QSPRPRED

A FLEXIBLE OPEN-SOURCE QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING TOOL

LACDR

Universiteit Leiden
The Netherlands

Helle W. van den Maagdenberg, Martin Šícho, David Araripe, Sohvi Luukkonen, Linde Schoenmaker, Michiel Jespers, Olivier J. M. Béquignon, Marina Gorostiola González, Remco L. van den Broek, Andrius Bernativicius, J.G. Coen van Hasselt, Piet H. van der Graaf, and Gerard J. P. van Westen

WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods

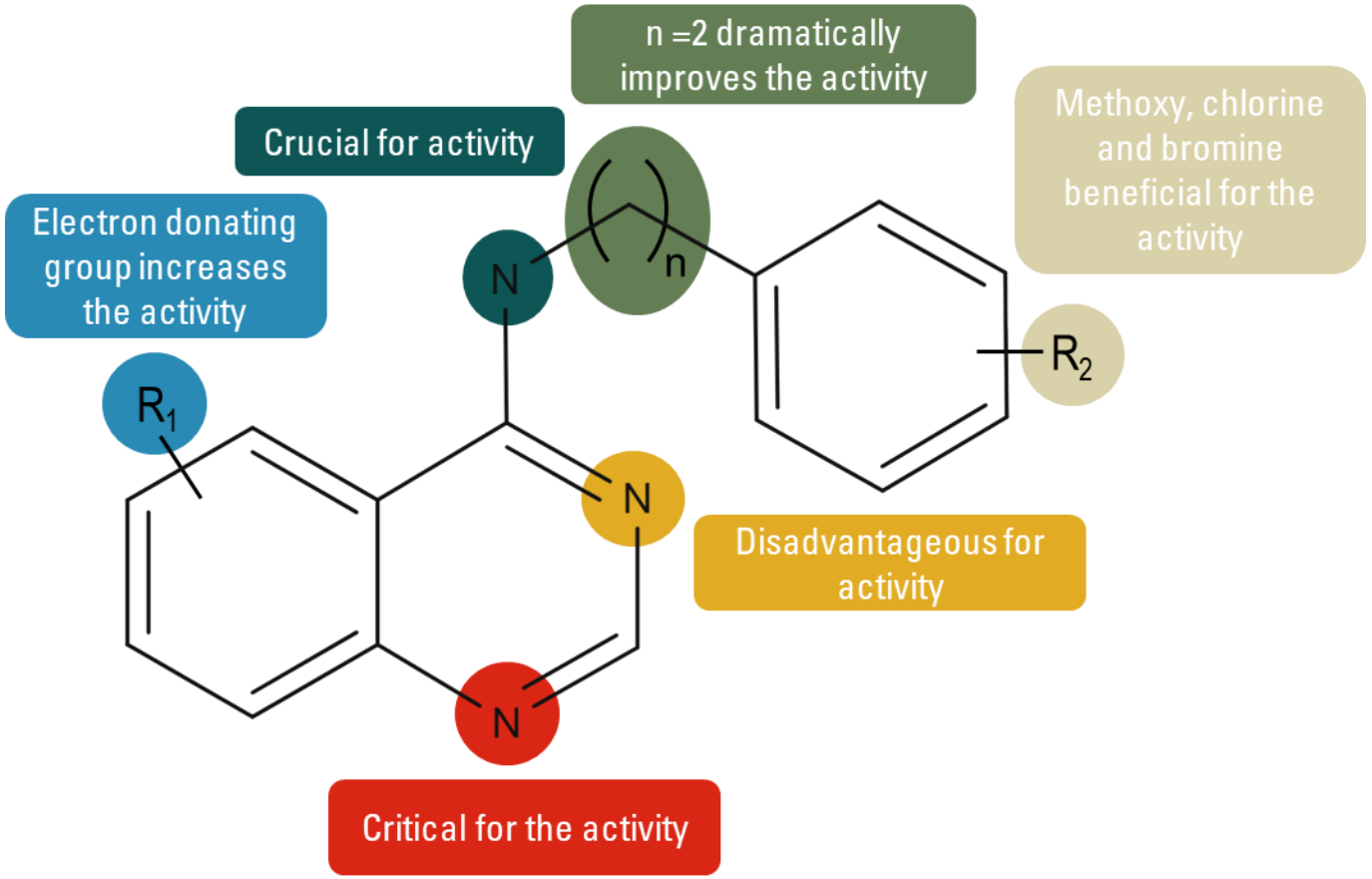


*Image adapted from Elsocht et al. *Int J Mol Sci.* 2020

WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods

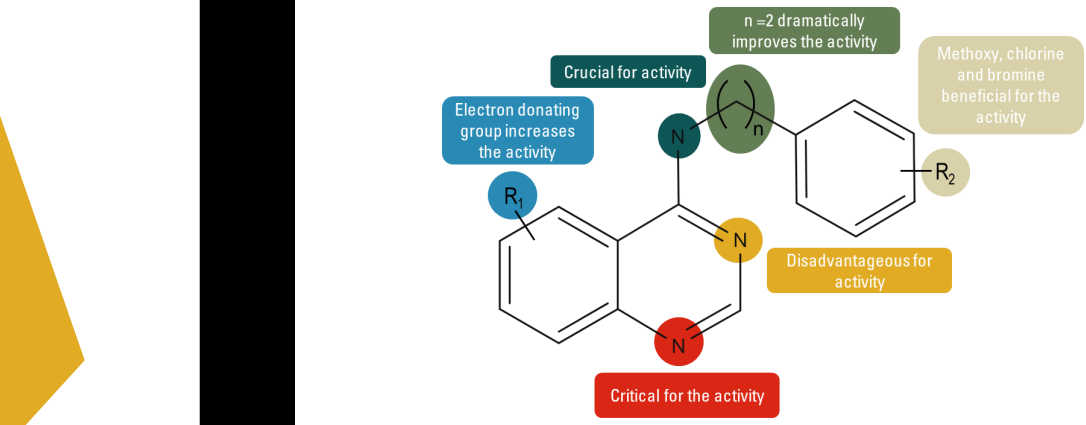
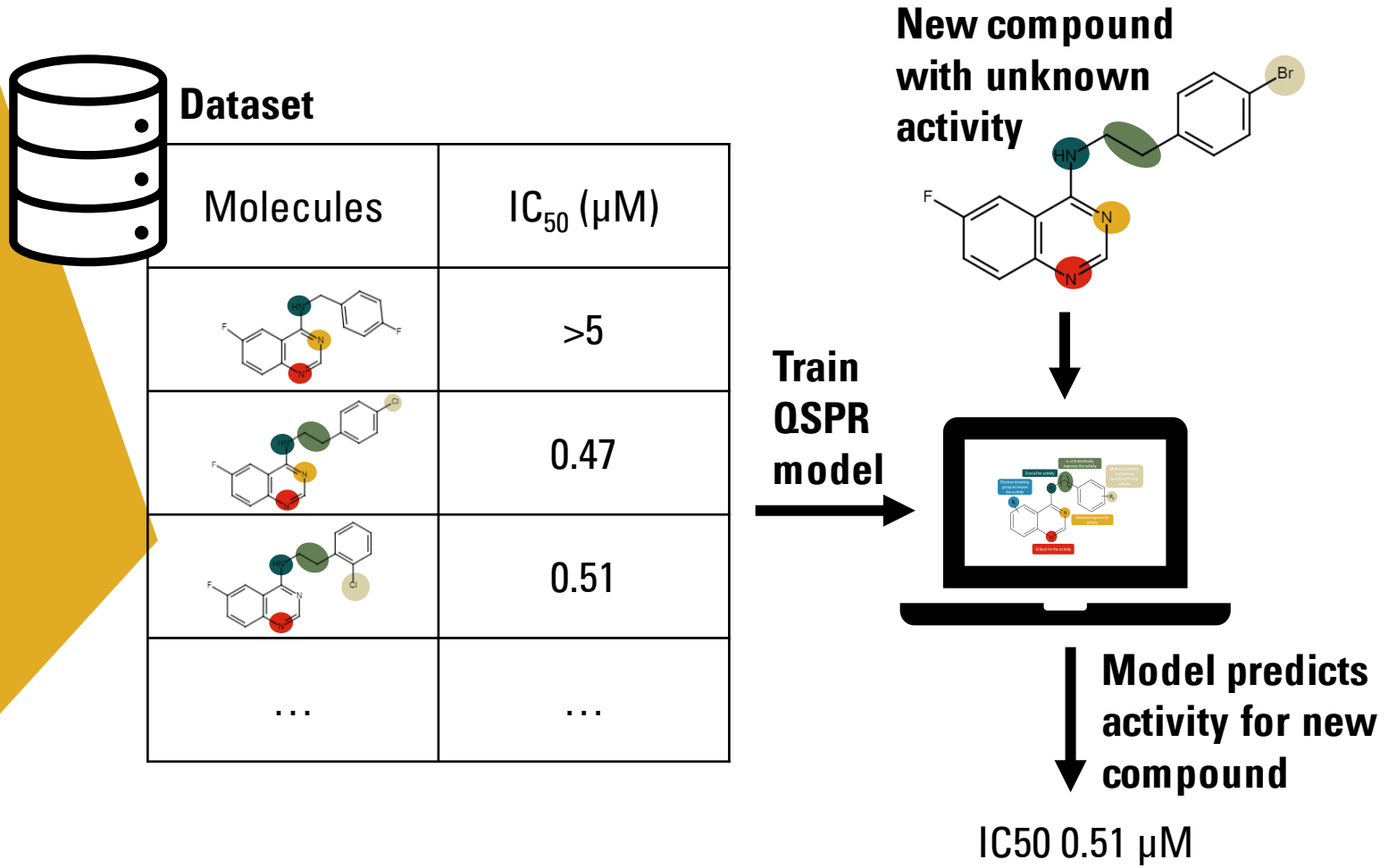


