Enhancing Drug Discovery with Machine Learning: ADMET Property Modeling

Marcin Kowiel, PhD | April 06, 2024

RYVU THERAPEUTICS | Developing therapeutics at the forefront of oncology

Agenda

01 About Ryvu

02 Drug discovery process

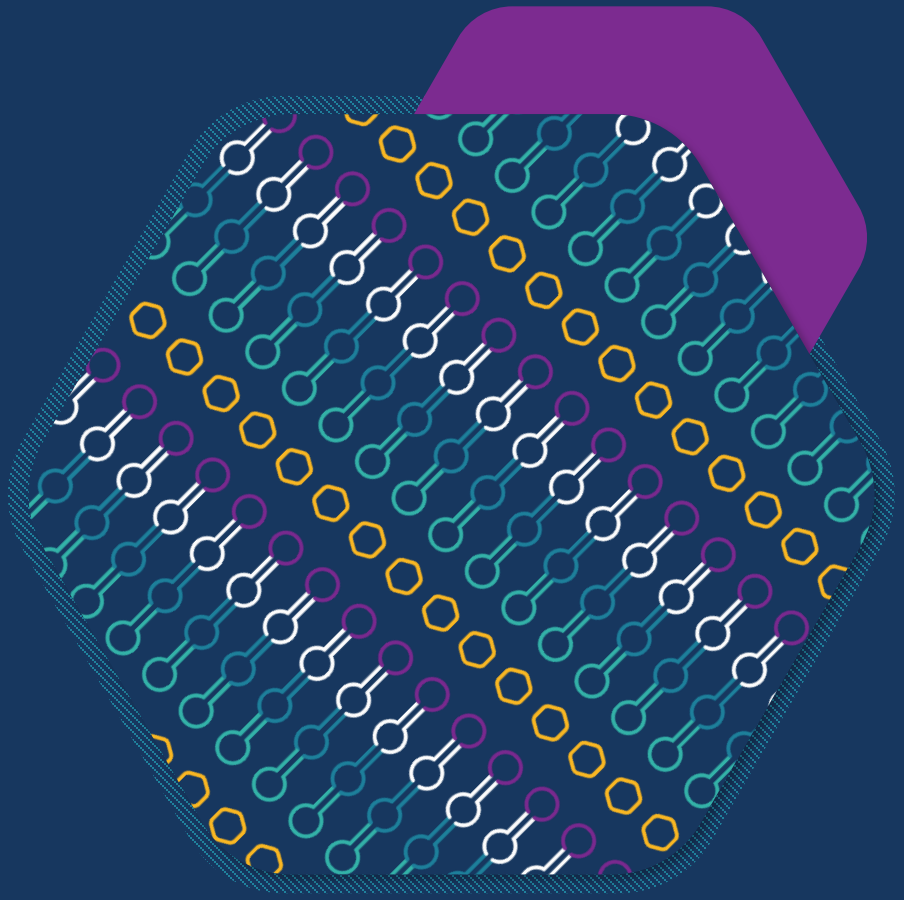
03 AI in drug discovery

04 Property prediction model training pipeline at Ryvu

05 Pipeline stages

06 Summary

Takeaway points



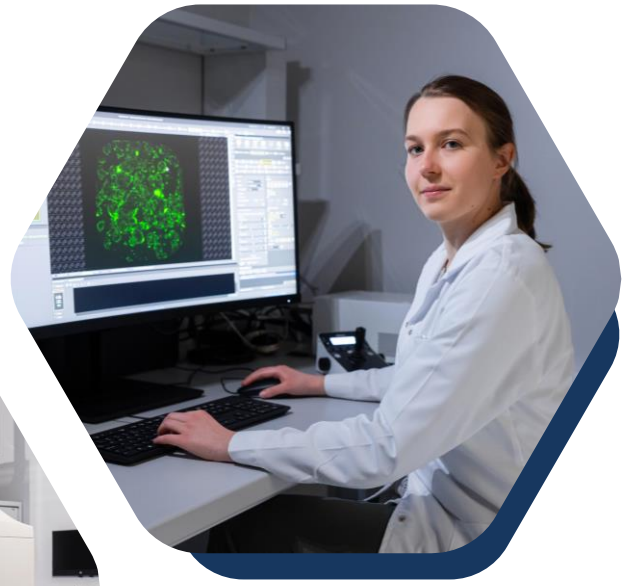
- Ryvu Therapeutics is leading Polish drug discovery company
- AI can improve every stage of the drug discovery process
- Property prediction models can improve lead optimization process
- High quality models can be achieved through combination of data science, data engineering and understanding of the field
- Models’ interpretability is important for medicinal chemists
- It is possible to automate many steps of the property prediction model training pipeline

RYVU THERAPEUTICS

Ryvu Therapeutics – Developing therapeutics at the forefront of oncology

Ryvu Therapeutics is a clinical stage **biopharmaceutical** company developing novel small molecule therapies addressing emerging targets in **precision oncology**.

The company was founded in **2007** and has its headquarters in **Kraków**. It was previously known as Selvita until it was renamed Ryvu Therapeutics after the spinning out of the services segment.



290+
Employees

200
Scientists

20
Machine Learning, Cheminformatics and Bioinformatics Engineers

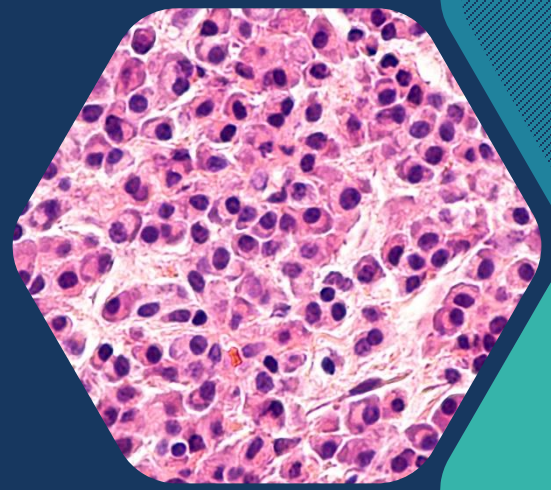
Broad pipeline addressing emerging targets in oncology

PROGRAM	INDICATION	DISCOVERY	PRECLINICAL	PHASE I	PHASE II	PARTNER
CLINICAL PROJECTS						
RVU120 (CDK8/19)	R/R AML/HR-MDS (RIVER-52) (monotheraphy)					LEUKEMIA & LYMPHOMA SOCIETY
	R/R AML (RIVER-81) (combination therapy)					
	Other Hematology (LR-MDS, MF)					
	Solid Tumors					
MEN1703 (SEL24) (PIM/FLT3)	DLBCL					MENARINI
DISCOVERY AND PRECLINICAL PROJECTS						
SYNTHETIC LETHALITY						
PRMT5	SOLID TUMORS					
WRN	SOLID TUMORS					
NOVEL TARGETS	ONCOLOGY					
IMMUNO-ONCOLOGY						
STING & MULTI-TARGET IMMUNE MODULATION COLLABORATION	ONCOLOGY					BIONTECH
STING ADC	ONCOLOGY					EXELIXIS

Study start-up

RYVU THERAPEUTICS

AI in drug discovery



RYUU THERAPEUTICS

The drug discovery process is costly ($1.7B) and time-consuming (10-15 years)

Discovery
4.5 years $650M

Preclinical
1 year $150M

Clinical
7.5 years $950M

- Target discovery
- Hit discovery
- Lead Optimization
- In Vitro
- In Vivo
- Clinical Trials
- Drug Production

100k-1M compounds → 100 compounds → 5 compounds → 1 drug

https://optim.tildacdn.net/tild3332-3535-4761-b562-373639353431/-/format/webp/noroot.png , https://doi.org/10.1016/j.apsb.2022.02.002 , https://www.genengnews.com/insights/what-is-the-real-drug-development-cost-for-very-small-biotech-companies/

