE MELLODDY OBJECTIVES

On average, bringing one drug to market costs €1.9 billion and 13 years[1].

The virtualization of parts of drug discovery by machine learning (ML) is a promising approach to improve efficiencies.

MELLODDY aims to show predictive benefits of modelling across tasks, data types and partners at the largest achievable scale.

[1] DiMasi JA et al., 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. Journal of Health Economics 47, 20-33.

In three yearly runs, the increasingly sophisticated platform will learn from:

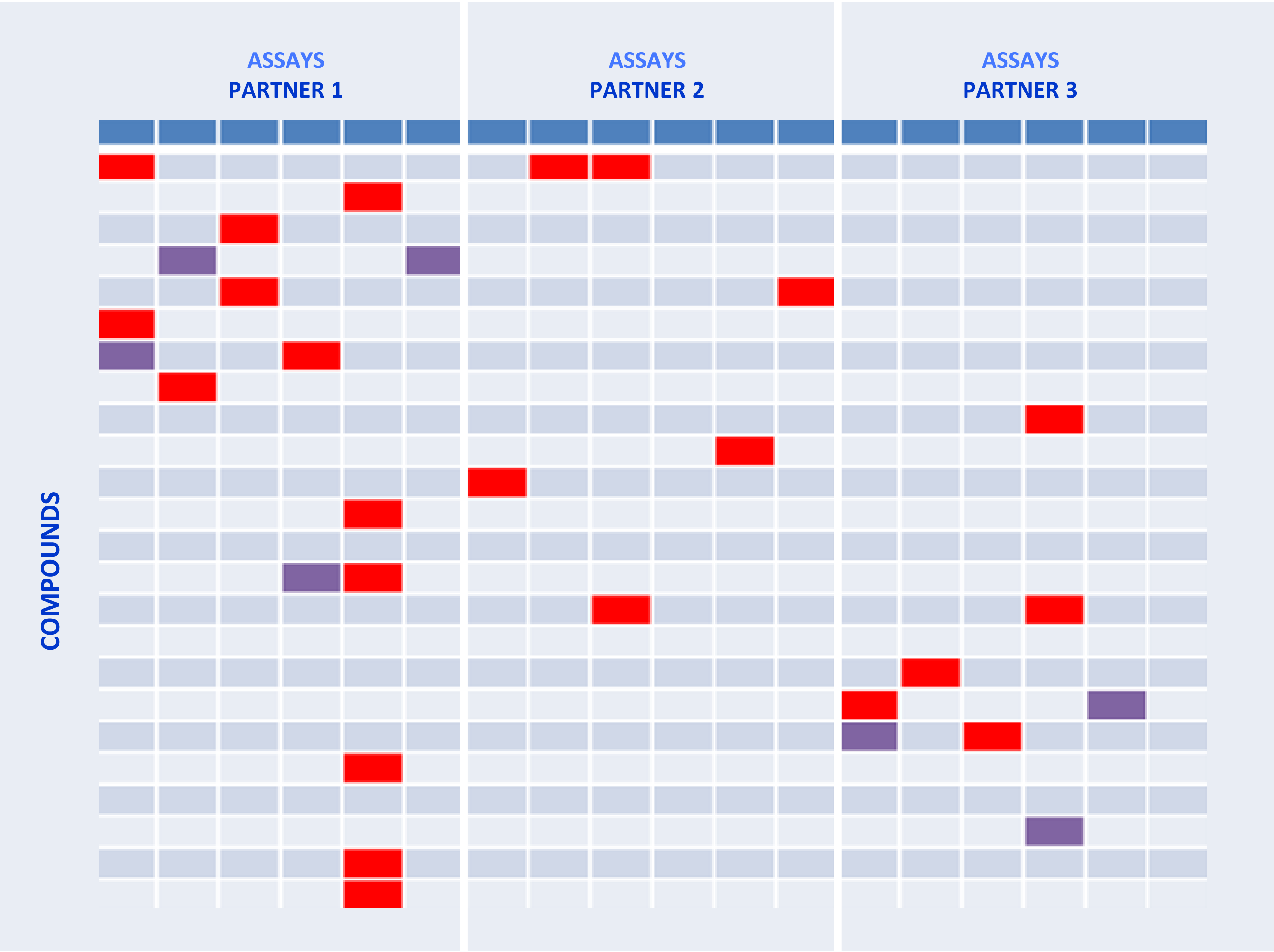
- **> 10 million annotated small molecules**
- **> 1 billion assay biological activity labels**
- **Multiple high-complexity phenotypes at high throughput**
- **Multiple high-complexity phenotypes at high throughput**

Privacy preservation of data and federated models is paramount.

MULTI-TASK LEARNING
ACROSS PHARMA PARTNERS

Compound and activity data and assay-specific models remain under their owner’s control

Multi-task approach across partners to improve predictive performance and applicability