*Image adapted from Elsocht et al. *Int J Mol Sci*. 2020

WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

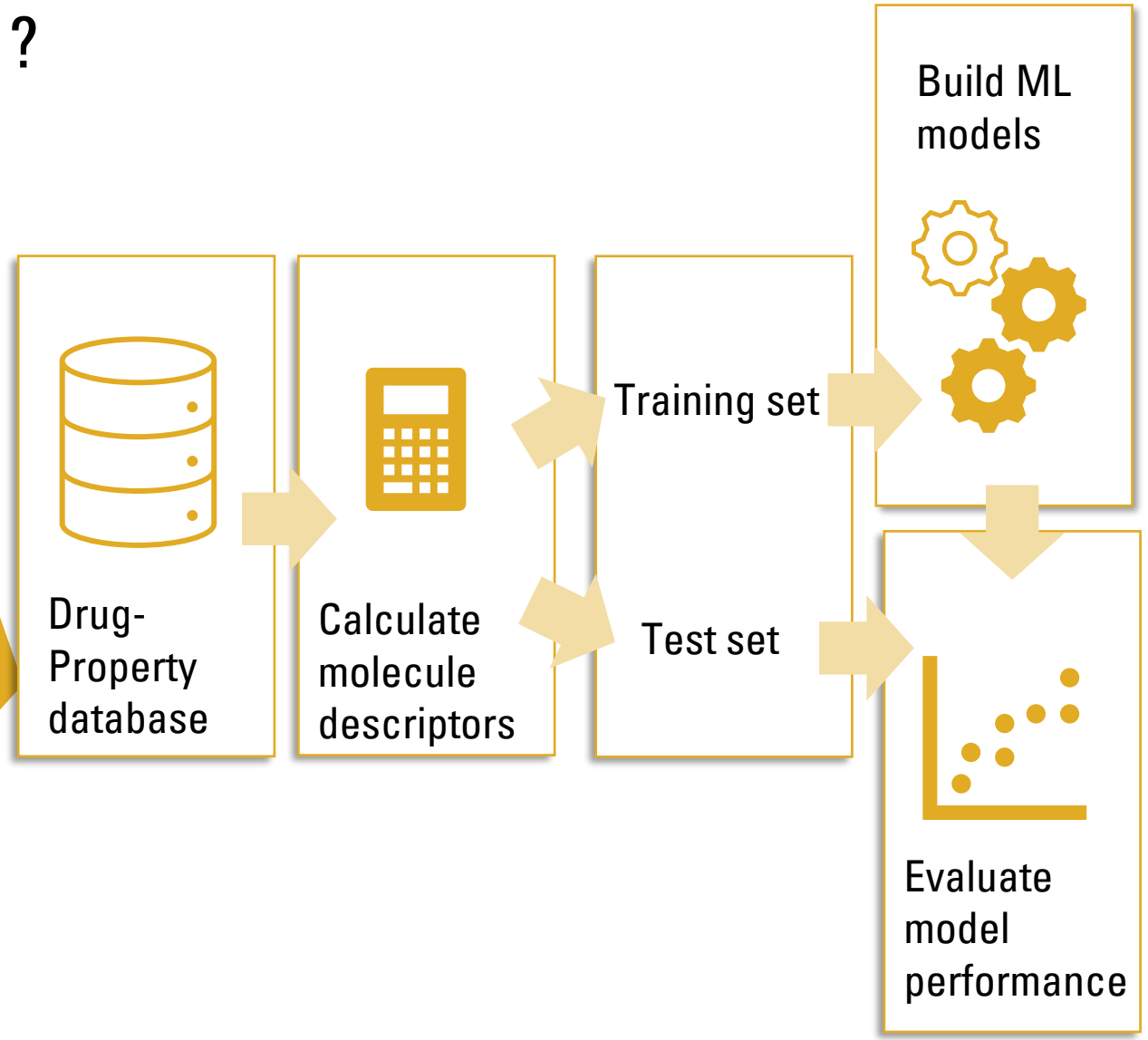
- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods



*Example dataset adapted from Elsocht et al. *Int J Mol Sci.* 2020

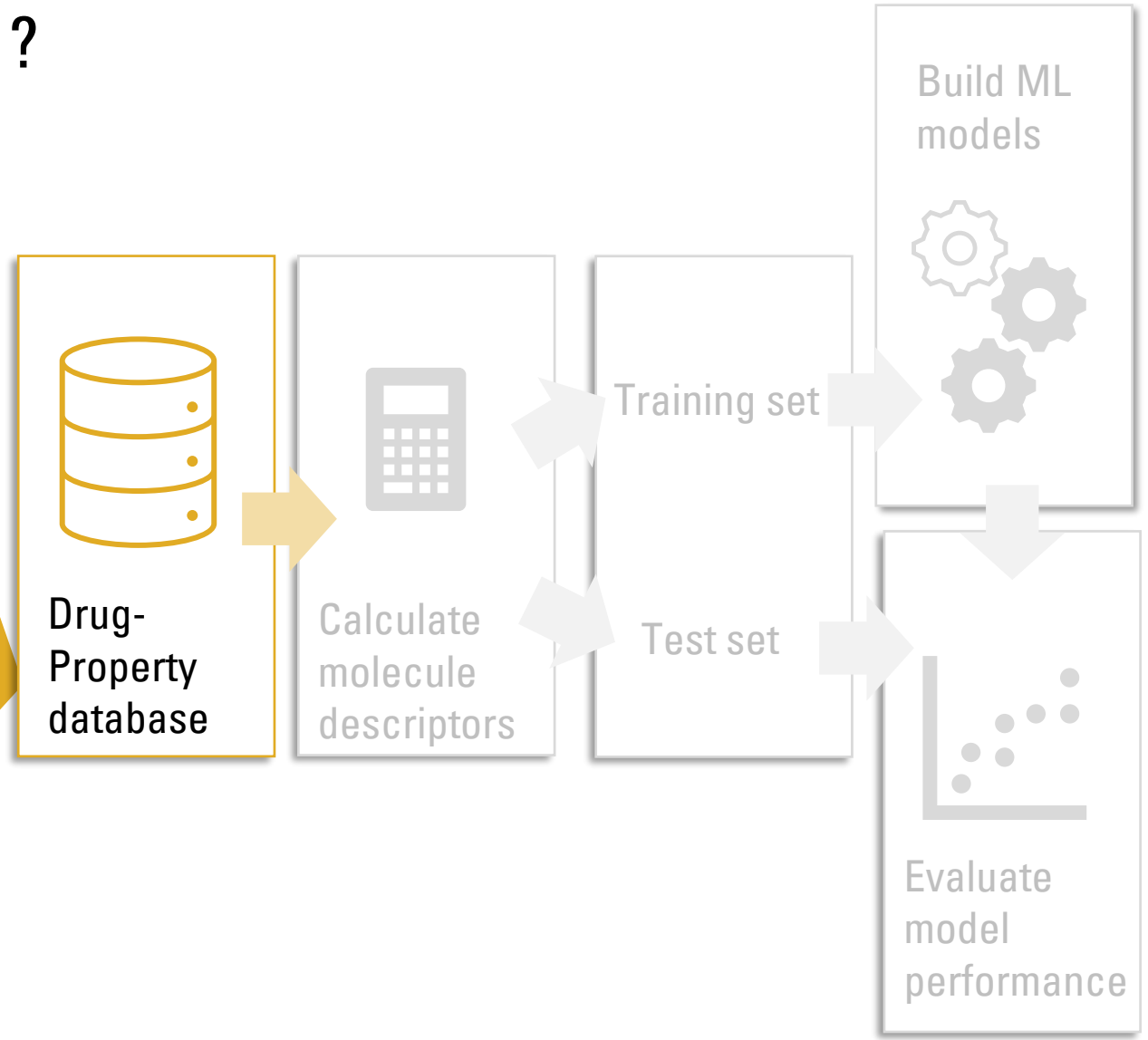
WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods



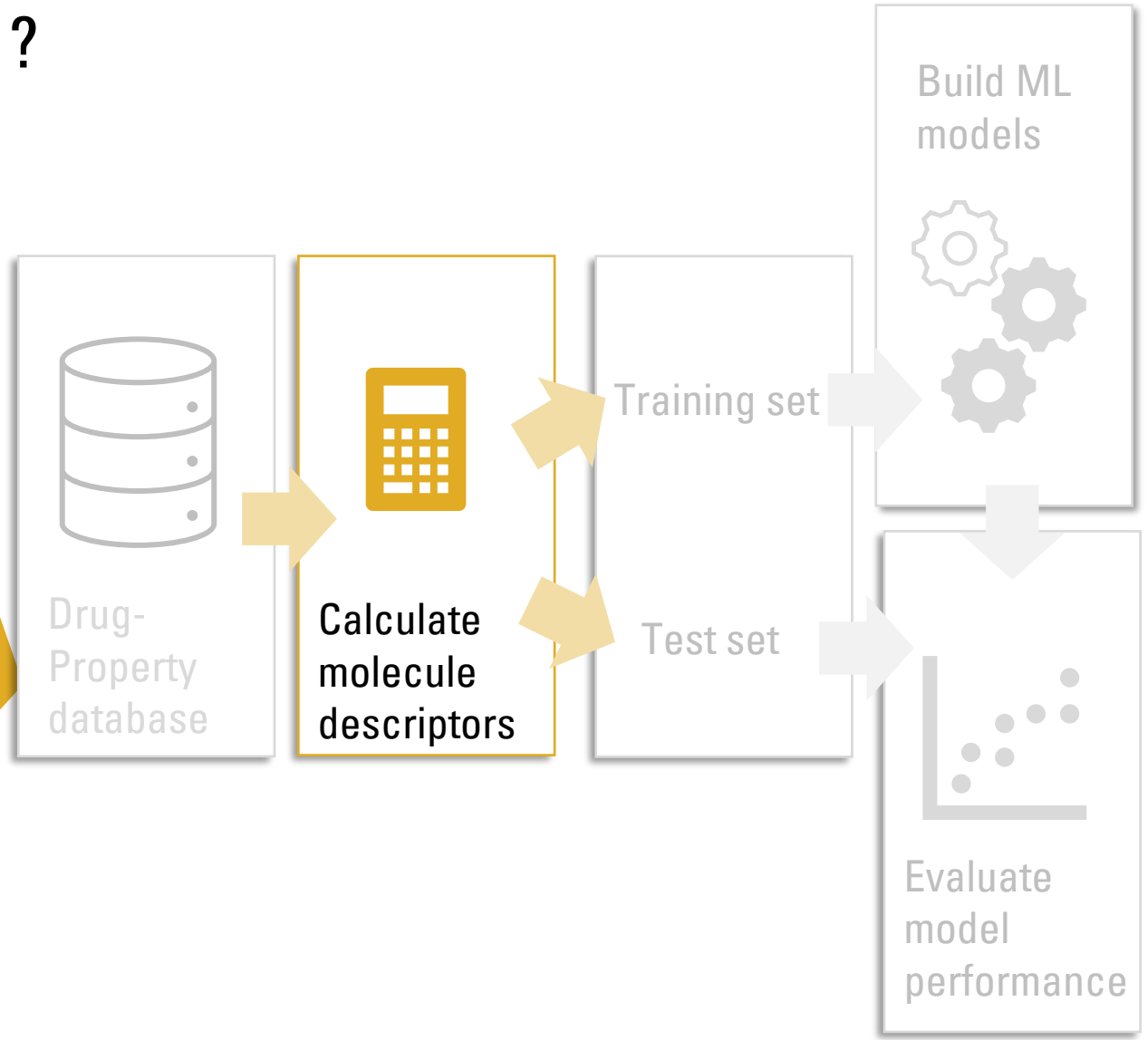
WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods



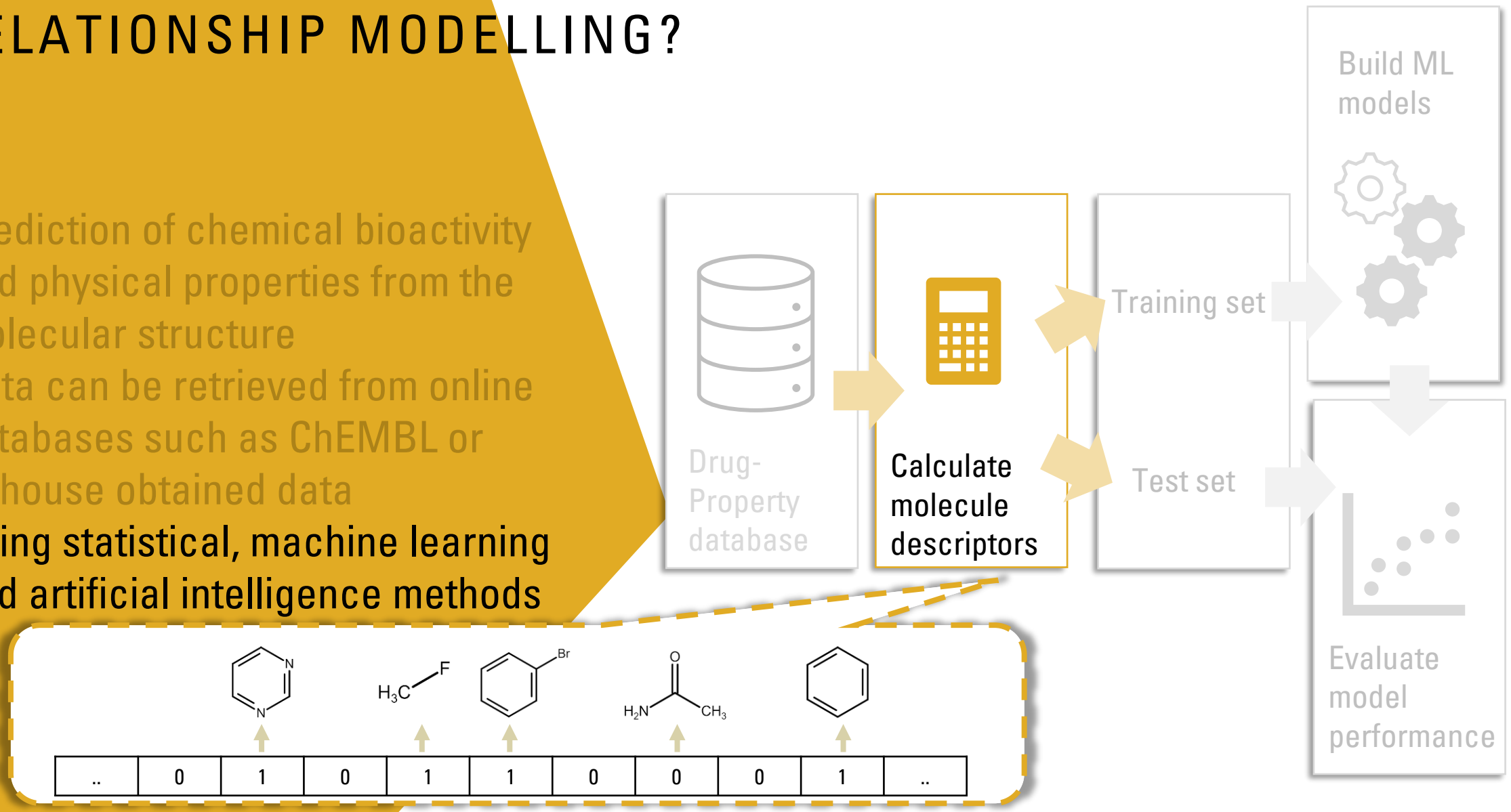
WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods



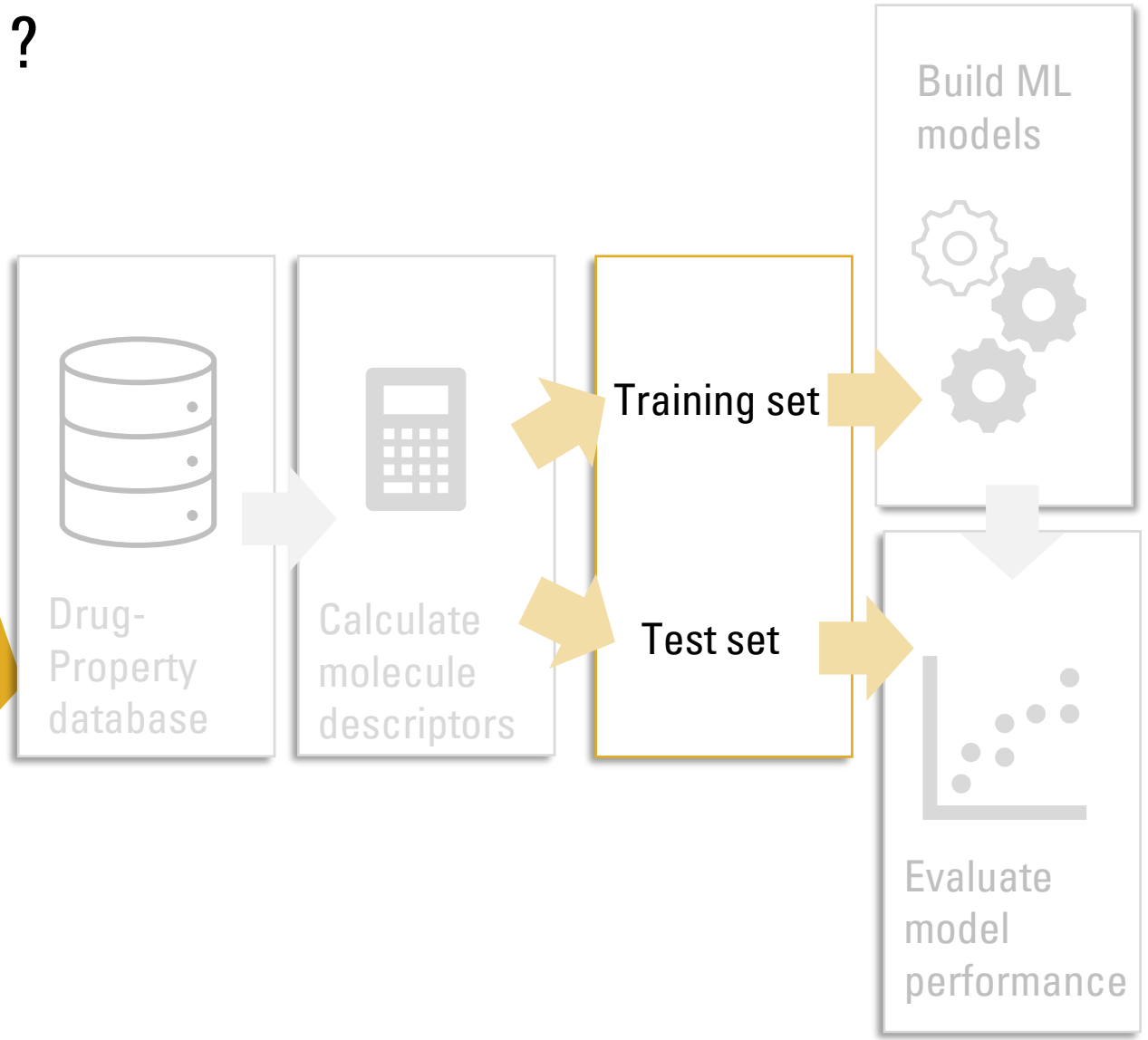
WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods



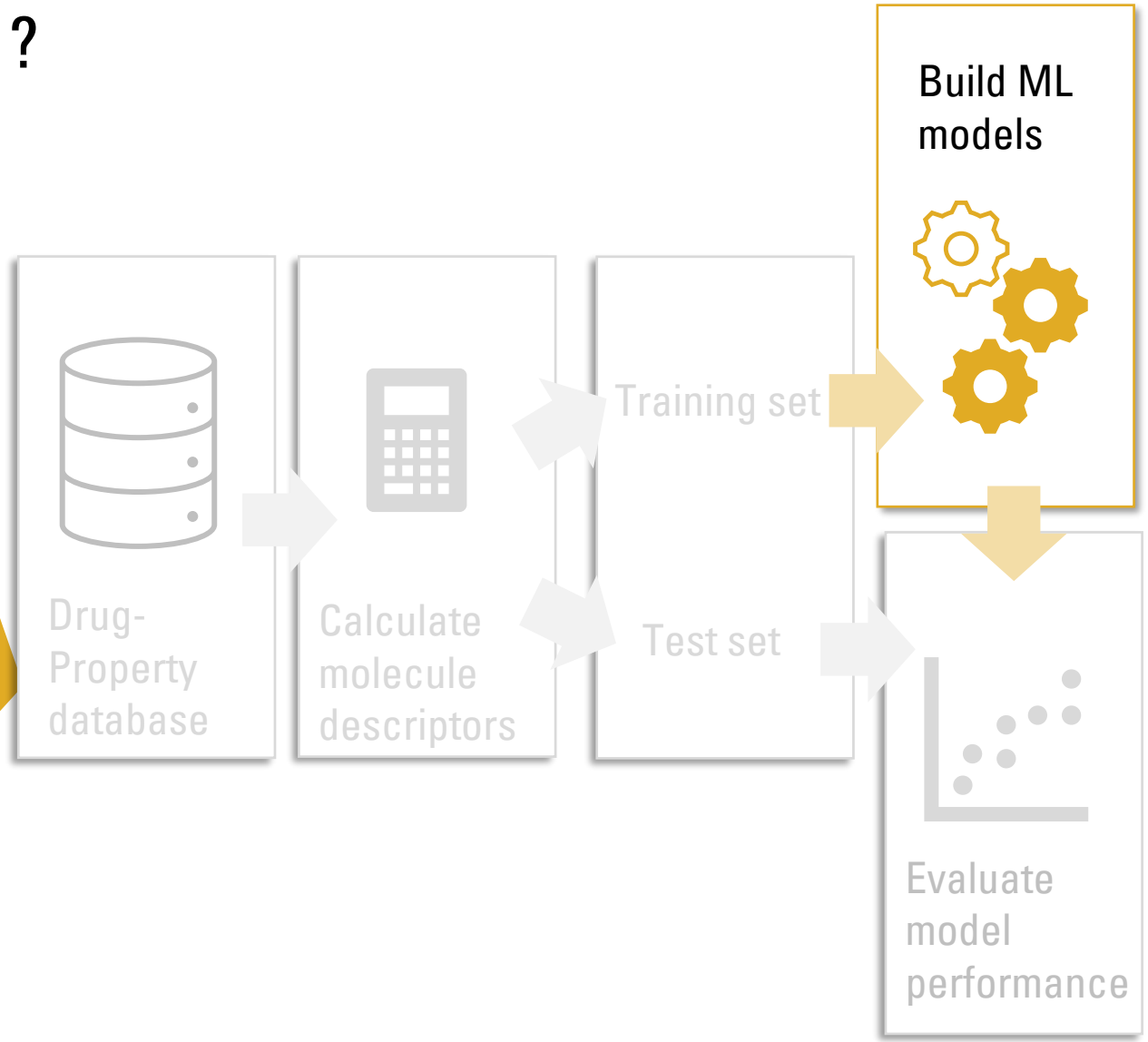
WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods



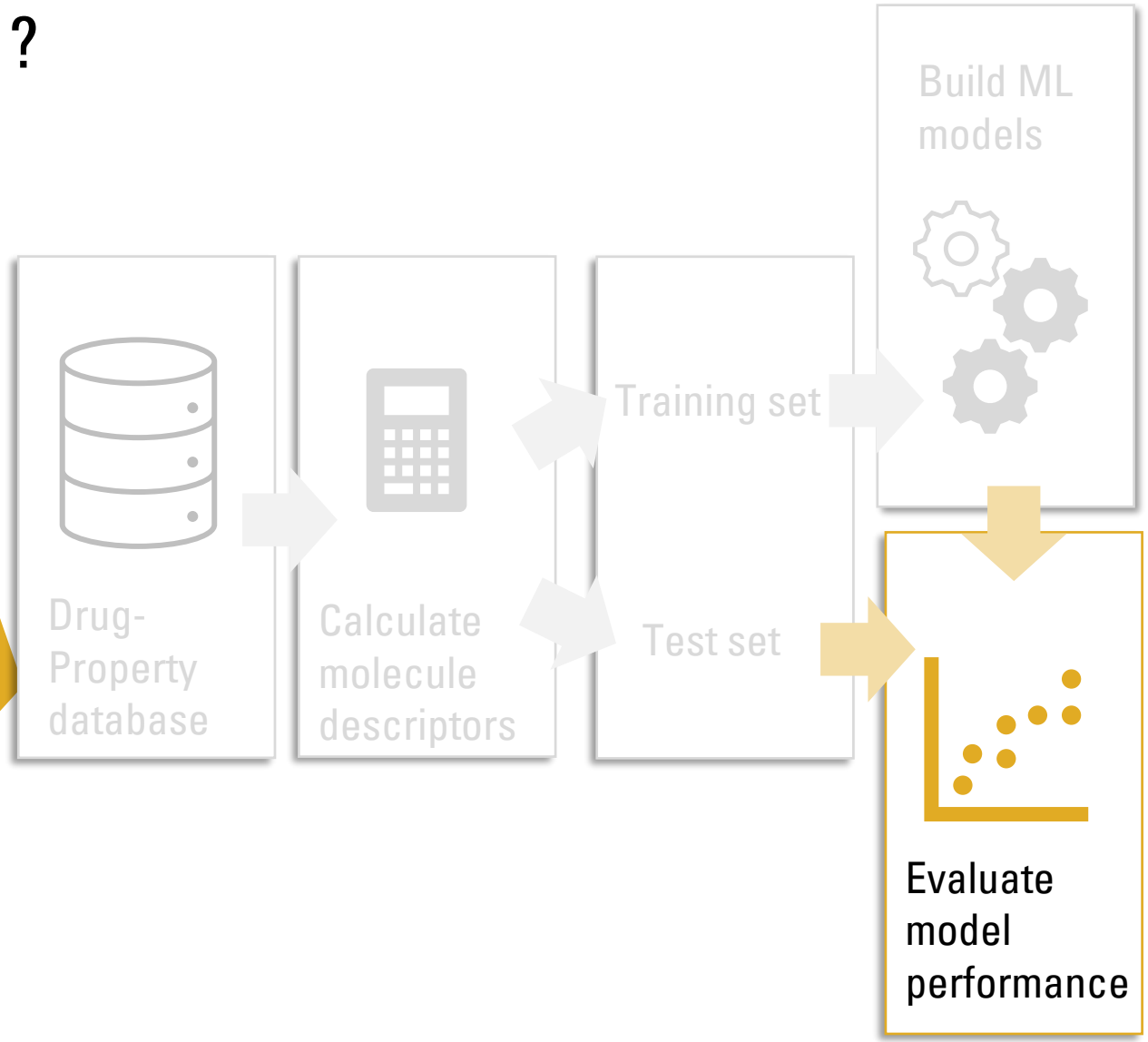
WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods



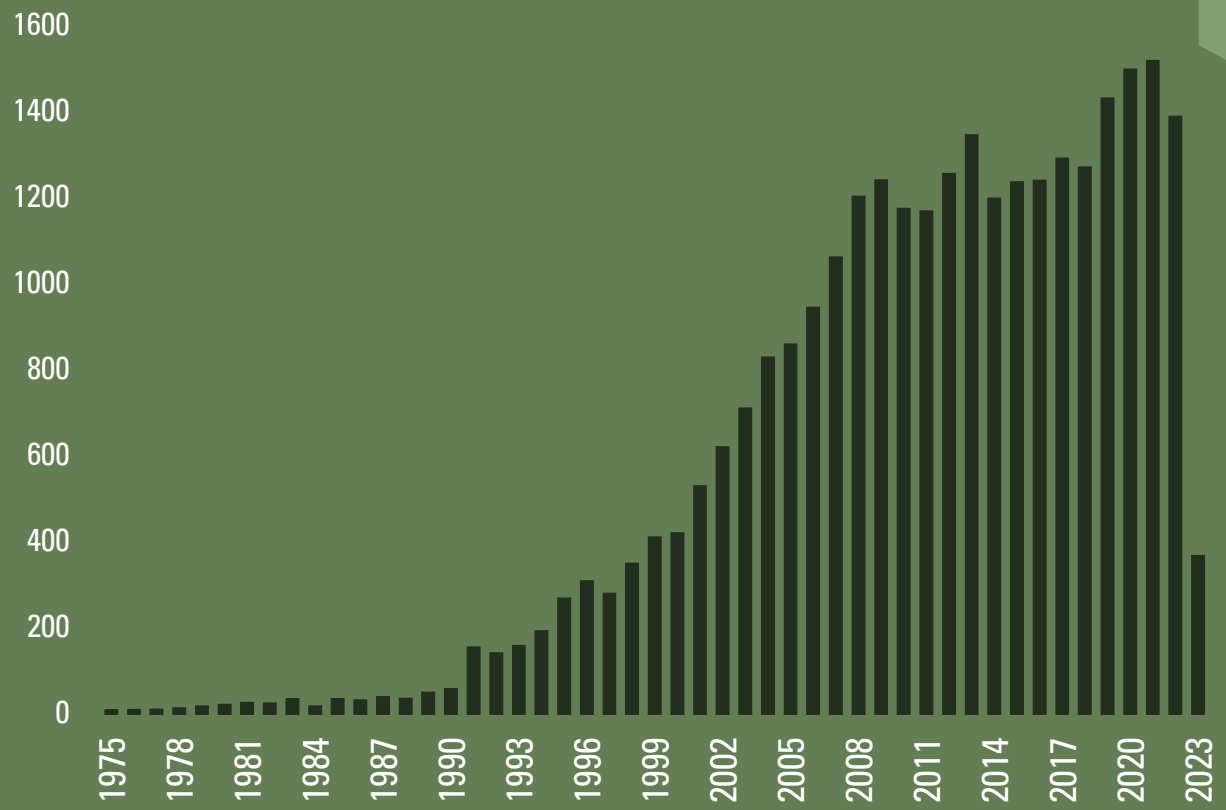
WHAT IS QSPR: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELLING?

- Prediction of chemical bioactivity and physical properties from the molecular structure
- Data can be retrieved from online databases such as ChEMBL or in-house obtained data
- Using statistical, machine learning and artificial intelligence methods



WHY DO WE NEED A TOOL FOR QSPR MODELLING?

Number of papers on QSPR modelling from Web of Science



- QSPR modelling is used in both industry and academia.
- Development typically involves common steps and components.
- Cheminformaticians prefer the flexibility of Python over available tools.
- Experimenting with different models and workflows increases code complexity.