Broad application of Machine Learning in Drug Discovery

Target discovery

Protein folding: AlphaFold, ESM2

Protein-Protein interaction prediction

Biological data analysis: genomic, RNAseq

Cell painting: image processing

Knowledge graphs

Hit discovery

Binding pocket detection: diffusion models

Virtual Screening: active learning docking

Fragment growing

Drug repurposing

Lead optimization

ADME/Tox property prediction

Molecule generative models

Synthetic accessibility

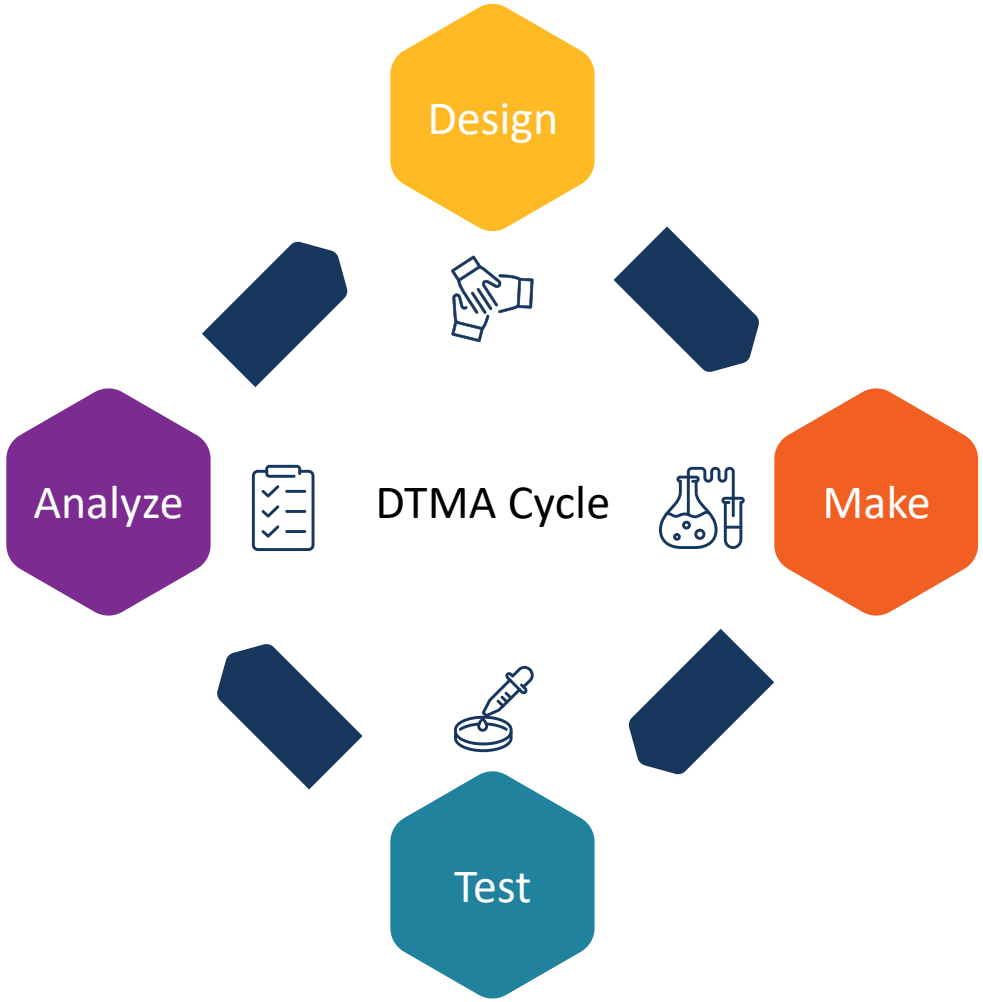
In vitro

Understanding the mechanism of action

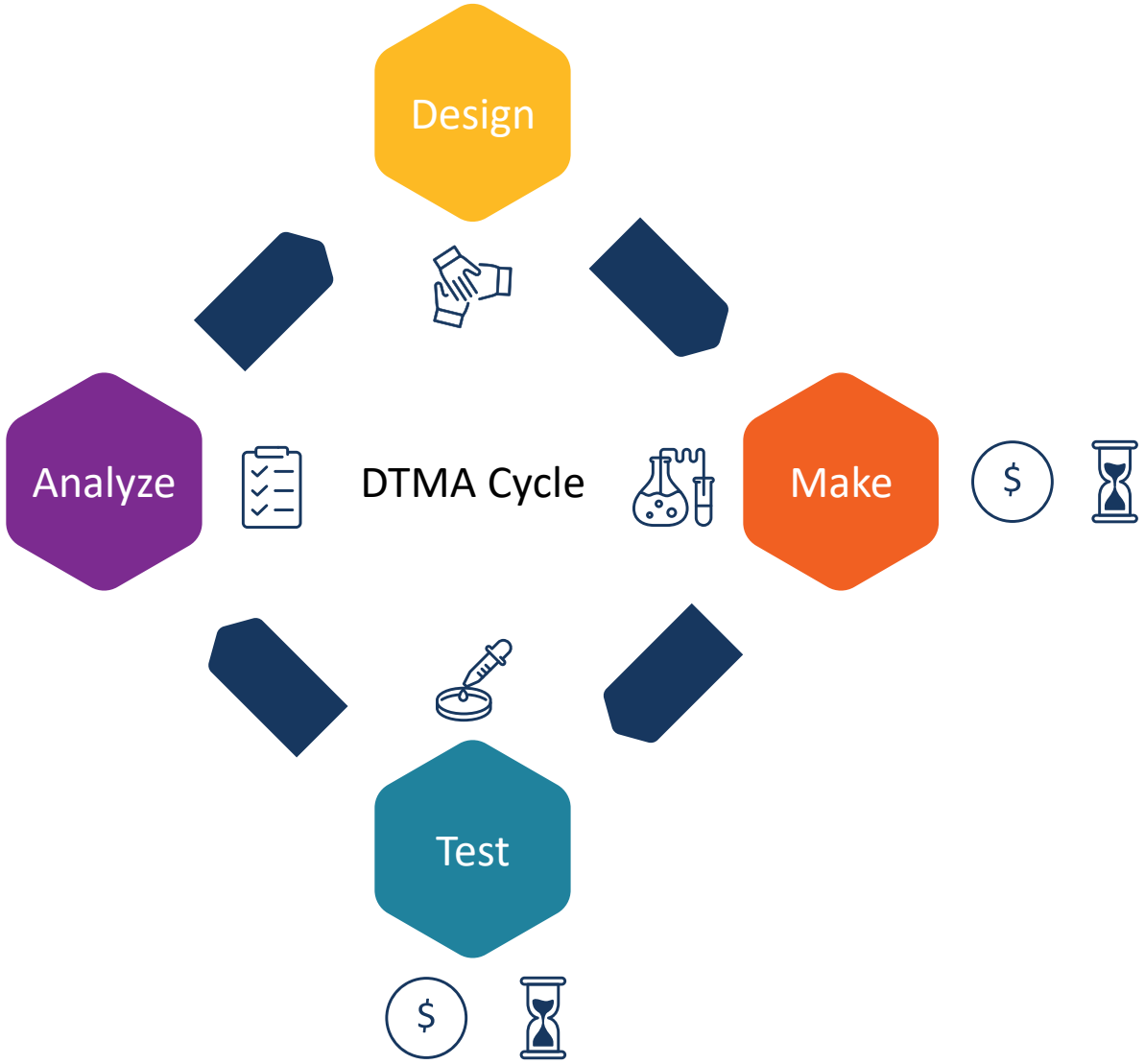
Biomarkers discovery for patient selection