AMGEN

astellas

AstraZeneca

BAYER

Boehringer Ingelheim

gsk

janssen
PHARMACEUTICAL COMPANIES OF Johnson&Johnson

MERCK

NOVARTIS

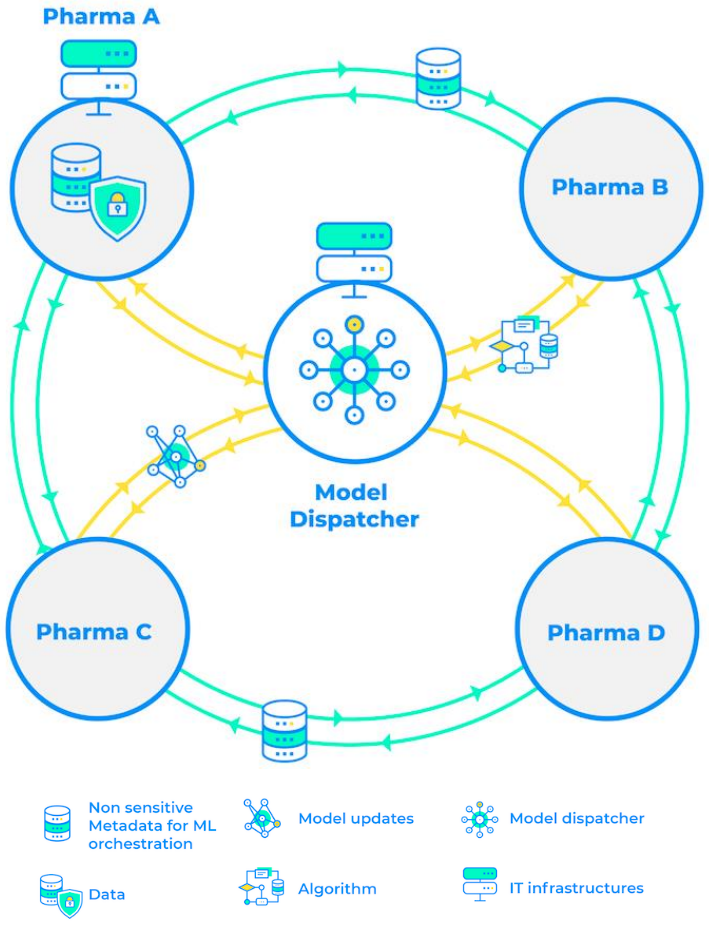
SERVIER

COMBINED PRIVACY-PRESERVING
FEDERATED MACHINE LEARNING PLATFORM

Sensitive data and assay-specific models remain locked on each pharma's server

Lower level model components are securely exchanged and trained over the network

Complex but transparent pre-agreed access arrangements are strictly enforced



IKTOS

KU LEUVEN

Kubermatic

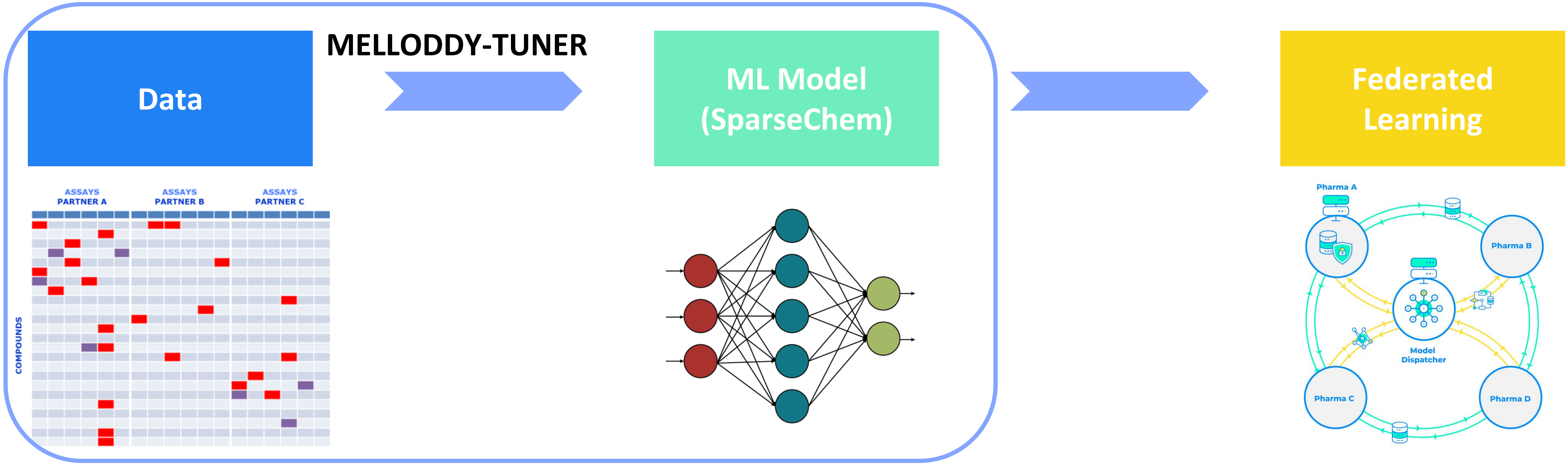
MŰEGYETEM 1782

NVIDIA.

OWKIN

Substra Foundation

MELLODDY-TUNER:
DATA STANDARDIZATION FOR FEDERATED LEARNING



Data

MELLODDY-TUNER

ML Model (SparseChem)

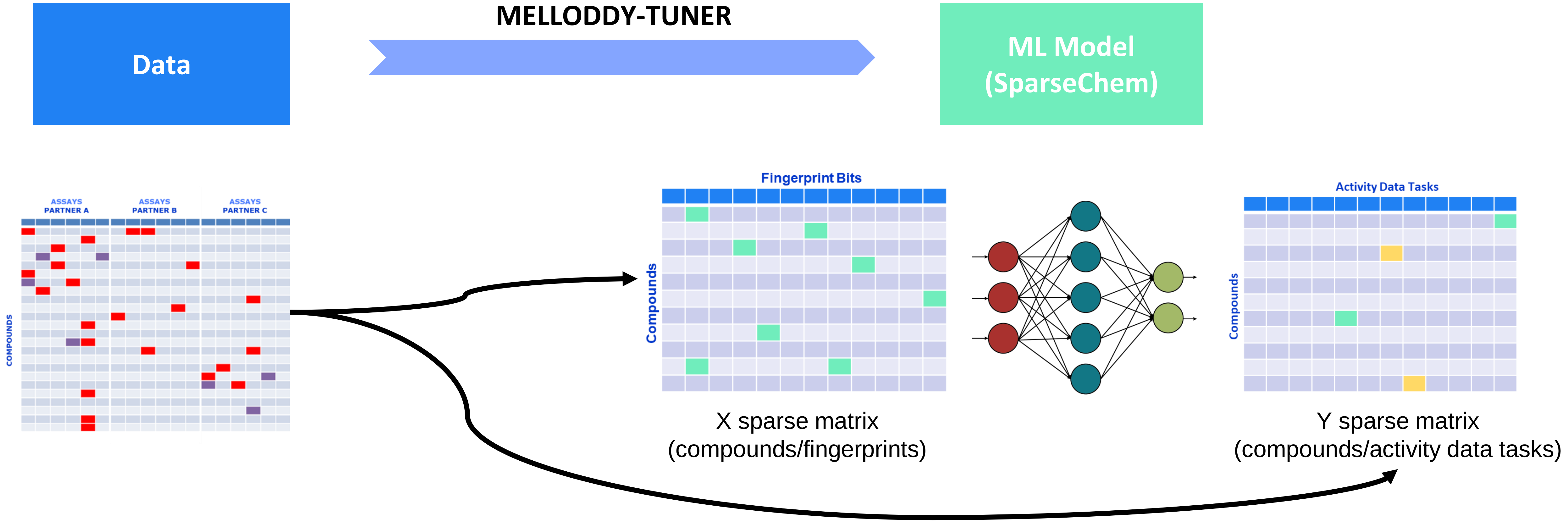
Federated Learning

Standardize data locally at multiple partners

Provide suitable input files for machine learning algorithm

Guarantee uniform processing within consortium while preserving privacy of partner’s data

MELLODDY-TUNER:
TECHNICAL OBJECTIVE



X sparse matrix
(compounds/fingerprints)

Y sparse matrix
(compounds/activity data tasks)

Standardize structures & activity data to create sparse matrices compatible with SparseChem

MELLODDY-TUNER:
DATA STANDARDIZATION FOR FEDERATED LEARNING

Standardize smiles

- Standardize structures:

  charge_parent

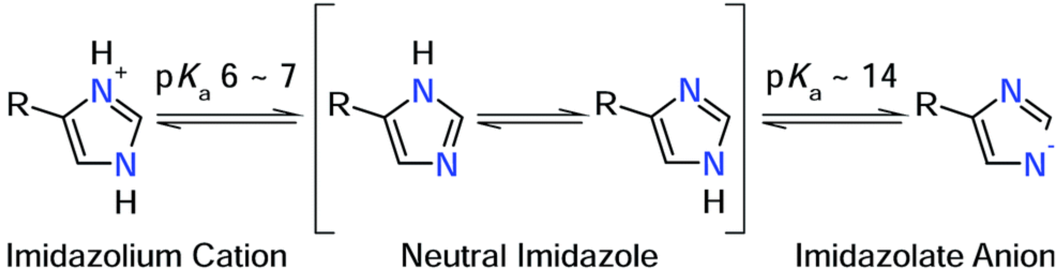
  isotope_parent

  stereo_parent

  tautomer_parent

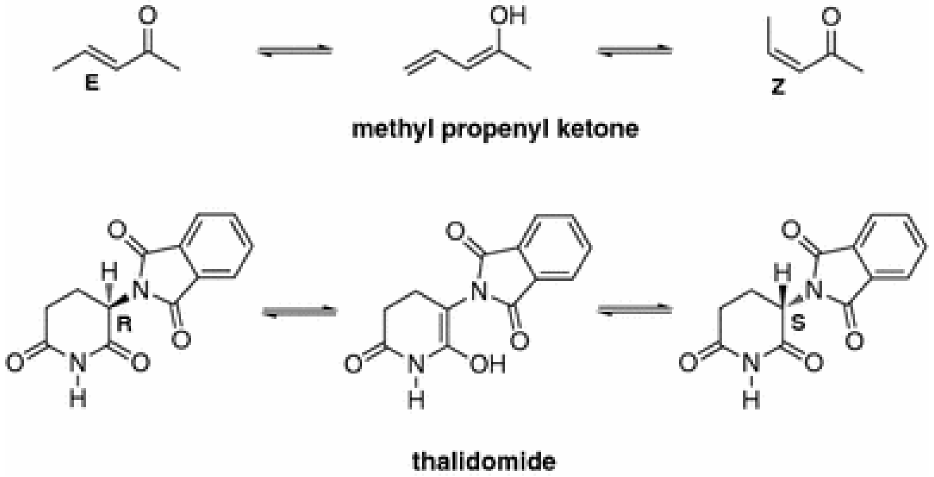
STRUCTURE STANDARDIZATION

One structure may not be enough



Horch et al., *RSC Adv.*, **4**, 54091-54095 (2014)