A flexible and easy to use Quantitative Structure-Property Relationship Modelling Framework: QSPRpred

- Modular: simple to add new models, descriptors, etc.
- Easy to use: Includes Command Line Interface, Python API and tutorials.
- New features are regularly added
- Compatible with *de novo* drug design package: DrugEx*
- Code available on GitHub and pip installable

* Liu, X., et al. J Cheminf. (2019) * Šícho M. *et al.*, *J. Chem. Inf. Model.* 2023

THE QSPRPRED WORKFLOW: FUNCTIONALITIES

Input in-house and/or public datasets

Pre-processing
- SMILES sanitization & standardization
- Data transformation
- Exclusion by value
- Class creation

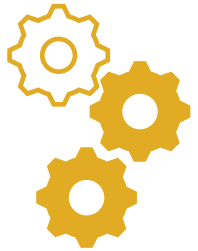
Descriptor calculation:
- Morgan Fingerprints
- RDKit descriptors
- Mordred descriptors
- Etc.
- **Protein descriptors**

Training set

Test set

Split types:
- Random
- Temporal
- Scaffold

Build ML models



Model types:
- Scikit-learn & PyTorch
- Regression
- Classification
- Multi-class
- Multi-task
- **Proteo-chemometrics**

Evaluate model performance

1. Hyperparameter optimization
   - Grid search
   - Bayes optimization
2. Cross-validation
3. Visualization

THE QSPRPRED WORKFLOW: FUNCTIONALITIES

Data collection with Papyrus*

- Curated ligand-protein bioactivity data
- Multiple publicly available datasets such as ChEMBL
- Out-of-the-box usable with QSPRpred


Pre-processing

- SMILES sanitization & standardization
- Data transformation
- Exclusion by value
- Class creation

Descriptor calculation:

- Morgan Fingerprints
- RDKit descriptors
- Mordred descriptors
- Etc.
- **Protein descriptors**

Training set

Test set

Split types:

- Random
- Temporal
- Scaffold

Build ML models

Model types:

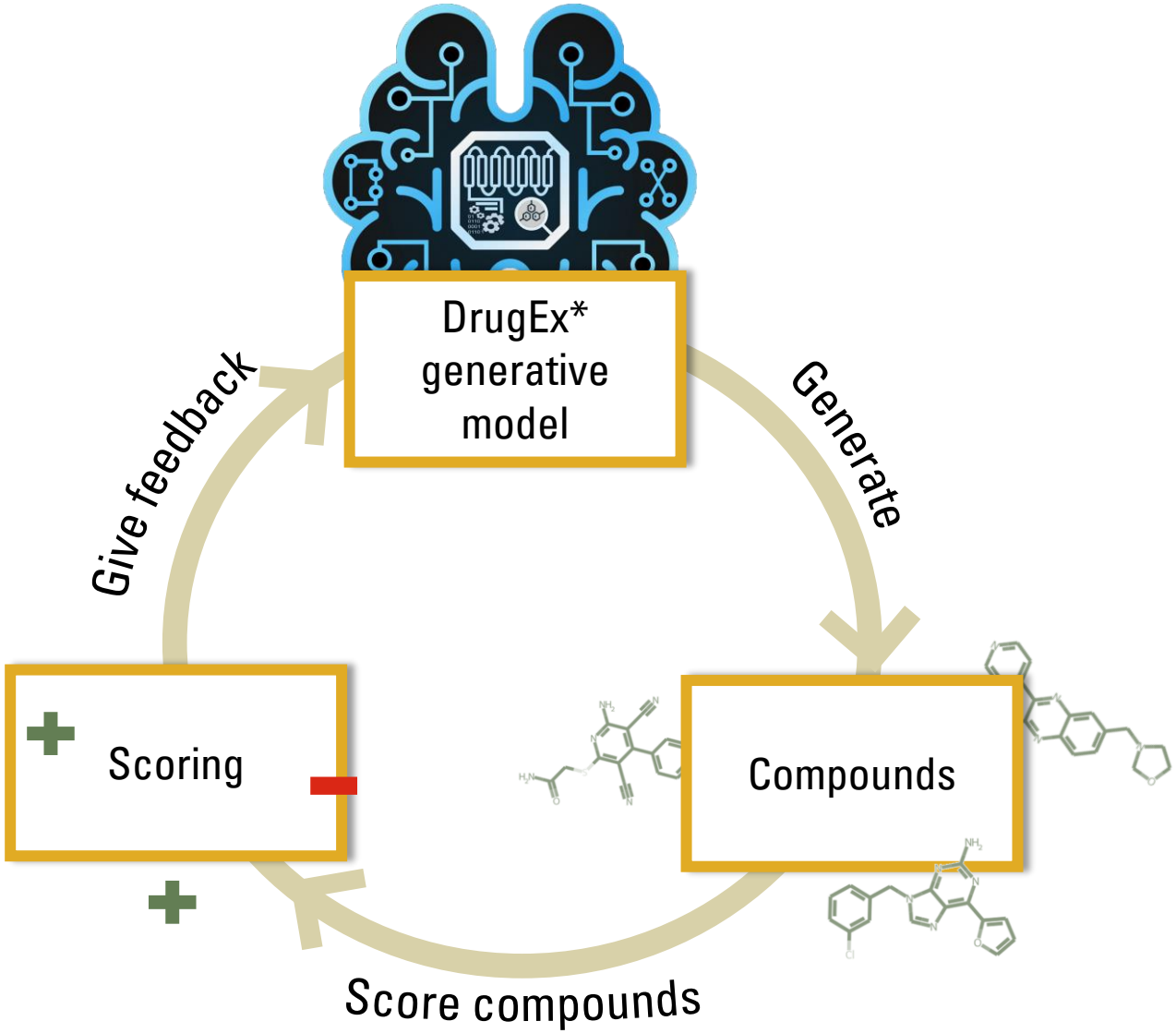
- Scikit-learn & PyTorch
- Regression
- Classification
- Multi-class
- Multi-task
- **Proteo-chemometrics**

Evaluate model performance

1. Hyperparameter optimization
   - Grid search
   - Bayes optimization
2. Cross-validation
3. Visualization

*Béquignon, O.J.M., *et al. J. Cheminf. 2023*

DRUGEX: GENERATE NOVEL MOLECULES



* Liu, X., et al. J Cheminf. (2019) * Šícho M. *et al.*, *J. Chem. Inf. Model.* 2023

DRUGEX: GENERATE NOVEL MOLECULES

Scoring

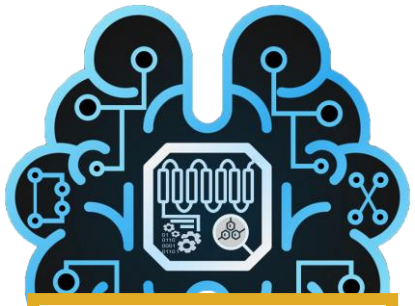
QSPR models
QSPRpred model predictions can be used to score molecules, e.g., for high binding affinity

Synthetic accessibility* (e.g. Led3Score)
Alan Kai Hassen

In development: docking

Spock: High throughput docking
Martin Sicho
& Andrius Bernativicius

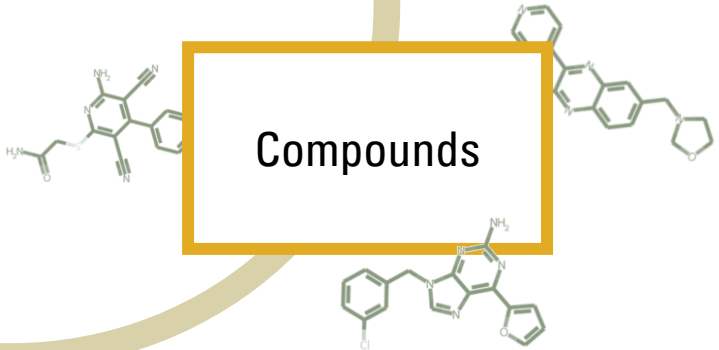
Any custom score, e.g
Physciochemical properties



DrugEx generative model

Generate

Compounds



Score compounds

Scoring

Give feedback

*Preprint. Hassen A.K. *ChemRxiv*. 2024. doi:10.26434/chemrxiv-2024-wtjt6.

HELPFUL TUTORIALS TO GET STARTED

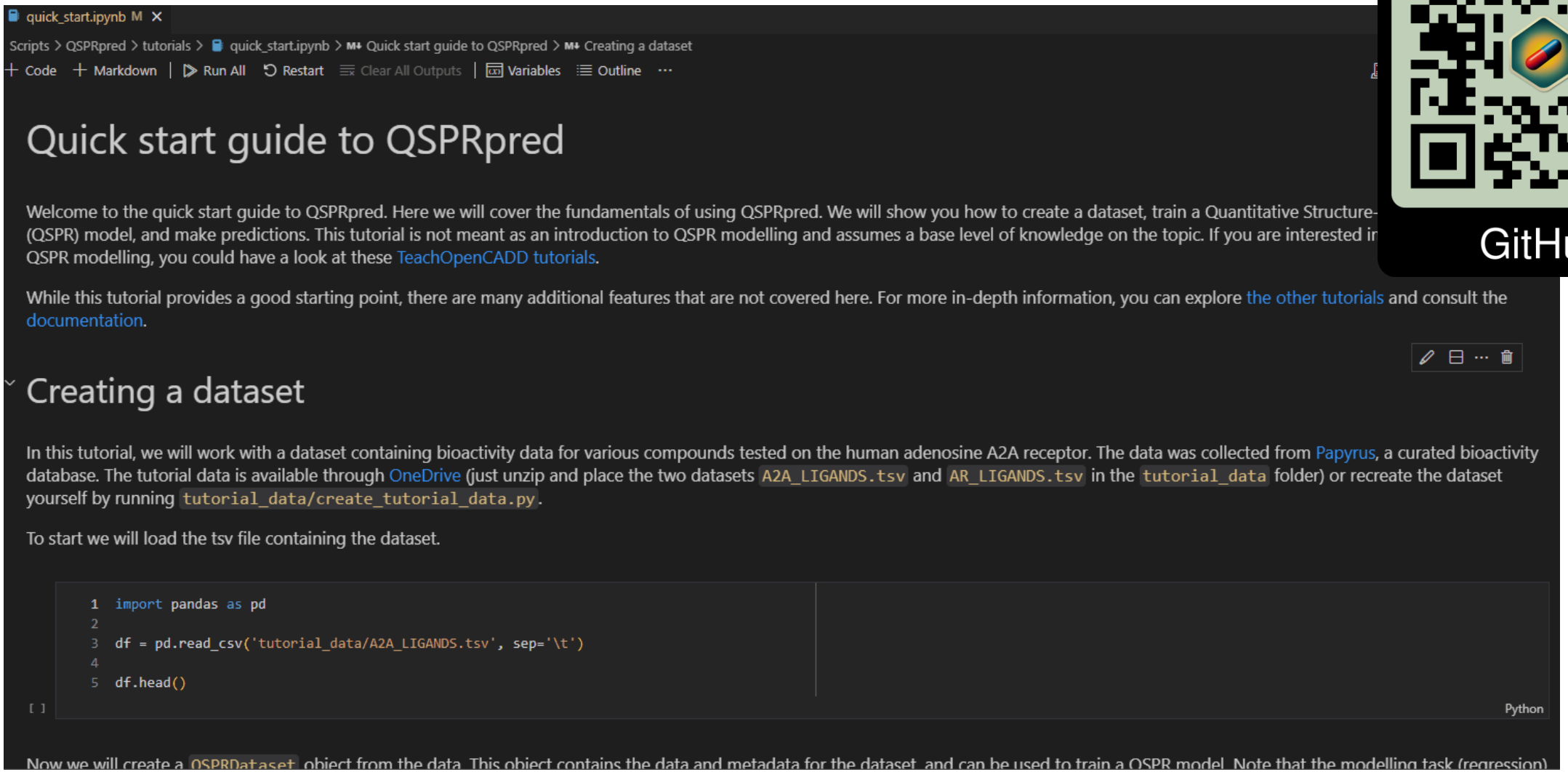
Table of Contents

This tutorial tries to cover the most important topics on the various features of QSPRpred, but it is not exhaustive. For more detailed information on the features of the package, please refer to the documentation.The tutorial data is available through OneDrive (just unzip and place the two datasets `A2A_LIGANDS.tsv` and `AR_LIGANDS.tsv` in the `tutorial_data` folder) or recreate the dataset yourself by running `tutorial_data/create_tutorial_data.py` after you have installed QSPRpred.