Identification of drug combination therapies

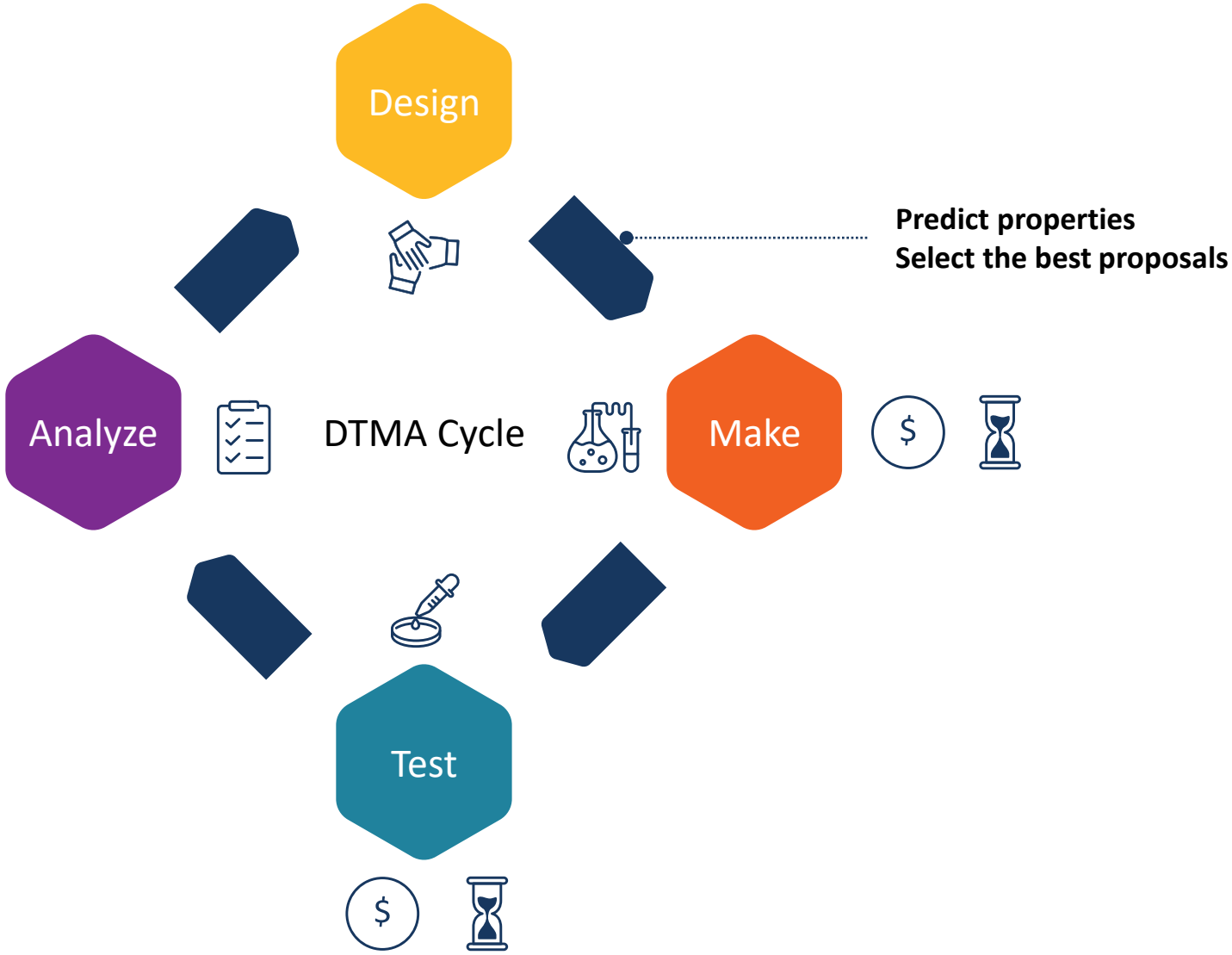
Problem: Rank molecule proposals before synthesis



Problem: Rank molecule proposals before synthesis



Solution: ADMET property prediction models

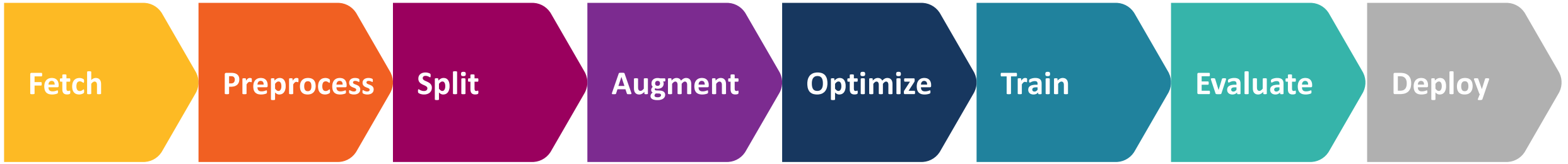


ADMET (absorption, distribution, metabolism, and excretion–toxicity)

Property prediction model training pipeline at Ryvu

RYVU THERAPEUTICS

Typical machine learning model training pipeline



Multiple tools used in Ryvu property prediction model training pipeline

RYVU THERAPEUTICS

Data versioning is a key to experiment reproducibility

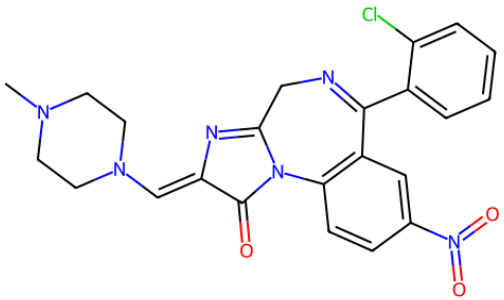
Fetch

Internal data
(train + test)

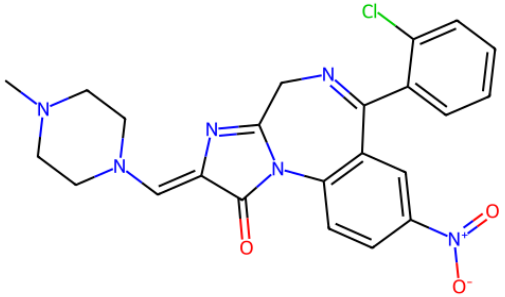
External data
(train only)

Different SMILES but the same compound: Importance of molecule standardization and data cleaning

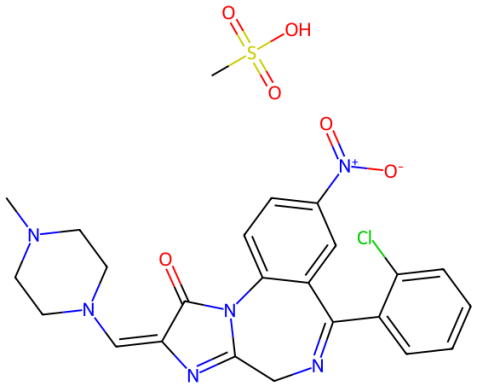
Preprocess



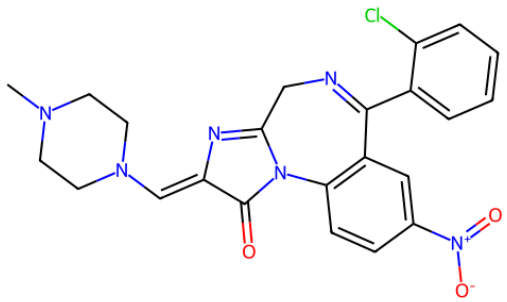
CN1CCN(CC1)\C=C1/N=C2CN=C(C3=CC=CC=C3Cl)C3=C(C=CC(=C3)**[N](=O)=O)**N2C1=O



CN1CCN(CC1)\C=C1/N=C2CN=C(C3=CC=CC=C3Cl)C3=C(C=CC(=C3)**[N+]([O-])=O)**N2C1=O



CS(O)(=O)=O.CN1CCN(CC1)\C=C1/N=C2CN=C(C3=CC=CC=C3Cl)C3=CC(=CC=C3N2C1=O)**[N+]([O-])=O**

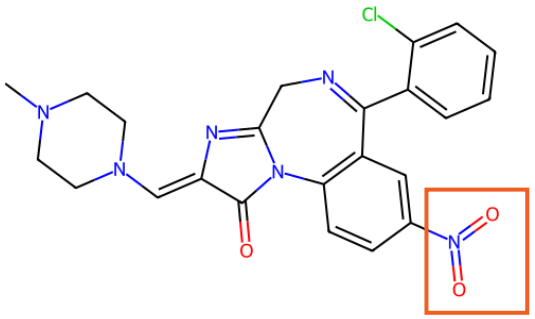


Loprazolam

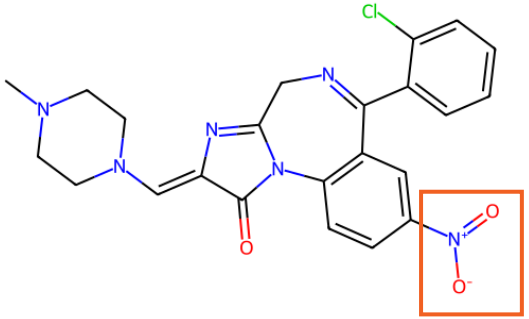
CN1CCN(C=C2N=C3CN=C(c4ccccc4Cl)c4cc([N+](=O)[O-])ccc4N3C2=O)CC1