Tautomerization can change stereochemistry



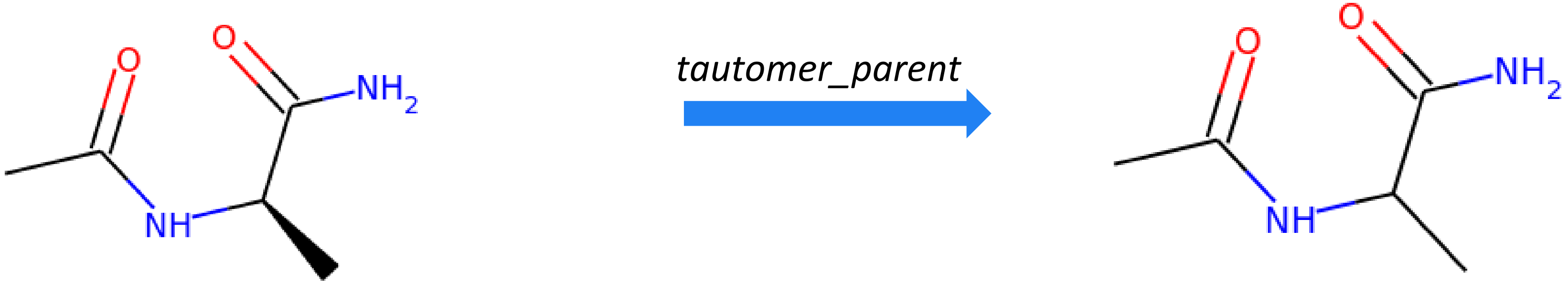
Sitzmann, M. et al., *J Comput Aided Mol Des* **24,** 521–551 (2010)

Objective: Make most consistent choice for standardization among several partners

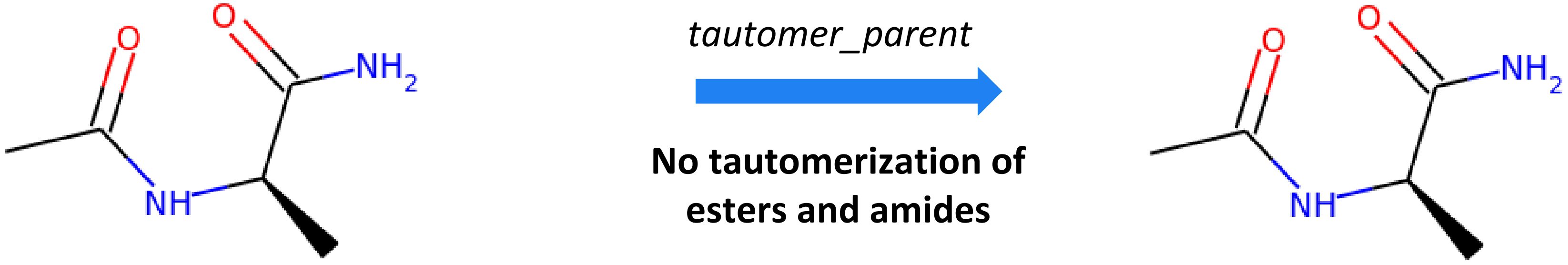
STRUCTURE STANDARDIZATION
TAUTOMERIZATION

RDKit2019.03

MolVS



Updated “TautomerTransform”



No tautomerization of esters and amides

```
TautomerTransform('1,3 (thio)keto/enol f', '[CX4!H0]-[C;!$([C]([CH1])(=[O,S,Se,Te;X1])-[N,O])]=[O,S,Se,Te;X1]')
```

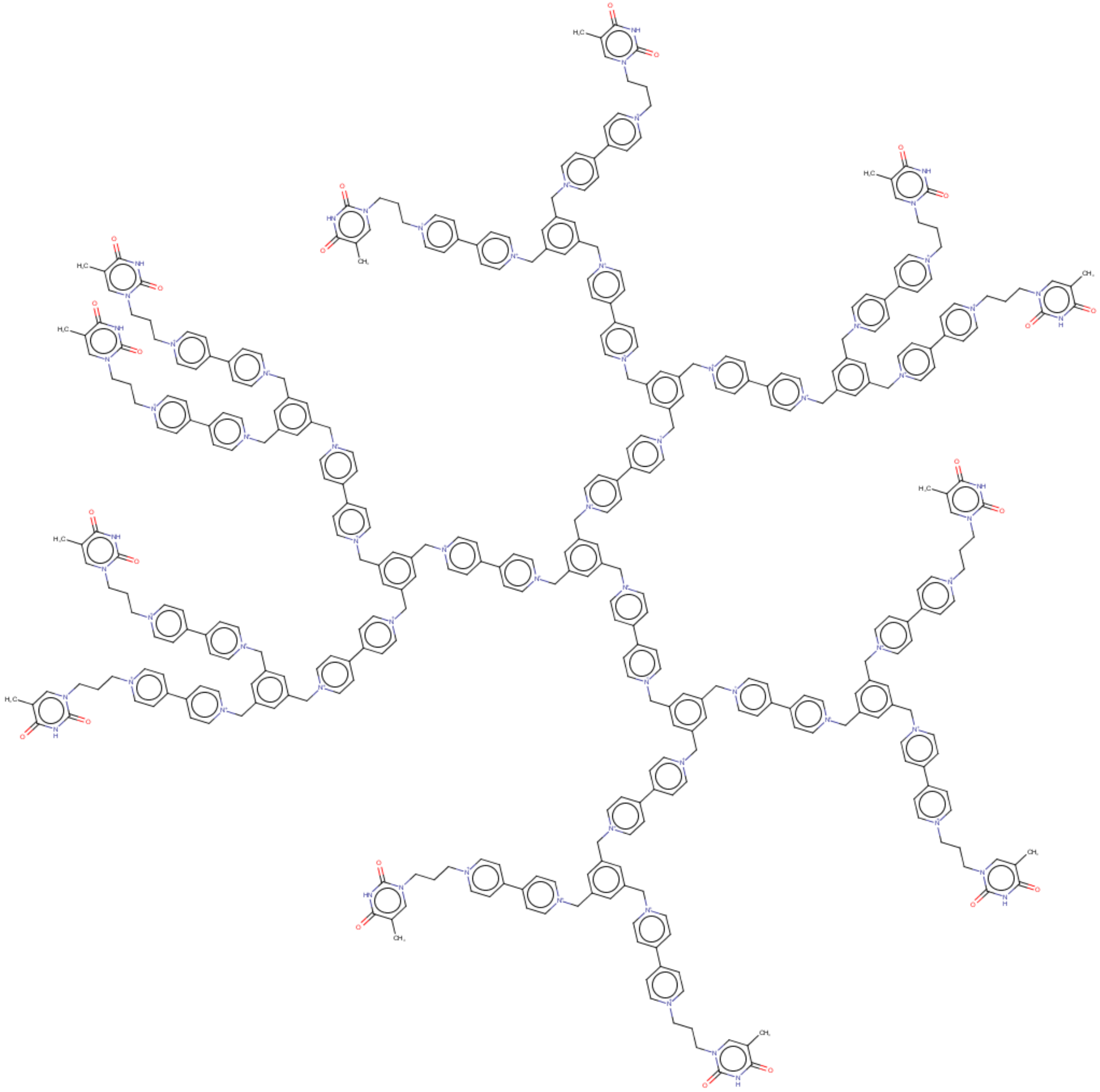
Moving Hydrogen from first atom to last atom

STRUCTURE STANDARDIZATION

Standardization can take time:

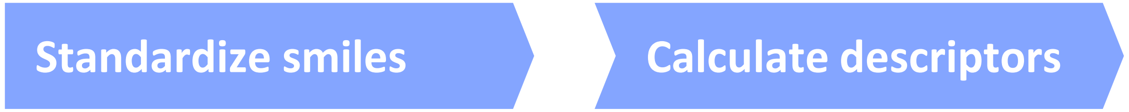
1. Enumerate all possible tautomers
2. Score all enumerated tautomers
3. Return canonical tautomer

Limit molecule size
(max. number heavy atoms)
Limit number of enumerated tautomers



Molecule from ChEMBL25
Standardization time: ~30 min

MELLODDY-TUNER:
DATA STANDARDIZATION FOR FEDERATED LEARNING



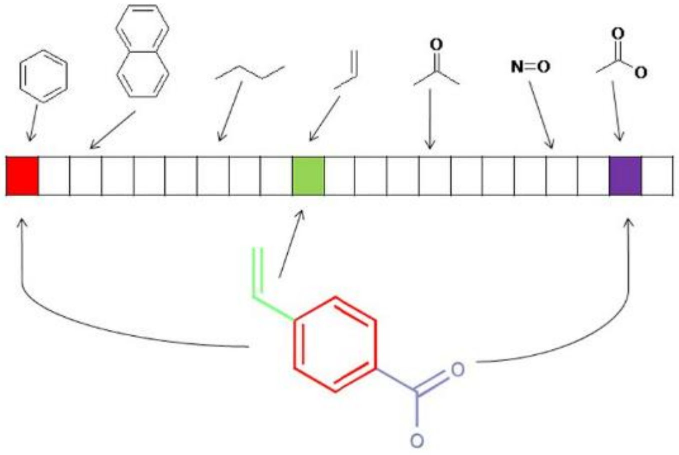
Standardize smiles

Calculate descriptors

- Calculate fingerprint
- **Cluster fingerprints into folds using locality-sensitive hashing (LSH)**