The Quick Start tutorial is designed to get you up and running with QSPRpred as quickly as possible while the rest dedicates more time to explain each feature in more detail. The Basics cover the most commonly used functionality of QSPRpred. The Advanced tutorials cover more advanced topics and are designed for users who are already familiar with QSPRpred more in depth or are looking for more niche features. For detailed description of all QSPRpred classes and functions, as well as examples of how to use the command line interface, see the documentation pages.

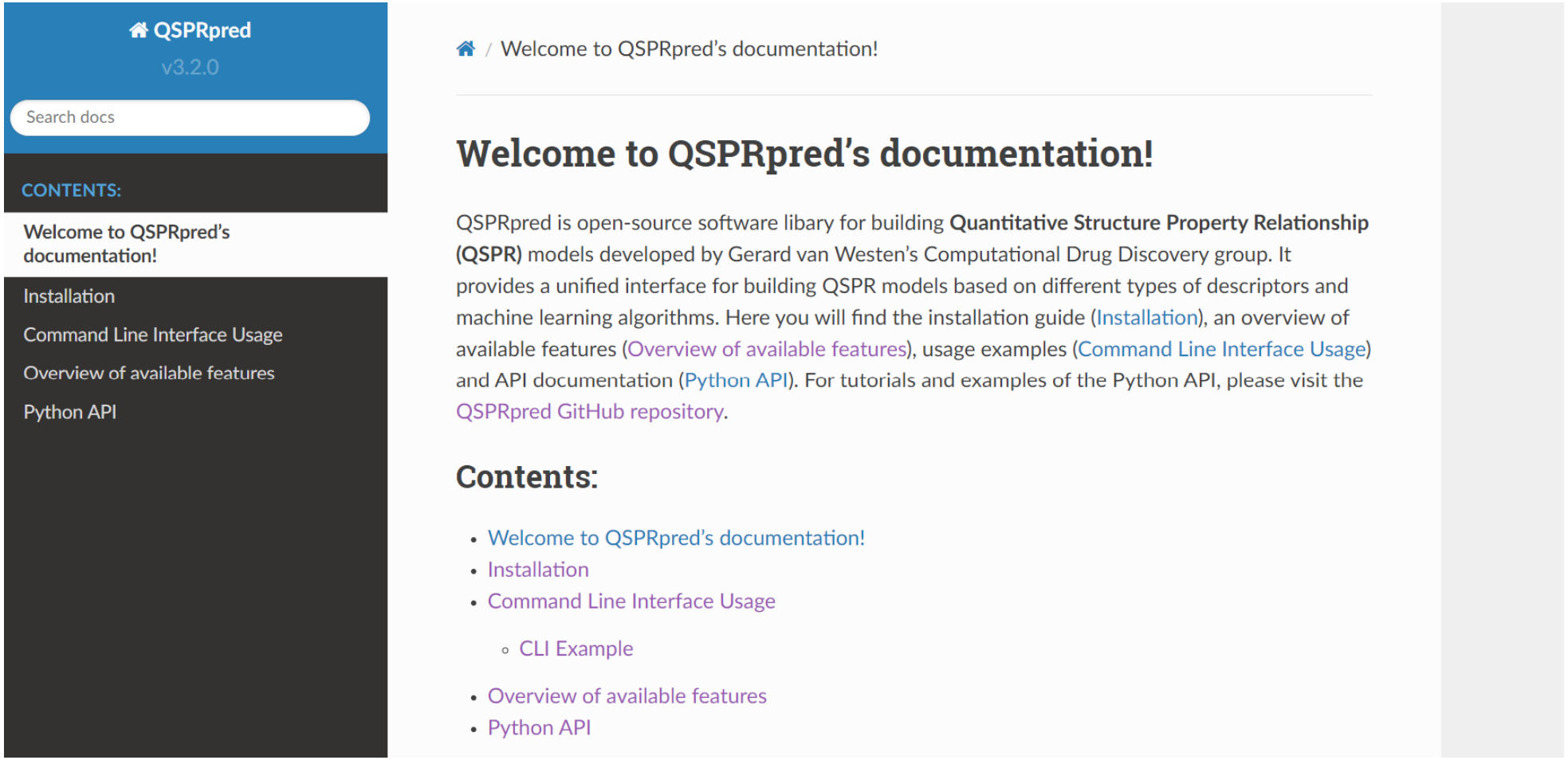
- **Quick Start**: A quick start guide to using QSPRpred.
- **Basics**
  - Data
    - Data Collection with Papyrus: How to collect data with Papyrus.
    - Data Preparation: How to prepare data for QSPRpred.
    - Data Representation: How data is represented in QSPRpred (MolTable, QSPRDataset, etc.).
    - Data Splitting: How to split data into training, validation, and test sets.
    - Descriptors: How to calculate descriptors for molecules.
    - Searching, Filtering and Plotting: How to search and filter data.
    - Applicability Domain: How to calculate the applicability domain of a model.
  - Modelling
    - Classification: How to train a classification model.
    - Logging: How to set-up logging.
    - Model Assessment: How to assess the performance of a model.
  - Other
    - Benchmarking: How to benchmark QSPRpred.
    - Serialization: How to save and load datasets and models

HELPFUL TUTORIALS TO GET STARTED



GitHub

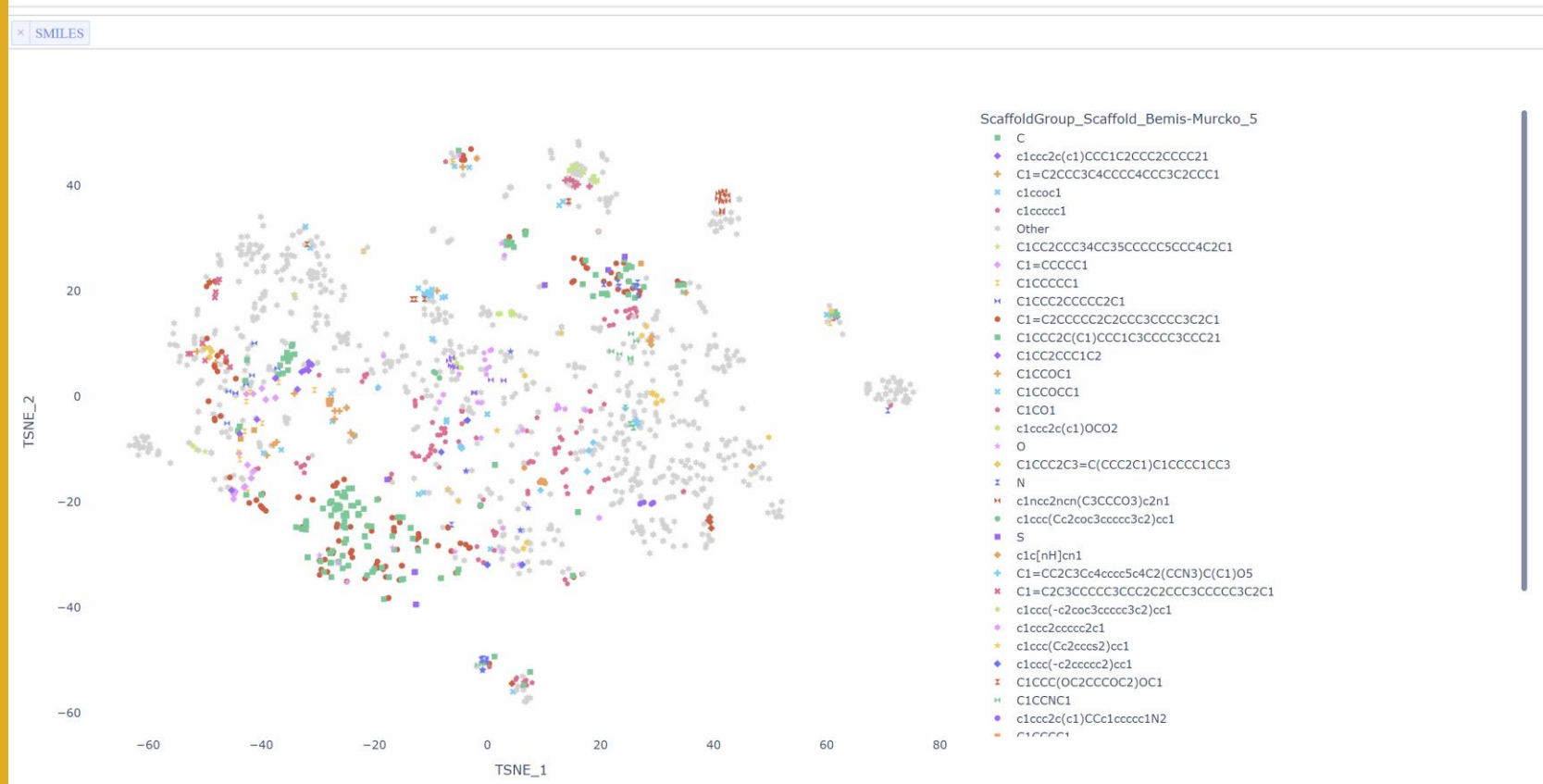
EXTENSIVE DOCUMENTATION



DATASET VISUALIZATION: SCAFFVIZ

- Dataset visualization: papyrus-scaffold-visualizer
- Build on top off Molplotly library*
- Developed by Martin Šícho