Different SMILES but the same compound: Importance of molecule standardization and data cleaning

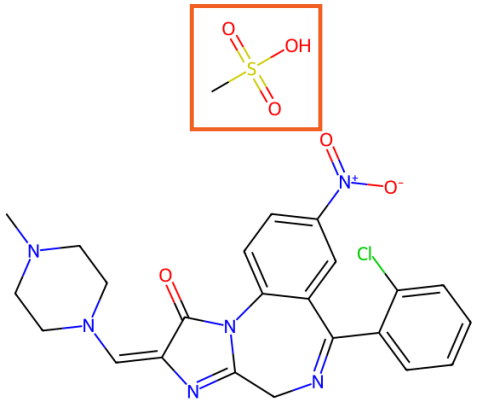
Preprocess



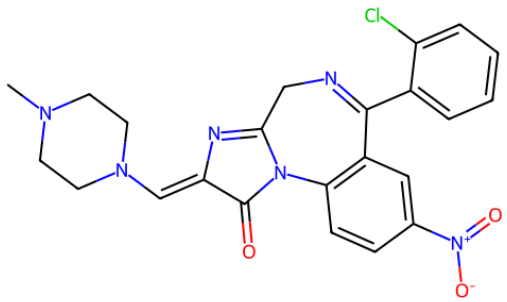
CN1CCN(CC1)\C=C1/N=C2CN=C(C3=CC=CC=C3Cl)C3=C(C=CC(=C3)**[N](=O)=O)**N2C1=O



CN1CCN(CC1)\C=C1/N=C2CN=C(C3=CC=CC=C3Cl)C3=C(C=CC(=C3)**[N+]([O-])=O)**N2C1=O



CS(O)(=O)=O.CN1CCN(CC1)\C=C1/N=C2CN=C(C3=CC=CC=C3Cl)C3=CC(=CC=C3N2C1=O)**[N+]([O-])=O**



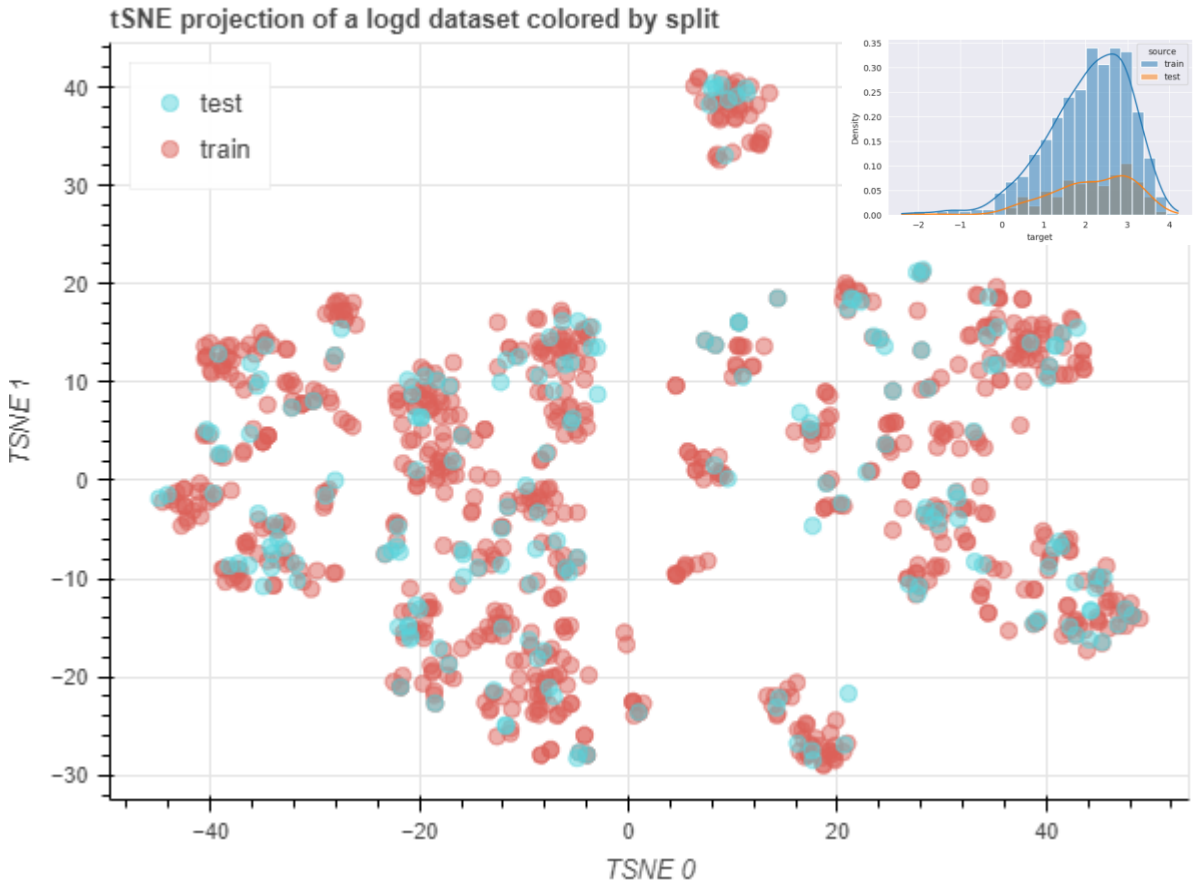
Loprazolam

CN1CCN(C=C2N=C3CN=C(c4ccccc4Cl)c4cc([N+](=O)[O-])ccc4N3C2=O)CC1

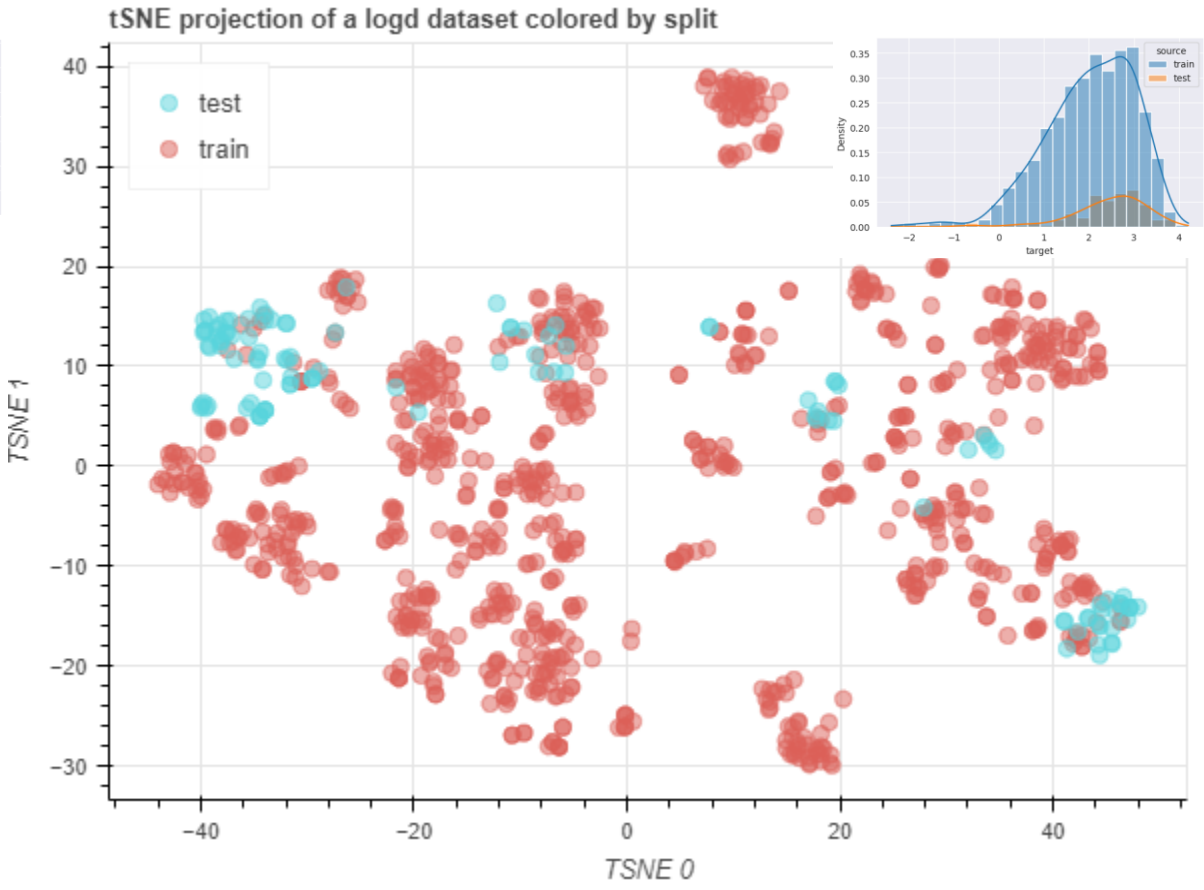
Use scaffold or time split to detect problems with generalization and have reliable estimation of model quality