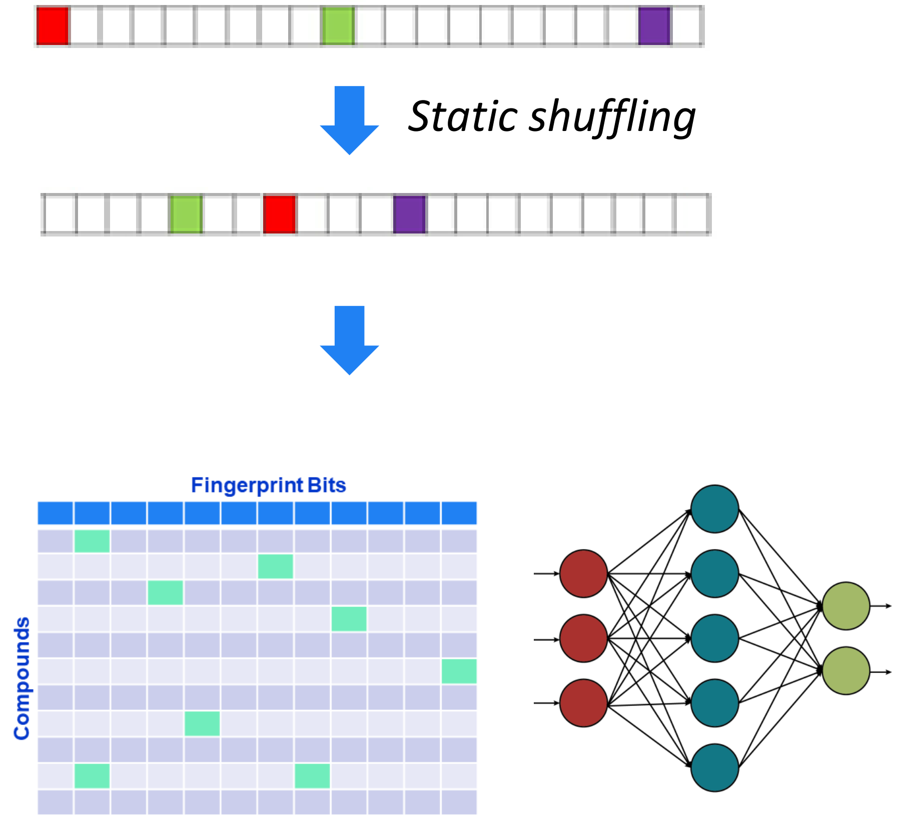
FINGERPRINT

1. Representation of molecules as bit vectors (Morgan Fingerprint with certain length):



➔ **Fingerprint can be „reverse-engineered“**
(Tuan Le‘ et al., ChemRxiv (2020))

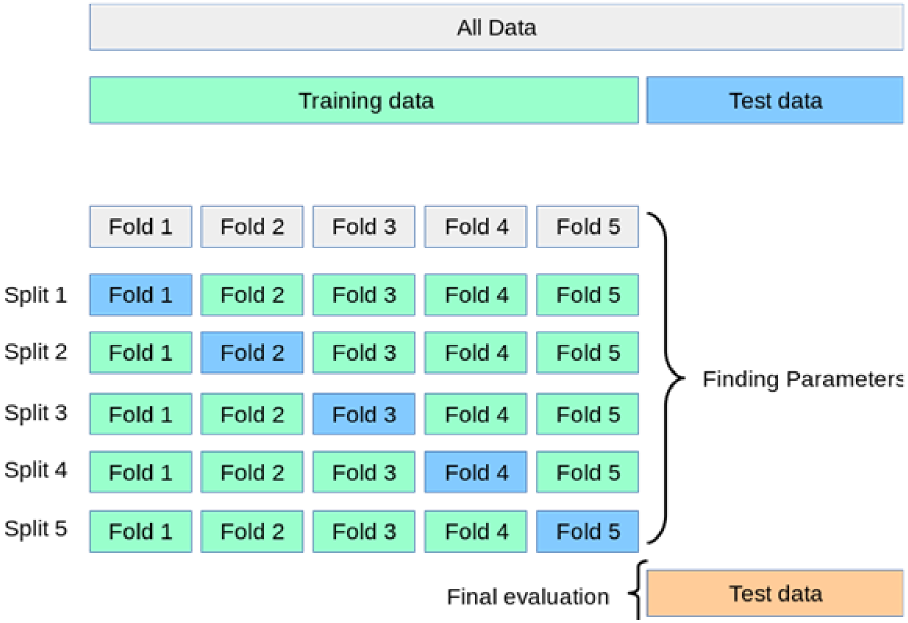
2. Static shuffling of bit positions using secret key



Molecular fingerprints are part of X matrix for SparseChem

FINGERPRINT
TRAINING, VALIDATION AND TEST SETS

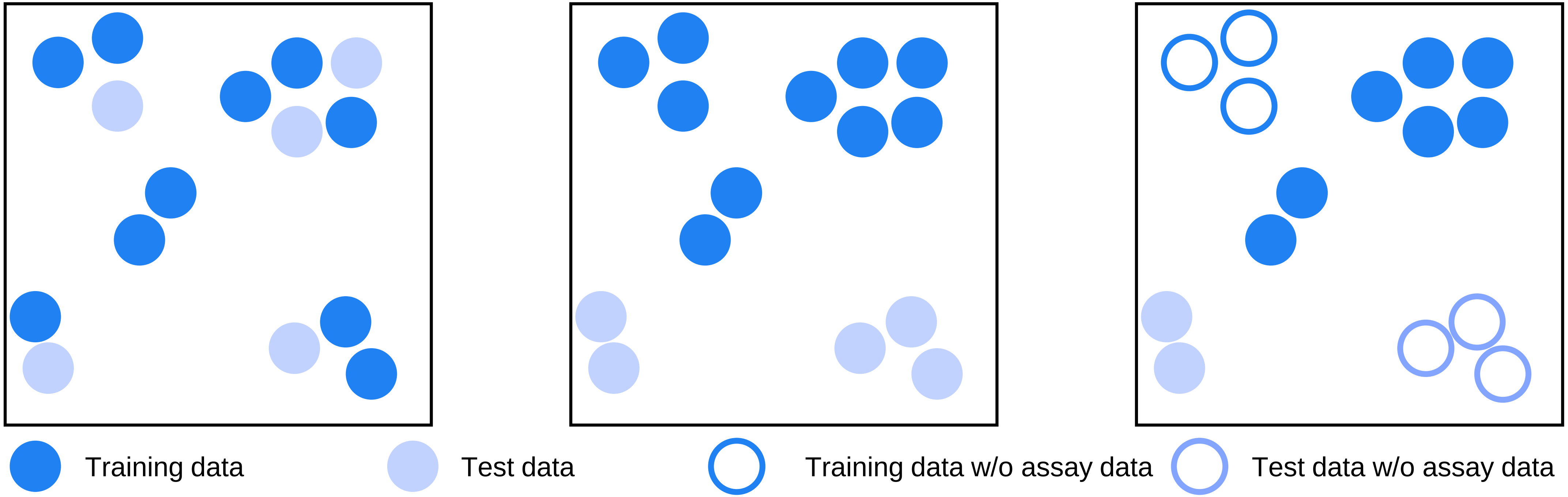
Evaluation of ML model performance requires data split into training, validation & test sets



Scikit-learn.org/stable/modules/cross_validation.html

How can we consistently assign compounds to folds across multiple partners?
How to guarantee that identical compounds from different partners land in the same fold?