Scaffviz



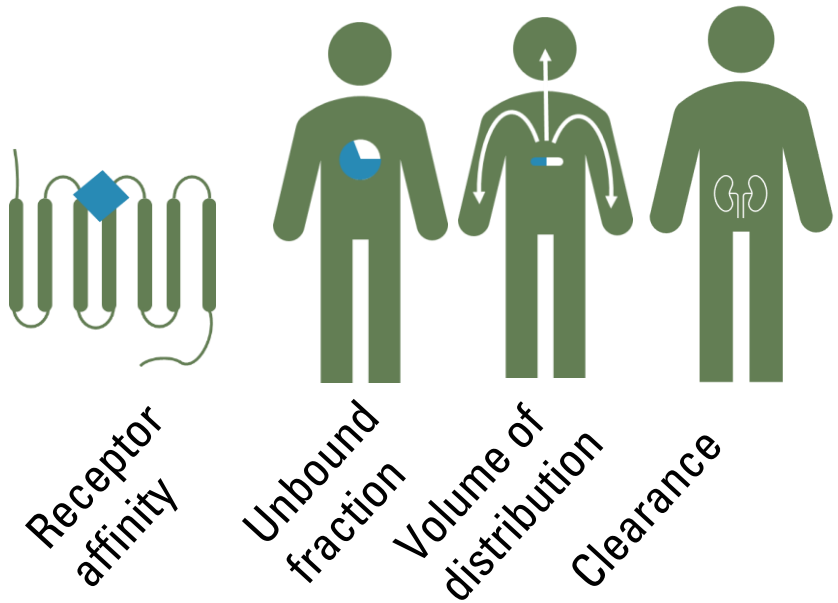
*https://github.com/wjm41/molplotly

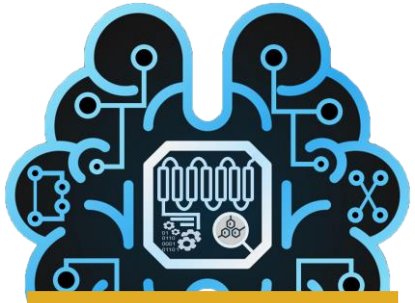
APPLICATION: INTEGRATING PHARMACOKINETICS IN GENERATIVE DRUG DESIGN

USING QSPRPRED PK & AFFINITY MODELS IN DRUGEX

Scoring

QSPR models for adenosine $A_{2A}R$ binding affinity & human pharmacokinetics

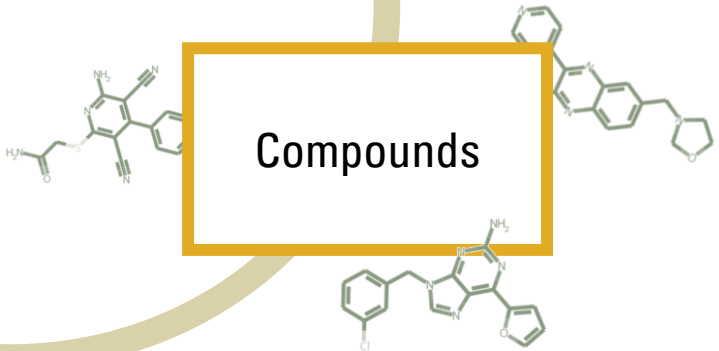




DrugEx generative model

Generate

Compounds



Score compounds

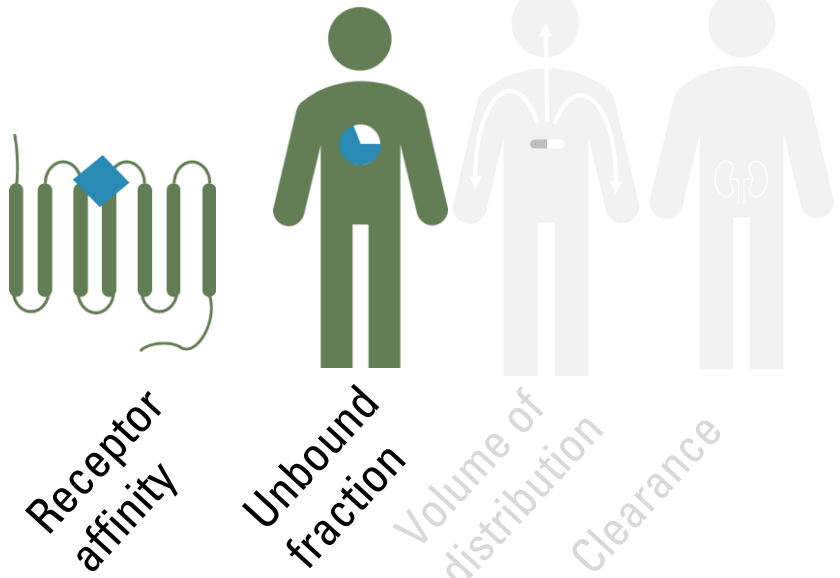
Scoring

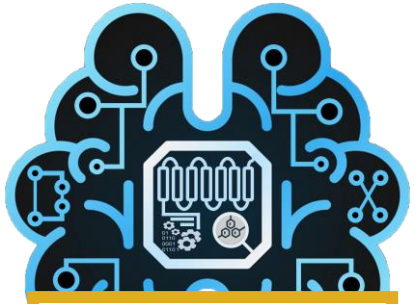
Give feedback

MAXIMIZE THE UNBOUND FRACTION & $A_{2A}R$ AFFINITY

Scoring

QSPR models for adenosine $A_{2A}R$ binding affinity & human pharmacokinetics

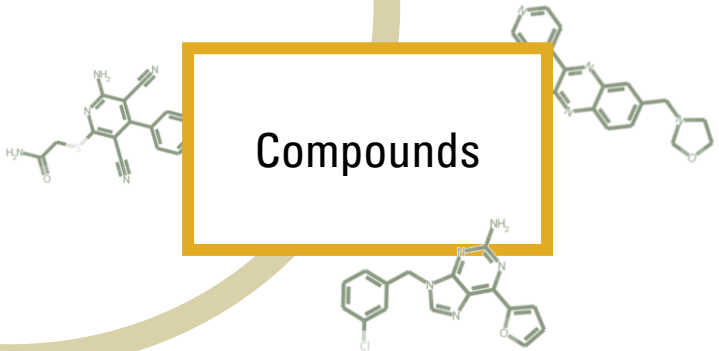




DrugEx generative model

Generate

Compounds

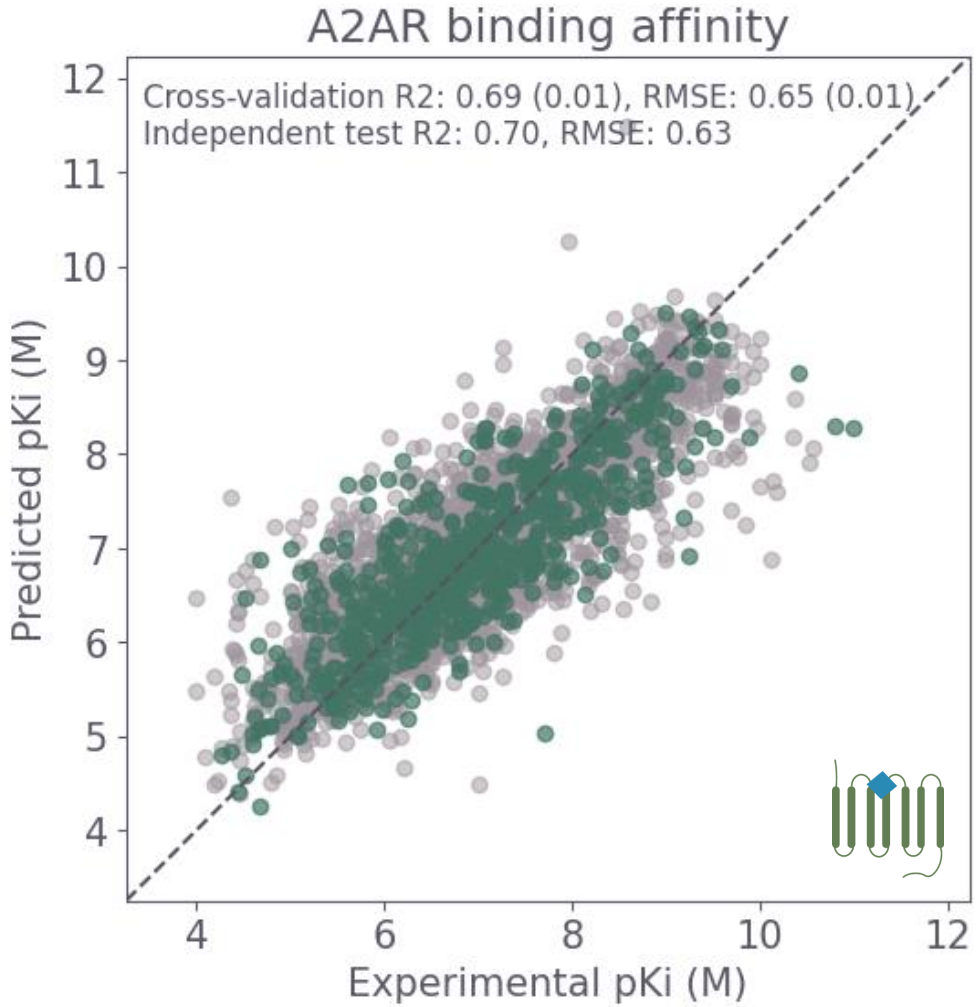


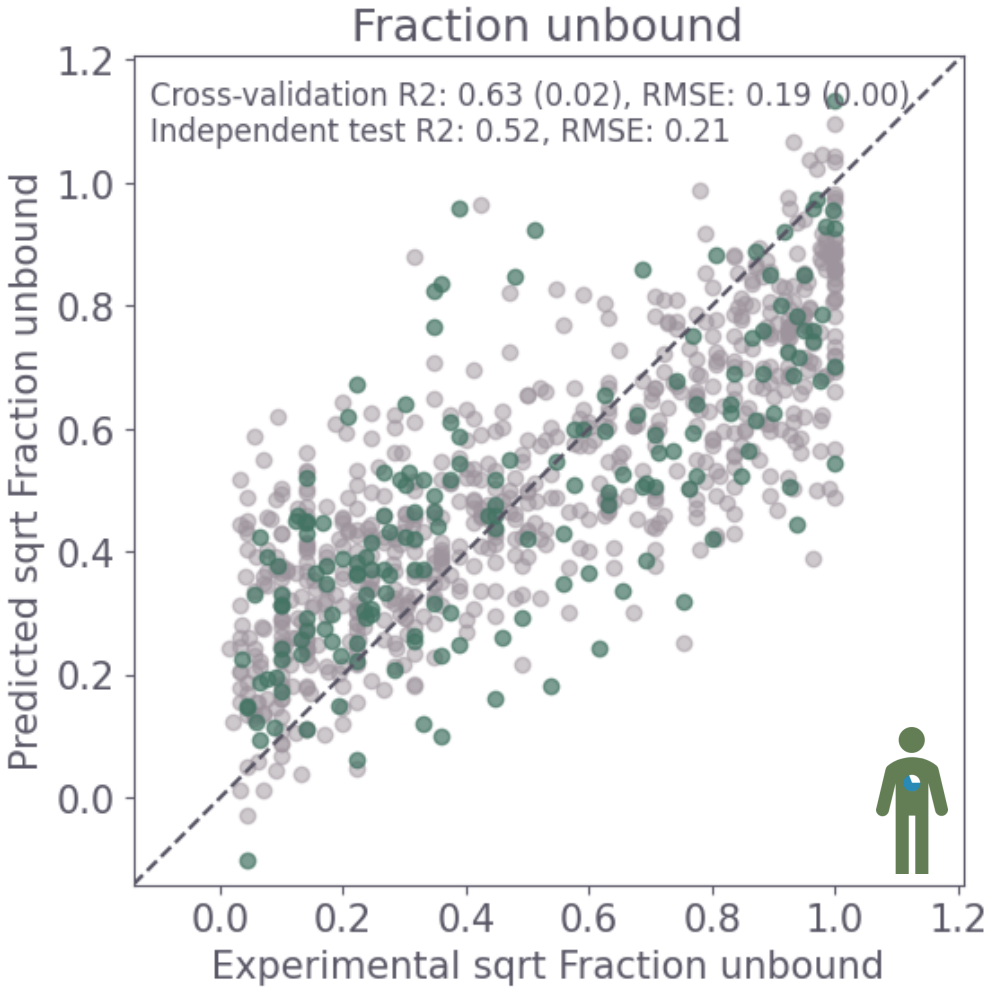
Score compounds

Scoring

Give feedback

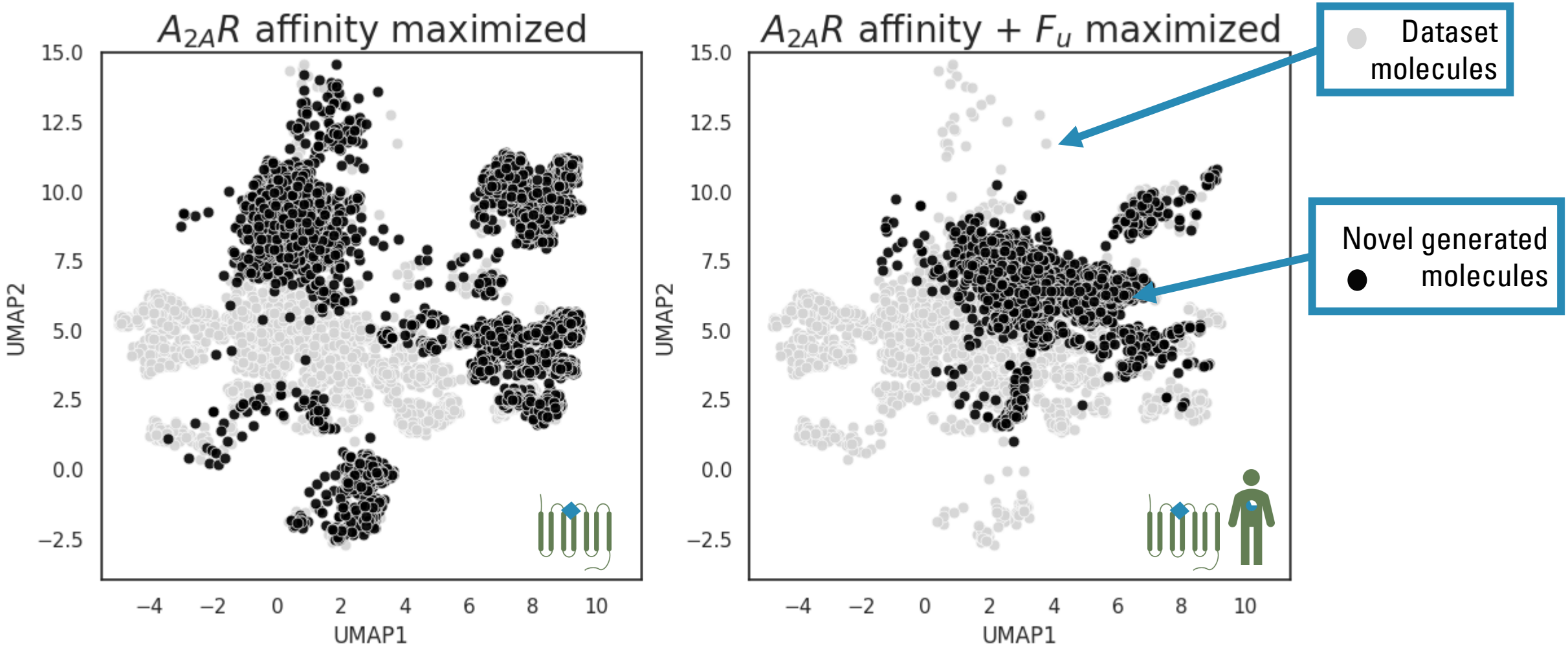
QSPRPRED MODELS PREDICT AFFINITY AND UNBOUND FRACTION



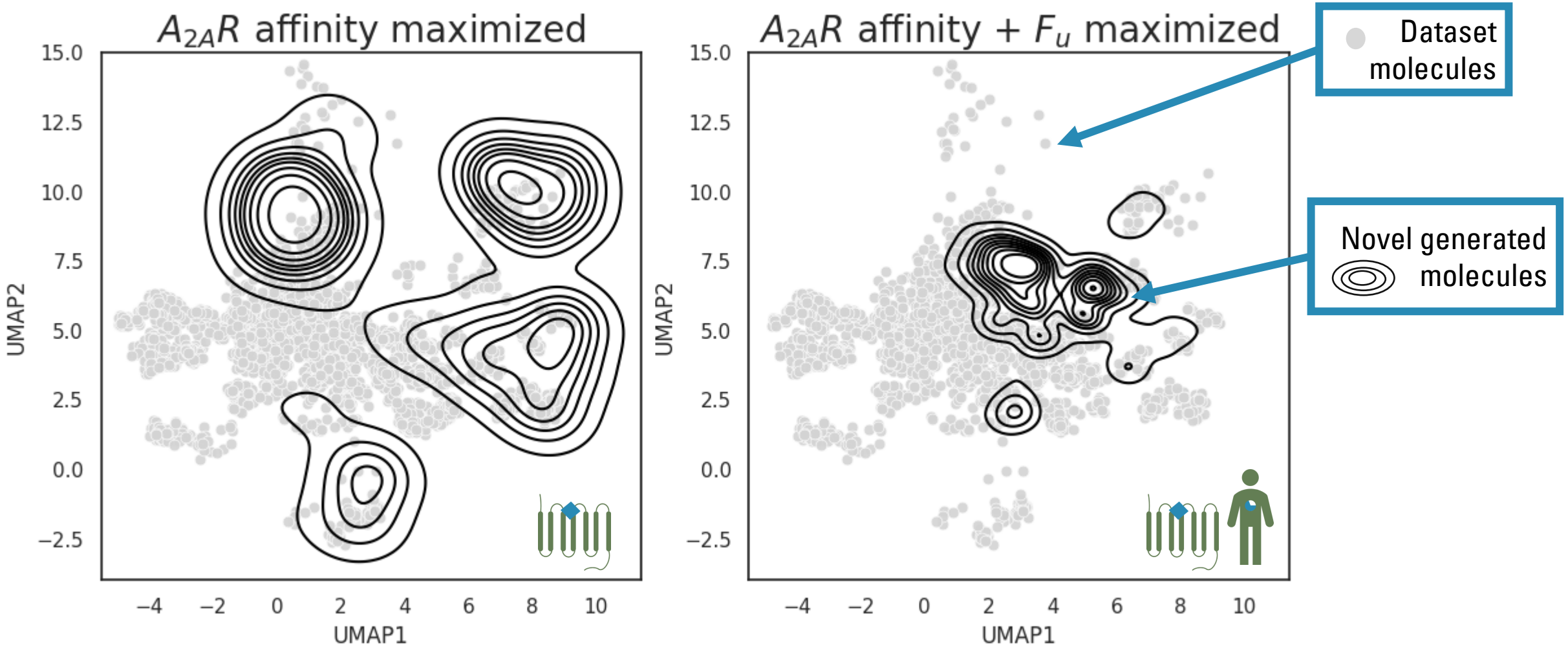


*PK data from Lombardo et al. *Drug Metab. Dispos.* 2018