Split

Random split

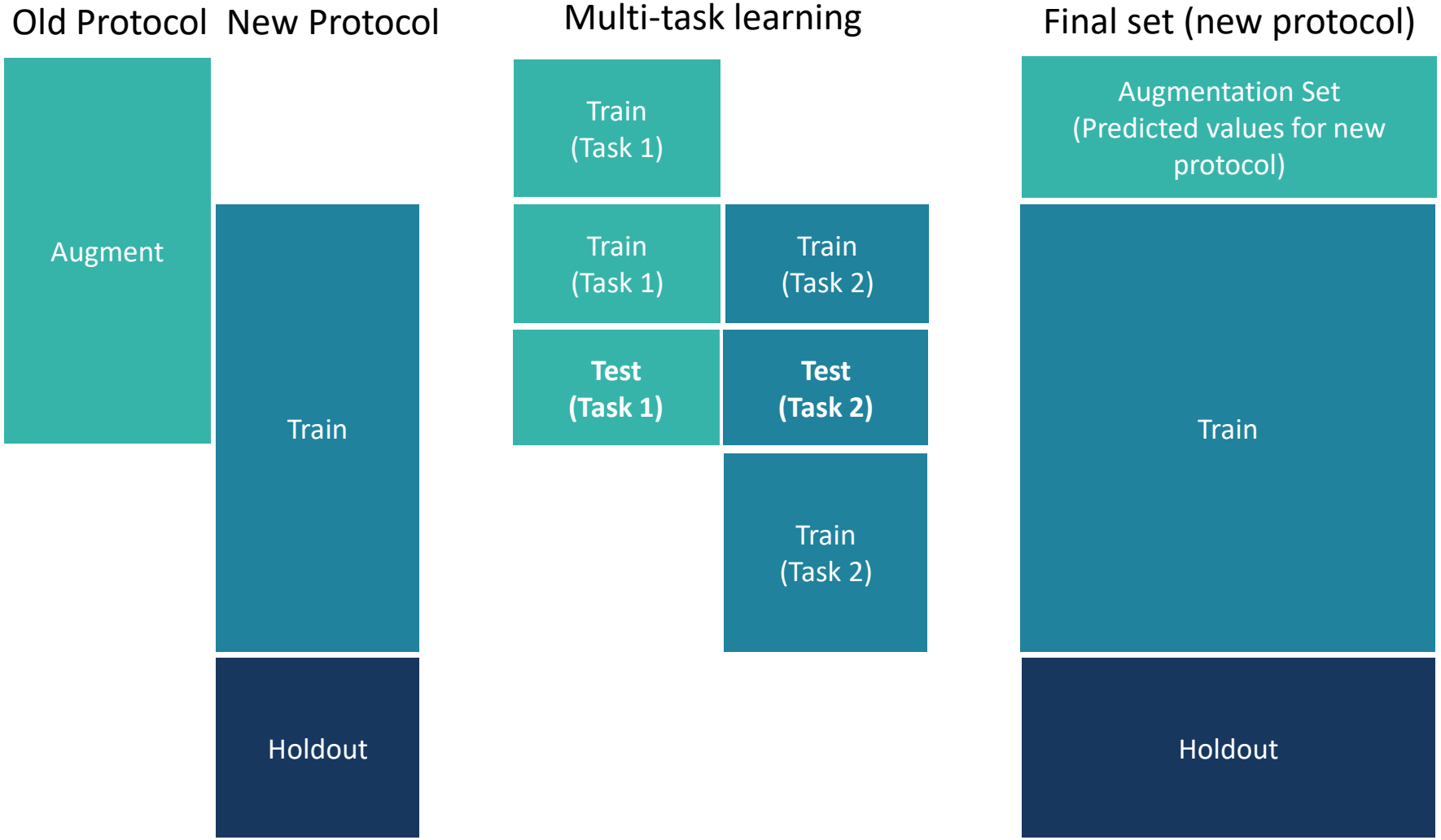


Time split (similar to scaffold split)