TRAIN/TEST SPLIT:
RANDOM VS CLUSTER BASED SPLIT



Random Split:
Overly optimistic
performance assessment

Cluster based split:
More realistic
performance assessment

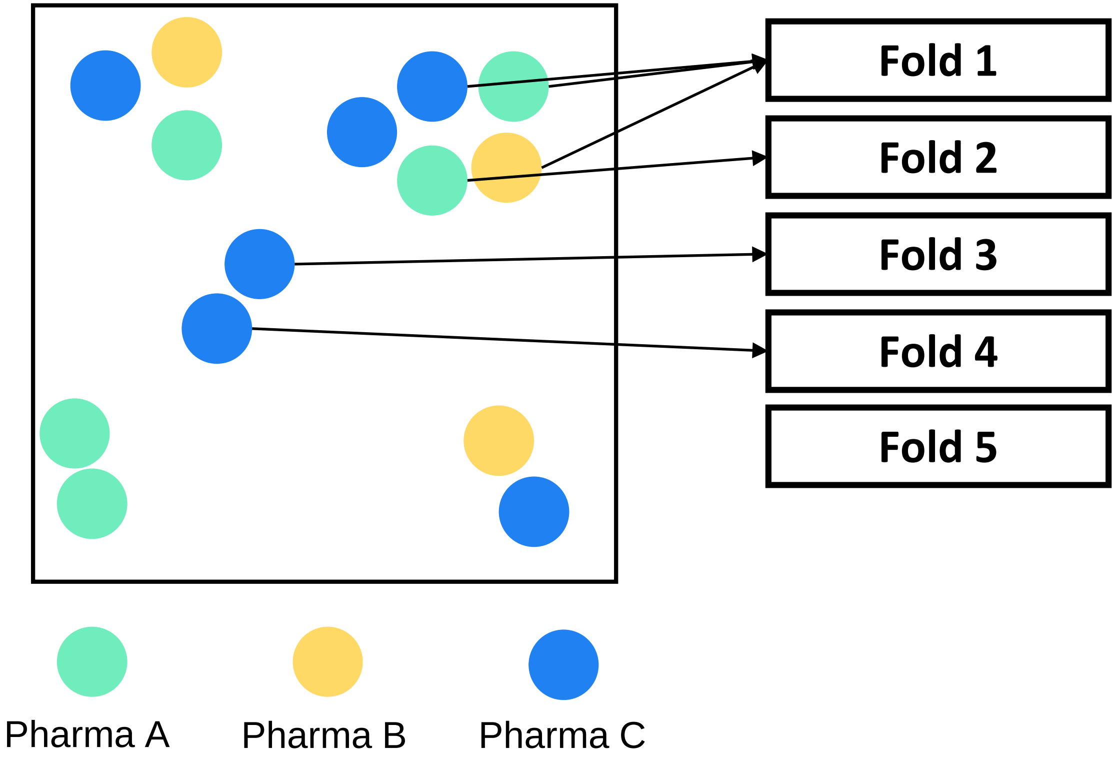
Cluster based split:
Uneven distribution of
assay data among clusters

perfect clustering not required, but privacy-preserving is necessary

TRAIN/VALIDATION/TEST FOLD:
LOCALITY-SENSITIVE HASHING

Proposed and implemented
by Jaak Simm (KU Leuven)

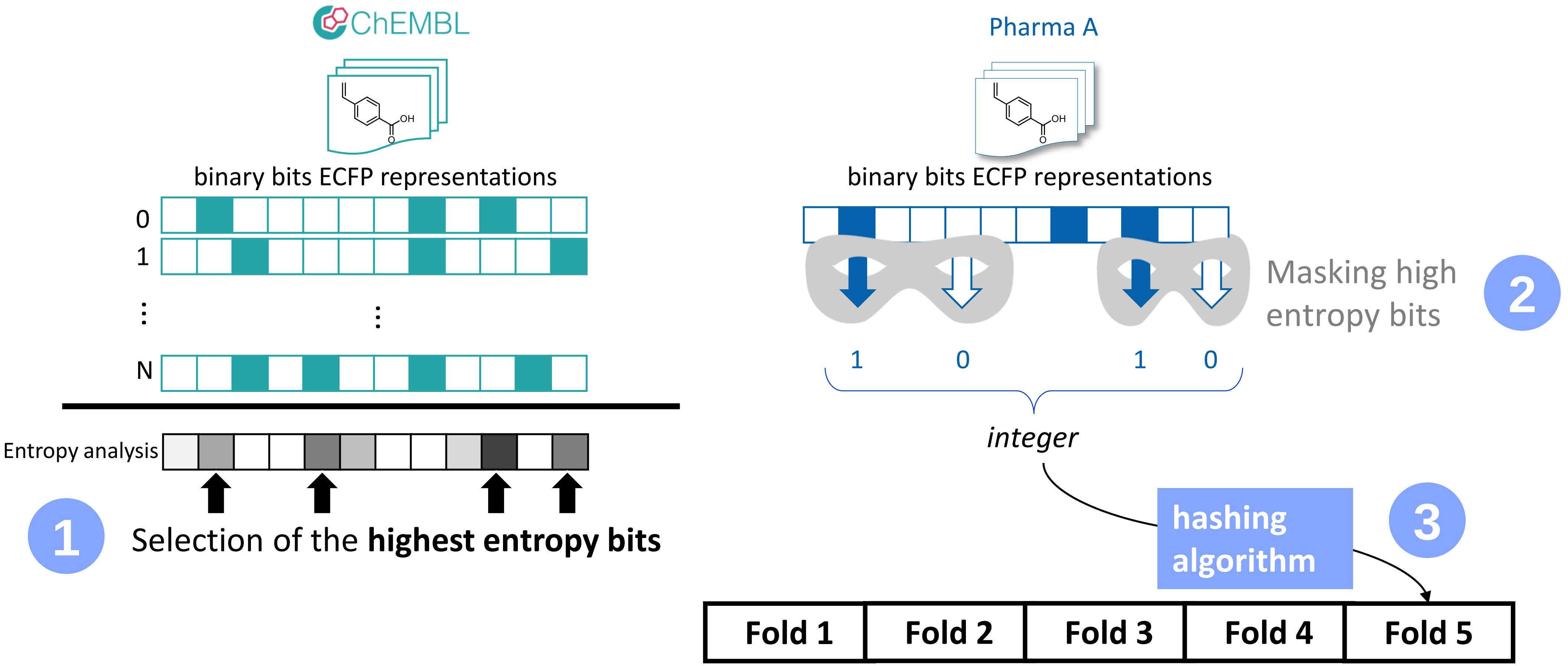
Locality-sensitive hashing:
Hashing similar items into same bucket(s) with high probability



How can we perform LSH clustering in the MELLODDY consortium?

TRAIN/VALIDATION/TEST FOLD:
LOCALITY-SENSITIVE HASHING

Proposed and implemented
by Jaak Simm (KU Leuven)


ChEMBL
binary bits ECFP representations
0
1
⋮
⋮
N
Entropy analysis
1
Pharma A
binary bits ECFP representations
2
1
0
1
0
integer
3

Selection of the **highest entropy bits**

Masking high entropy bits

hashing algorithm

Fold 1 | **Fold 2** | **Fold 3** | **Fold 4** | **Fold 5**

Figures from Noe Sturm (Novartis)