QSPR MODELS STEER THE GENERATION OF NOVEL MOLECULES TO DIFFERENT AREAS OF CHEMICAL SPACE



QSPR MODELS STEER THE GENERATION OF NOVEL MOLECULES TO DIFFERENT AREAS OF CHEMICAL SPACE



TAKE HOME MESSAGES

QSPRpred is a versatile tool for Quantitative Structure-Property Relationship modelling

Contains extensive data-preprocessing functionality.
Suitable for building single-task, multi-task and proteochemometric models.

QSPRpred is simple but flexible

Has a modular structure that allows for easily adding new functionalities.
Comprehensive tutorials available

QSPRpred is open-source

The code can be found on the Leiden Computational Drug Discovery group Github

ACKNOWLEDGEMENTS

- J.G. Coen van Hasselt, Piet H. van der Graaf, and Gerard J. P. van Westen
- The Quantitative Pharmacology group
- The Computational Drug Discovery group
- The QSPRpred dev team: Martin Šícho, David Araripe, Sohvi Luukkonen, Linde Schoenmaker, Michiel Jespers, Olivier J. M. Béquignon, Marina Gorostiola González, Remco L. van den Broek, Andrius Bernativicius




QSPRPRED DEV TEAM







QUESTIONS?

GitHub

Preprint