Multi-task learning or transfer learning can enhance model quality

Augment

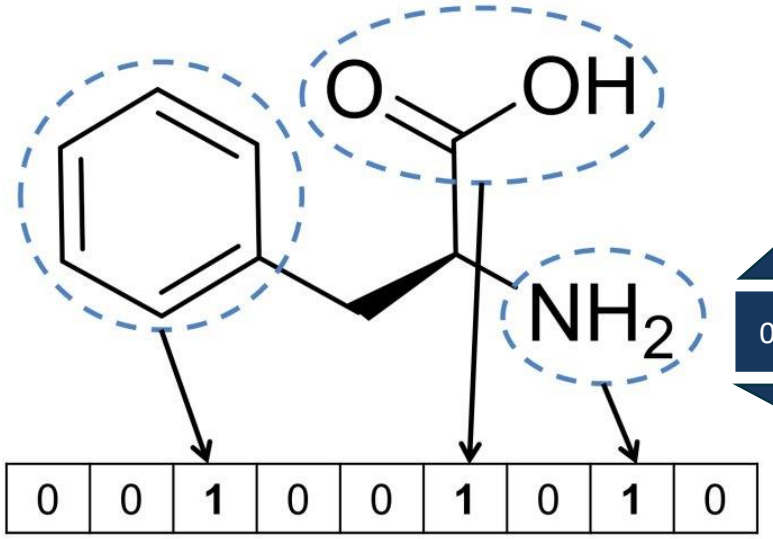


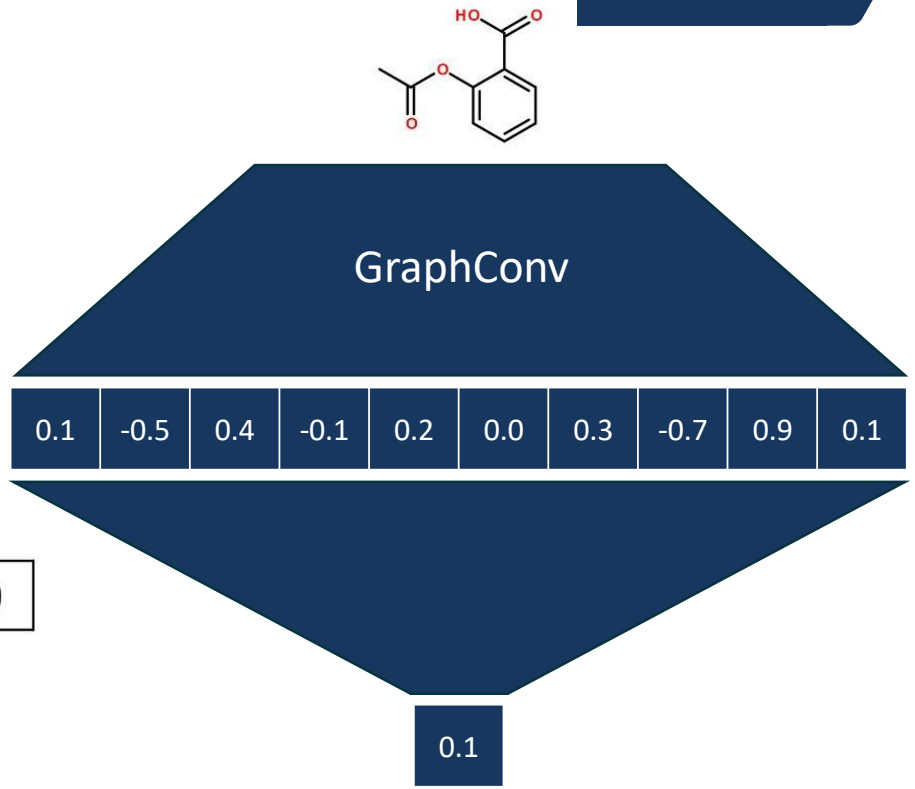
Needs special handling during CV

Find the optimal feature generation technique

Optimize

Name	Description
MW	Molecular weight
nHDon	Number of donor atoms for hydrogen bond (HB)
nHAcc	Number of acceptor atoms for HB
SA	Total surface area
TPSA	Topological polar surface area
nSK	Number of non-H atoms
nsp3	Number of sp^3 hybridized carbon atoms
RBN	Number of rotatable bonds
ARR	Aromatic ratio
cLogP	Calculated partition coefficient between octanol and water
nAR	Number of aromatic rings
Fsp^3	Fraction of sp^3 carbon atoms





- **RDKit** descriptors
- …
- **Mordred** descriptors

- **MACCS** keys
- **ECFP4** fingerprints

  Extended-Connectivity Circular Fingerprints
- …
- **PubChem** fingerprints

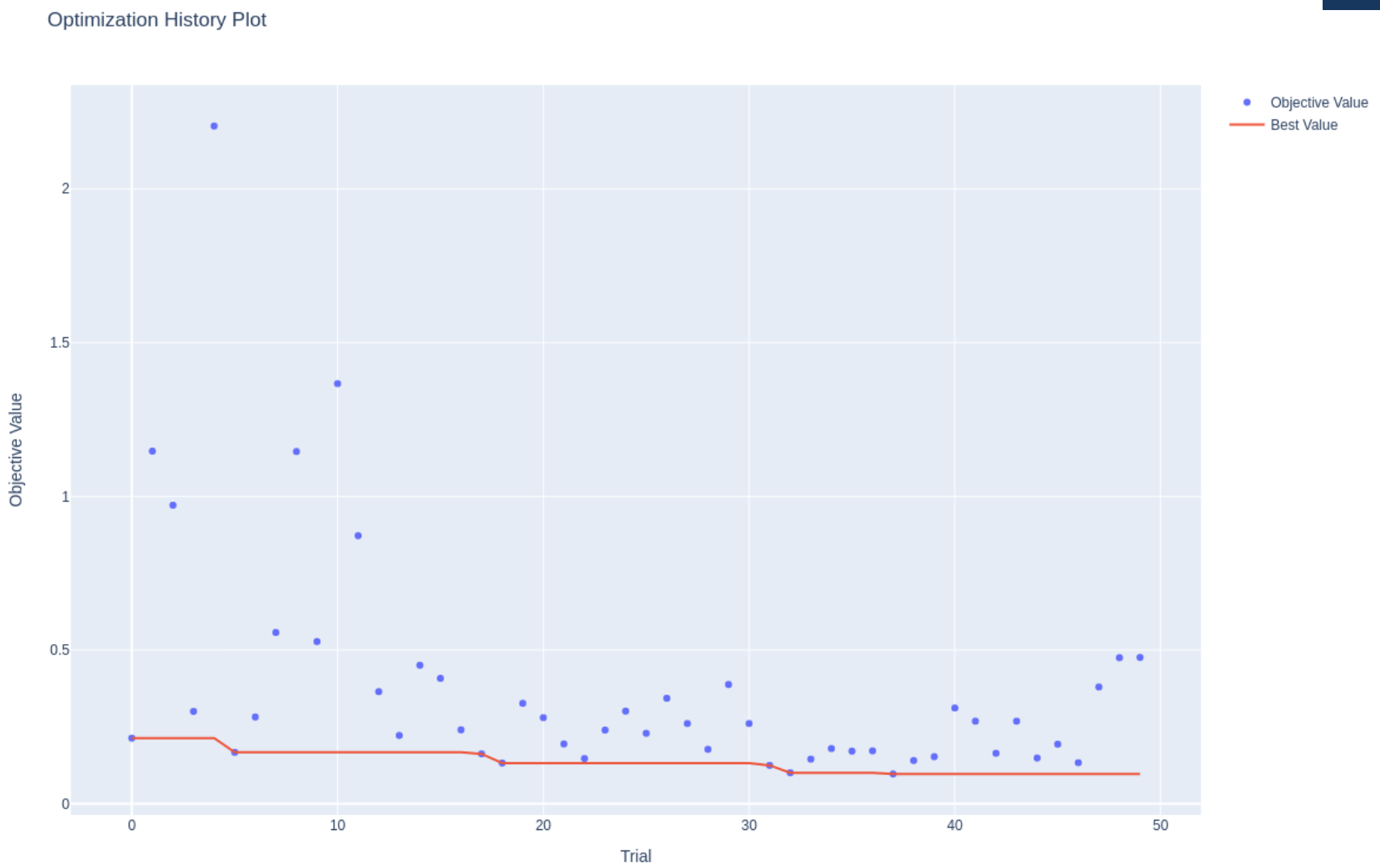
- Molecule **Embeddings** (trainable)
- …
- Sequence to sequence autoencoders

[1]https://www.researchgate.net/figure/In-substructure-key-based-fingerprints-bits-are-set-according-to-the-substructures-that_fig10_315513438

RYVU THERAPEUTICS

Avoid grid search – use Optuna instead

Optimize



Model zoo:
It is important to test various model architectures

Train

Random Forest

XGBoost

LightGBM

Graph Convolutional Networks

(Message Passing)