MELLODDY-TUNER:
DATA STANDARDIZATION FOR FEDERATED LEARNING

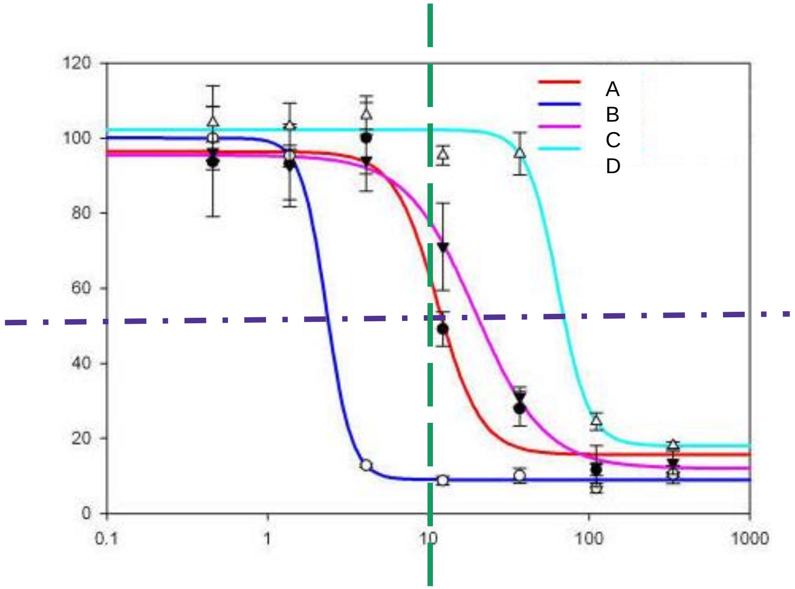
Standardize smiles

Calculate descriptors

Format activity data

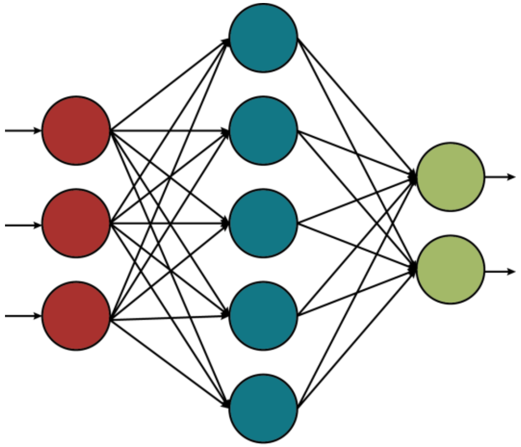
- Remove data from failed compounds
- Aggregate replicates
- Filter out tasks not fulfilling minimum number of actives/inactives

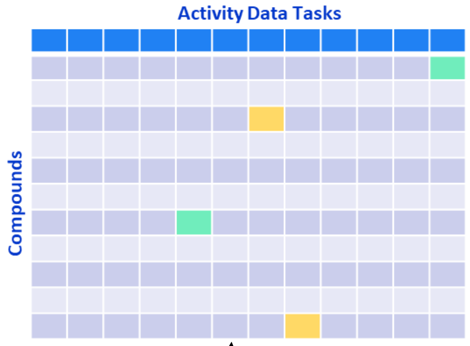
ACTIVITY DATA
BIT VECTOR REPRESENTATION



Assay	A	B	C	D
%Ctrl @ 10µM	70	8	82	100
Activity class	0	1	0	0

Activity fingerprints are part of Y matrix for SparseChem





ACTIVITY DATA
DATA AGGREGATION AND FILTERING

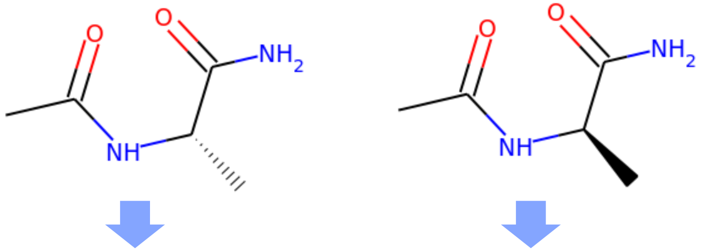
Remove data of „failed“ compounds

Aggregate data of replicates

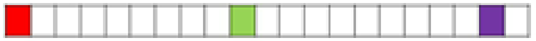
Filte out tasks not fulfilling criteria for SparseChem

„Failed“ compounds like

m1 = Chem.MolFromSmiles('CO(C)C')



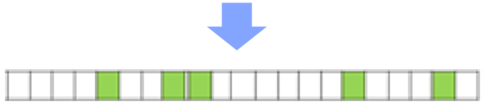
Same fingerprint for different compounds (e.g. stereochemistry not considered)



Assay	A	B	C	D
%Ctrl @ 10µM	70	8	82	100
Activity class	0	1	0	0

Assay	A	B	C	D
%Ctrl @ 10µM	90	23	15	100
Activity class	0	0	1	0

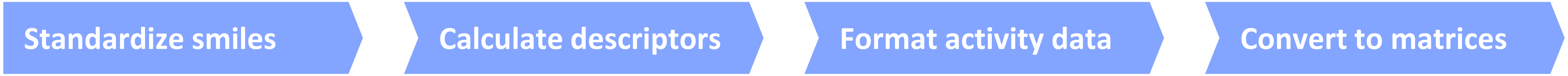
Majority voting approach:
„Majority“ class wins for tasks with multiple datapoints
„Minority“ class wins for tasks with draws



Guarantee sufficient amount of both classes in all folds

Analysis of ML performance is possible

MELLODDY-TUNER:
DATA STANDARDIZATION FOR FEDERATED LEARNING



Standardize smiles

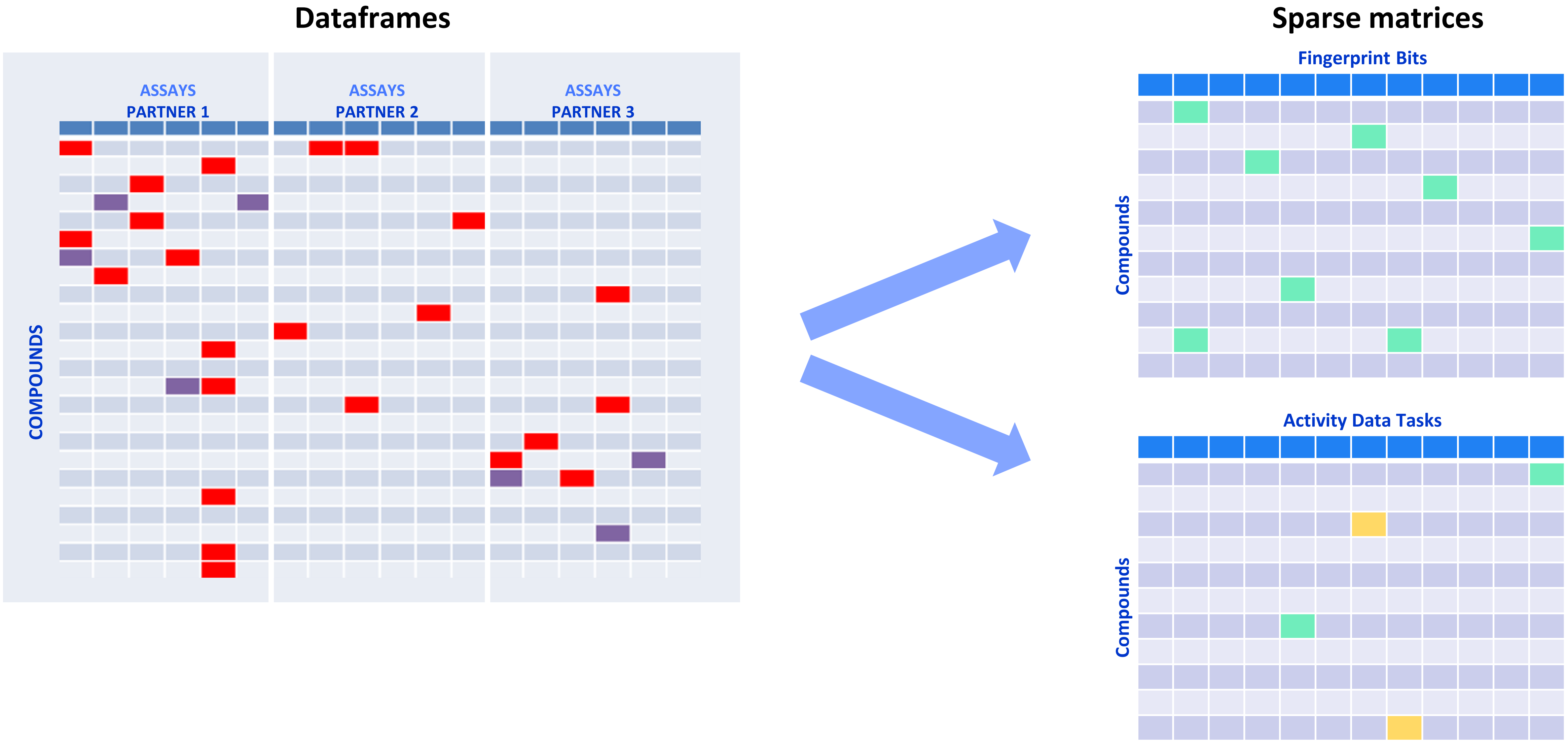
Calculate descriptors

Format activity data

Convert to matrices

- Create X sparse matrix (compound/fingerprint)
- Create Y sparse matrix (compound/activity data tasks)