Gaussian Process Regressor

PyG

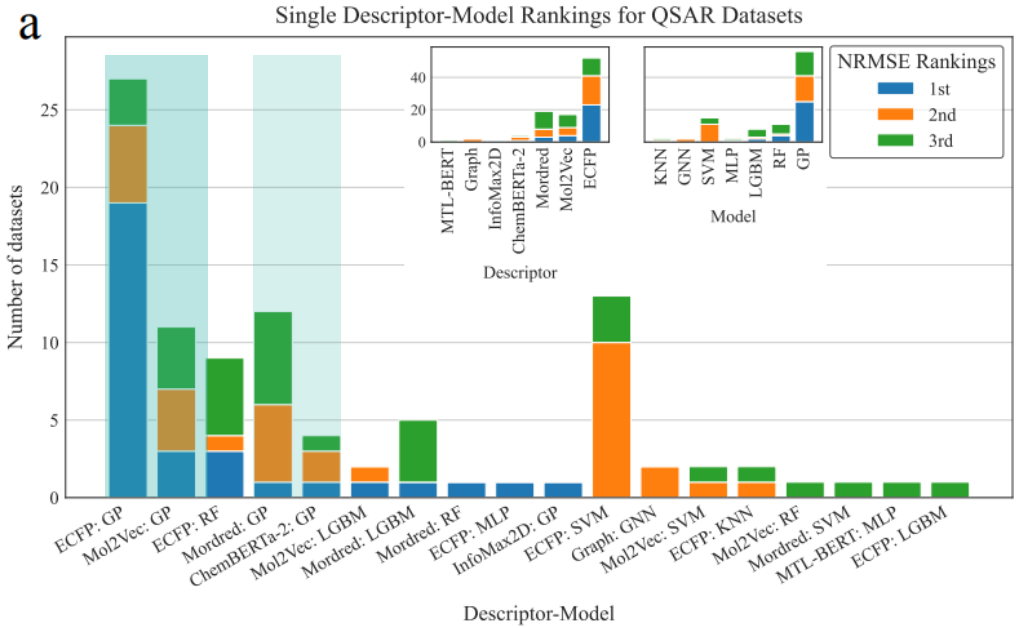
GPyTorch

RYVU THERAPEUTICS

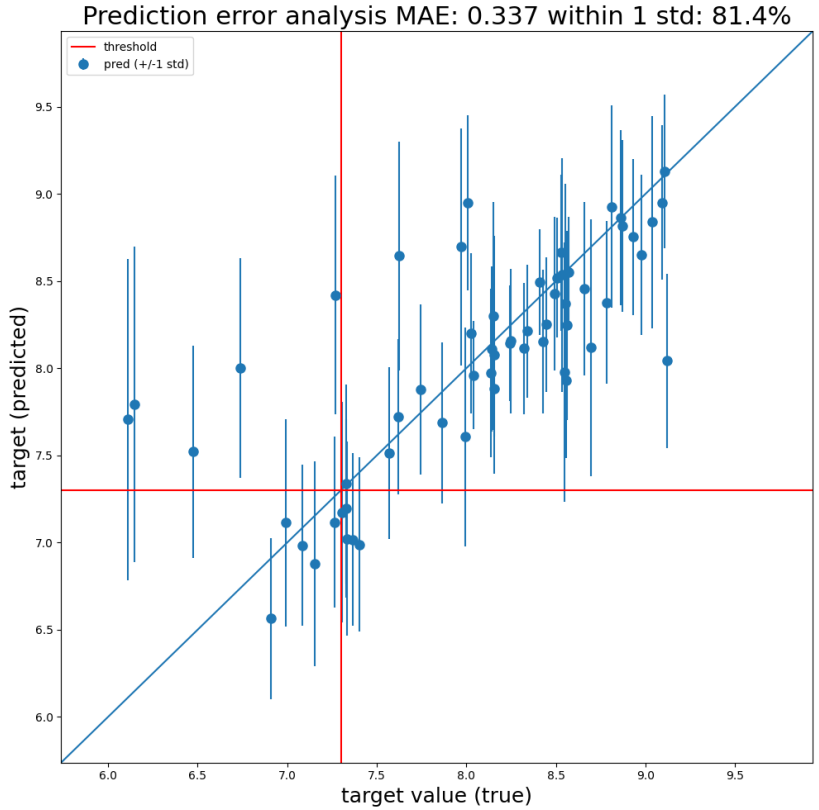
Gaussian process is effective for QSAR modeling

Train

- Gaussian Process Regressors are effective in Quantitative Structure-Activity Relationships (QSAR) modeling[1]
- Gaussian processes have embedded prediction **uncertainty estimation**



Single descriptor-model rankings based on NRMSE[1]



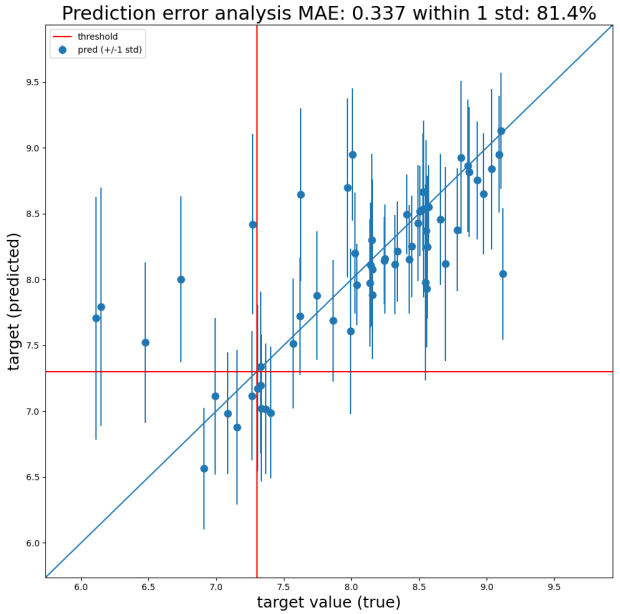
Gaussian processes have embedded prediction uncertainty estimation

[1] https://browse.arxiv.org/pdf/2309.17161.pdf

Pick the right metrics for the problem

Evaluate

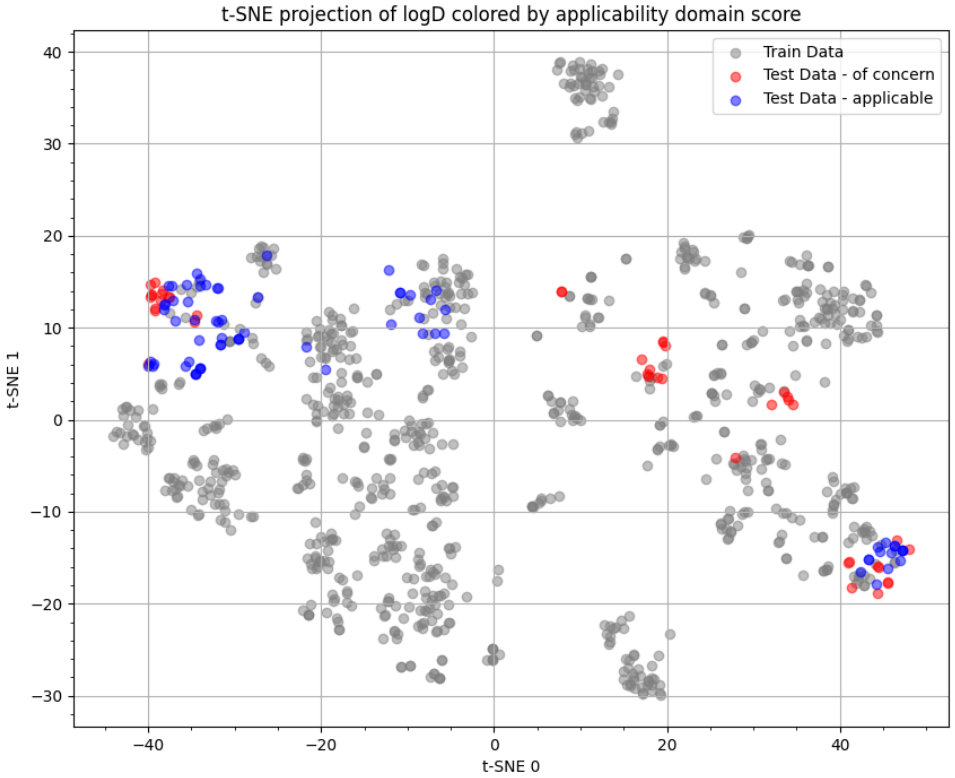
Binary classification	Ranking	Regression
Simple interpretation	Prioritization	Exact value prediction
ROC AUC	Spearman correlation	MSE
F1	NDCG	MAE
Accuracy		R2
Precision		Pearson Correlation
Recall		



Each user has different need – focus on regression

Estimation inference limits via applicability domain analysis

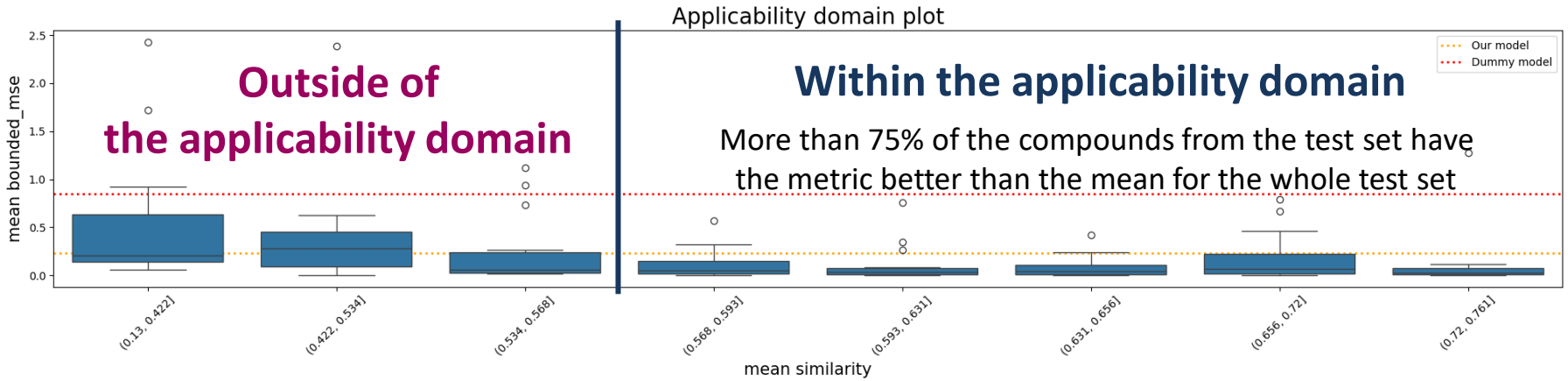
Evaluate



Problem

Chemists would like to know the model inference limitations.