SPARSE MATRICES



Dataframes

ASSAYS
PARTNER 1

ASSAYS
PARTNER 2

ASSAYS
PARTNER 3

COMPOUNDS

Sparse matrices

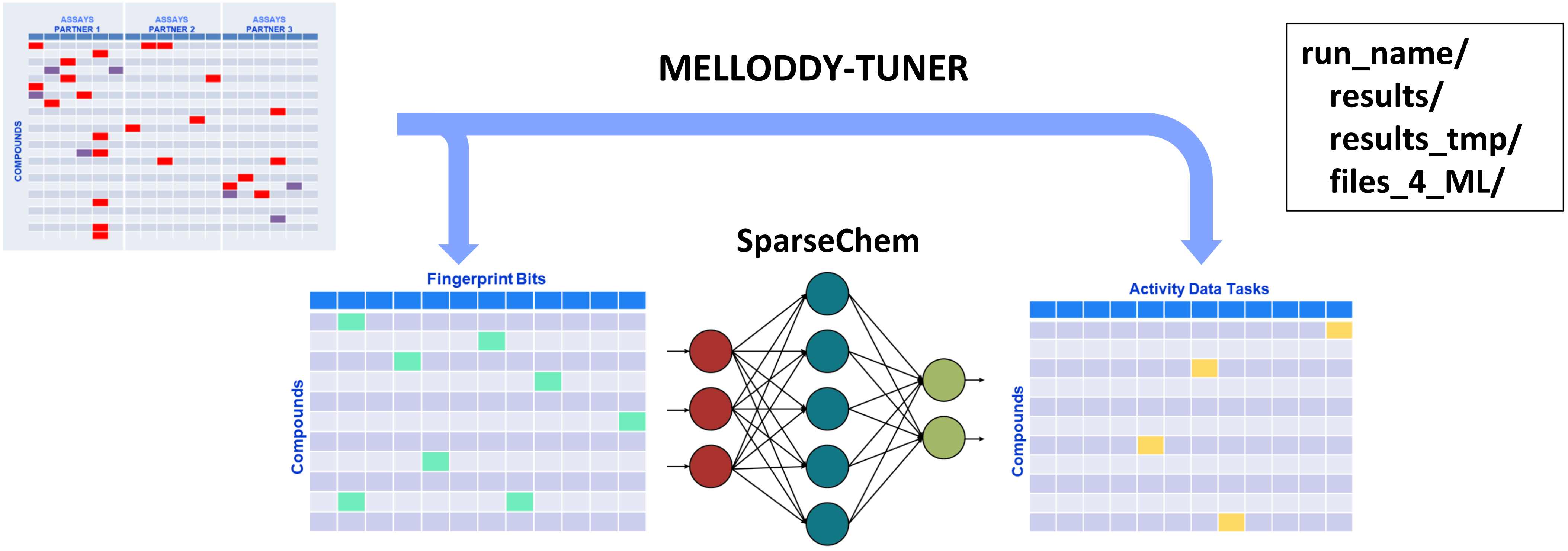
Fingerprint Bits

Compounds

Activity Data Tasks

Compounds

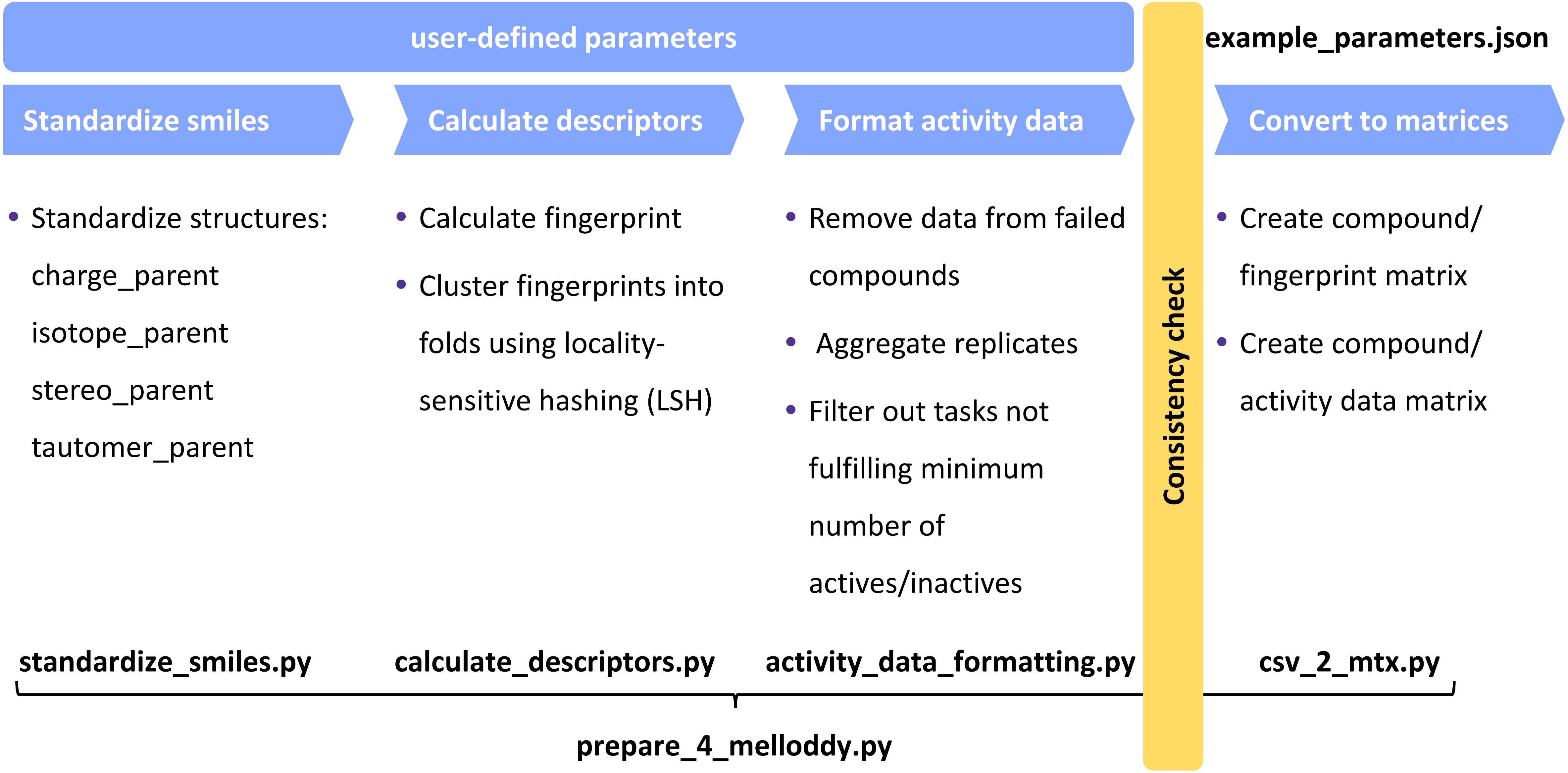
SPARSE MATRICES
MELLODDY-TUNER & SPARSECHEM



```
run_name/
  results/
  results_tmp/
  files_4_ML/
```

MELLODDY-TUNER provides dataframes of standardized data (results), mapping tables & excluded data (results_tmp) and SparseChem-compatible matrices (files_4_ML)

MELLODDY-TUNER
SUMMARY



user-defined parameters

example_parameters.json

Standardize smiles

- Standardize structures:
  charge_parent
  isotope_parent
  stereo_parent
  tautomer_parent

standardize_smiles.py

Calculate descriptors

- Calculate fingerprint
- Cluster fingerprints into folds using locality-sensitive hashing (LSH)

calculate_descriptors.py

Format activity data

- Remove data from failed compounds
- Aggregate replicates
- Filter out tasks not fulfilling minimum number of actives/inactives

activity_data_formatting.py

Consistency check

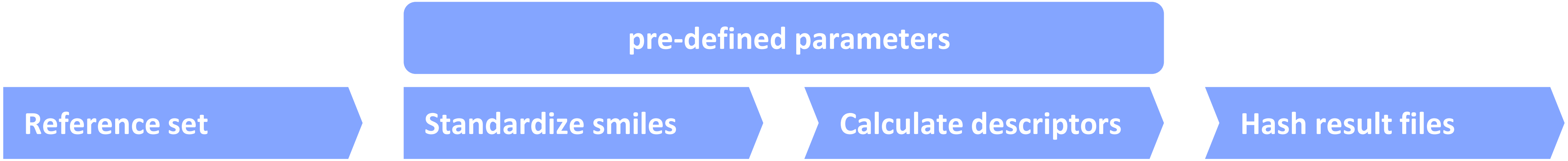
Convert to matrices

- Create compound/ fingerprint matrix
- Create compound/ activity data matrix

csv_2_mtx.py

prepare_4_melloddy.py

MELLODDY-TUNER
CONSISTENCY CHECK



Reference set

- 10 reference structures to be processed

Standardize smiles

- Standardize structures:
  charge_parent
  isotope_parent
  stereo_parent
  tautomer_parent