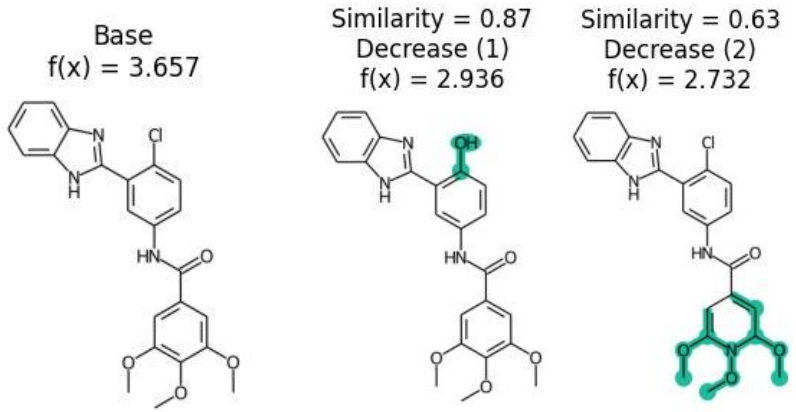
Often, models can't generalize to different chemical series, further away from the training set.



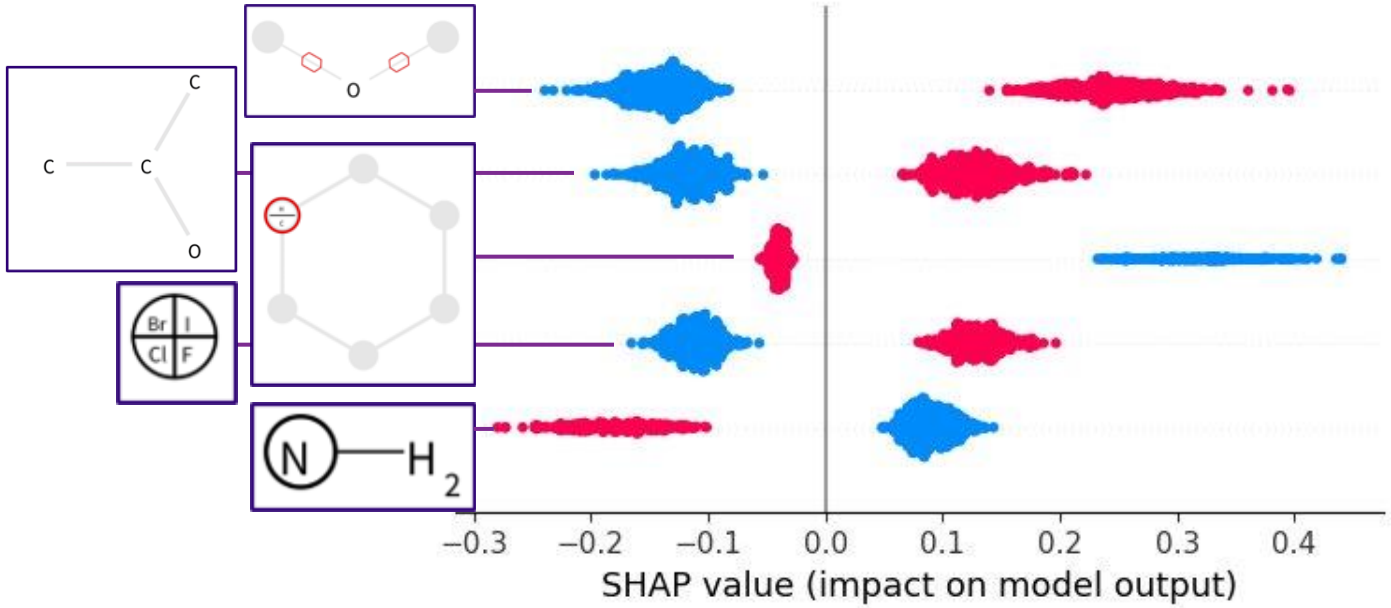
Model explainability and interpretability is important for chemists

Evaluate

Exmol



Shapley values



Explainability methods (SHAP, Lime, Exmol) can give hints about molecule changes to improve the property

https://ur-whitelab.github.io/exmol/

Save time with deployment and model updates automation

Deploy

- **Regular model updates**
  Ideally, once per DTMA cycle
- **Automation**
  Deployment of multiple models can take time
- **Integration**
  with software tools that chemists use for daily work

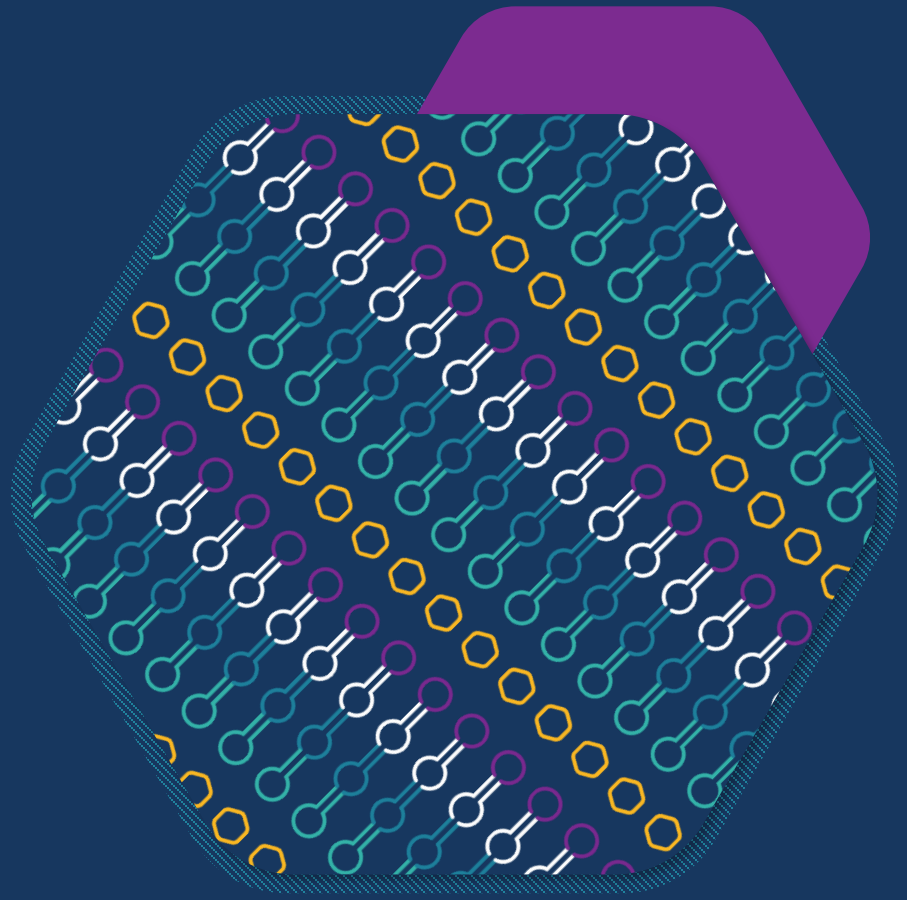
FastAPI

Dotmatics

Regular model retraining is a key to keep high model quality

RYVU THERAPEUTICS

Summary



- Ryvu Therapeutics is leading Polish drug discovery company
- AI can improve every stage of the drug discovery process
- Property prediction models can improve lead optimization process
- High quality models can be achieved through combination of data science, data engineering and understanding of the field
- Models’ interpretability is important for medicinal chemists
- It is possible to automate many steps of the property prediction model training pipeline

RYVU THERAPEUTICS

Thank you!

Marcin Kowiel

Team Lead Machine Learning Engineer

Data Science and AI Platforms

marcin.kowiel@ryvu.com

Open positions

https://ryvu.com/careers/job-offers/

Data Engineeing Intern

Data Scientist/AI Research Intern

Computer Aided Drug Design Intern

Bioinformatics Intern

Computer Aided Drug Design Senior Scientist

Senior IT Administrator

RYVU THERAPEUTICS