Calculate descriptors

- Calculate fingerprint
- Cluster fingerprints into folds using locality-sensitive hashing (LSH)

Hash result files

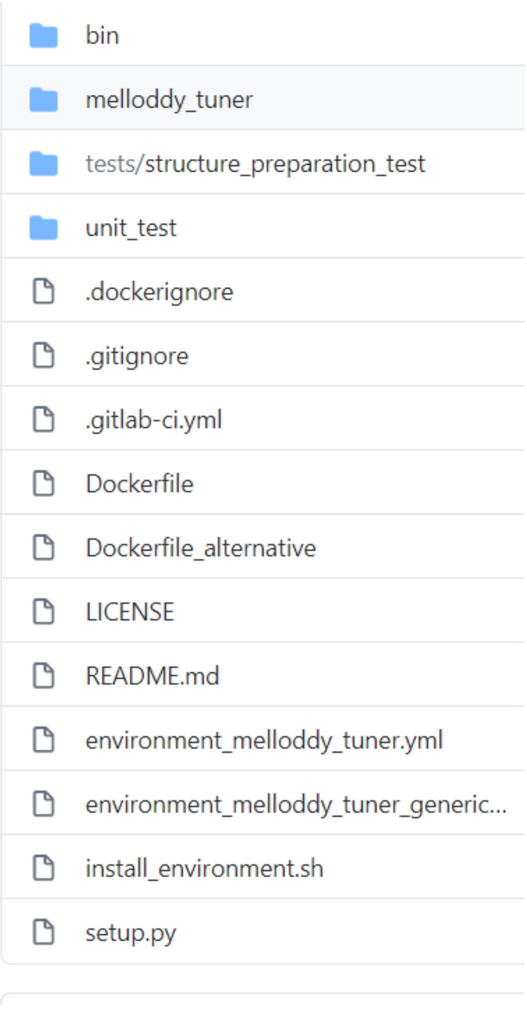
- Hash generated output files and parameter file
- Compare hash with a distributed reference hash

standardize_smiles.py **calculate_descriptors.py**

hash_reference_set.py

MELLODDY-TUNER:
TECHNICAL DETAILS

MELLODDY-TUNER@Github



- bin
- melloddy_tuner
- tests/structure_preparation_test
- unit_test
- .dockerignore
- .gitignore
- .gitlab-ci.yml
- Dockerfile
- Dockerfile_alternative
- LICENSE
- README.md
- environment_melloddy_tuner.yml
- environment_melloddy_tuner_generic...
- install_environment.sh
- setup.py

Python 3.6 or higher

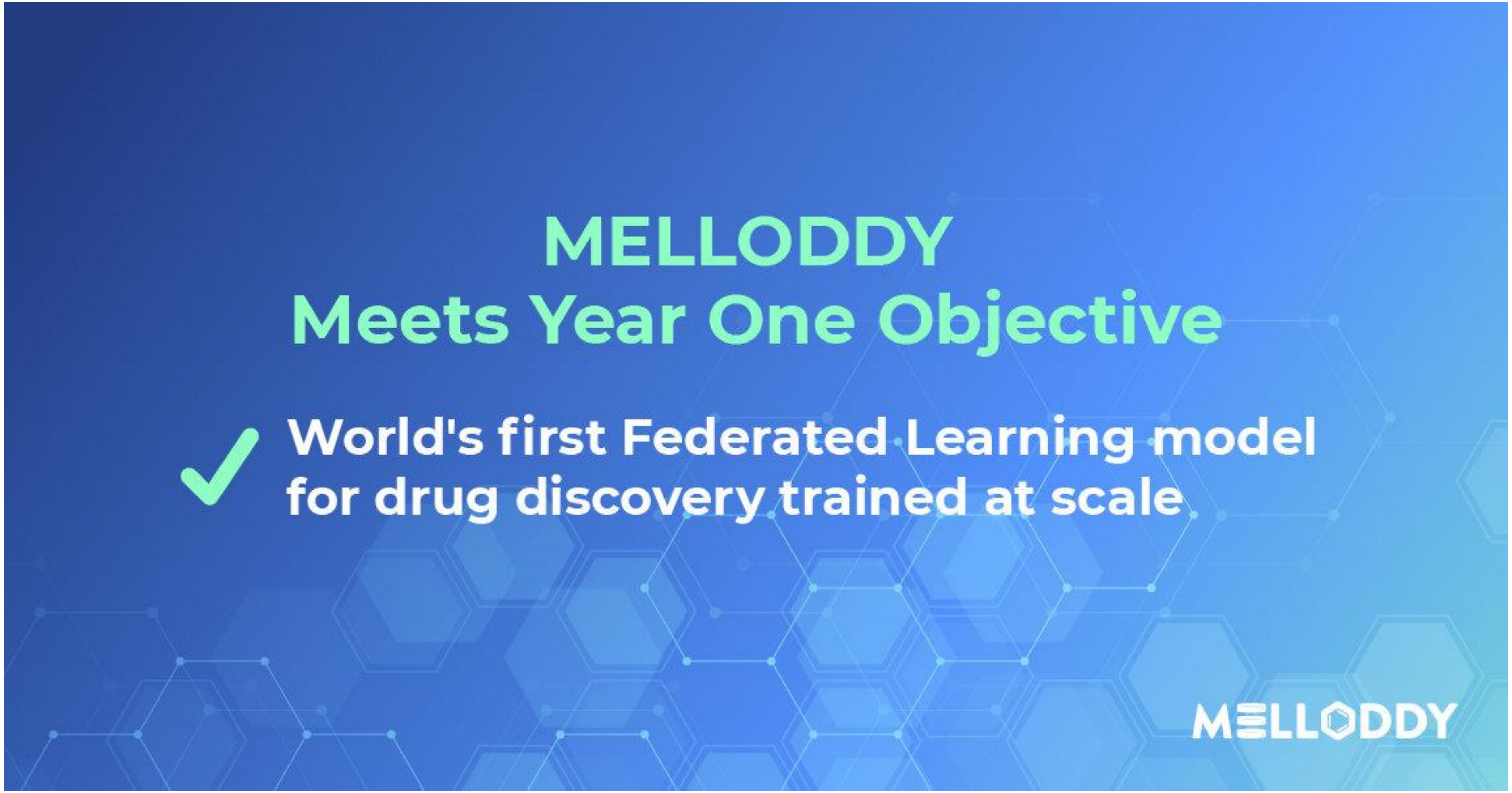
Conda environment

Docker image available

Machine Learning Code (local version):

SparseChem@Github

OBJECTIVE YEAR 1



MELLODDY
Meets Year One Objective

✓ World's first Federated Learning model for drug discovery trained at scale

Press release Sept. 17th , 2020

MELLODDY PROJECT
YEAR 2

Hugo Ceulemans (Janssen Pharmaceuticals):
“[…] Over the next year we’ll turn our focus on studying the hypothesis that multi-partnered modeling will yield superior predictive models for drug discovery.”

Press release Sept. 17th , 2020

MELLODDY

ACKNOWLEDGEMENTS

- Jaak Simm (KU Leuven)
- Adam Arany (Ku Leuven)
- Noé Sturm (Novartis)
- Ansgar Schuffenhauer (Novartis)
- Nikolas Fechner (Novartis)
- Anastasia Pentina (Bayer)
- Tobias Morawietz (Bayer)
- Lina Humbeck (Boehringer-Ingelheim)
- Peter Schmidtke (Servier-Discngine)
- Adam Zalewski (Amgen)
- Michael Krug (Merck KGaA)
- Wouter Heyndrickx (Janssen Pharmaceuticals)
- Lewis Mervin (AstraZeneca)
- Arnaud Gohier (Servier)