NIBR Translational Medicine
PKS M&S Data Science | GDC, CADD

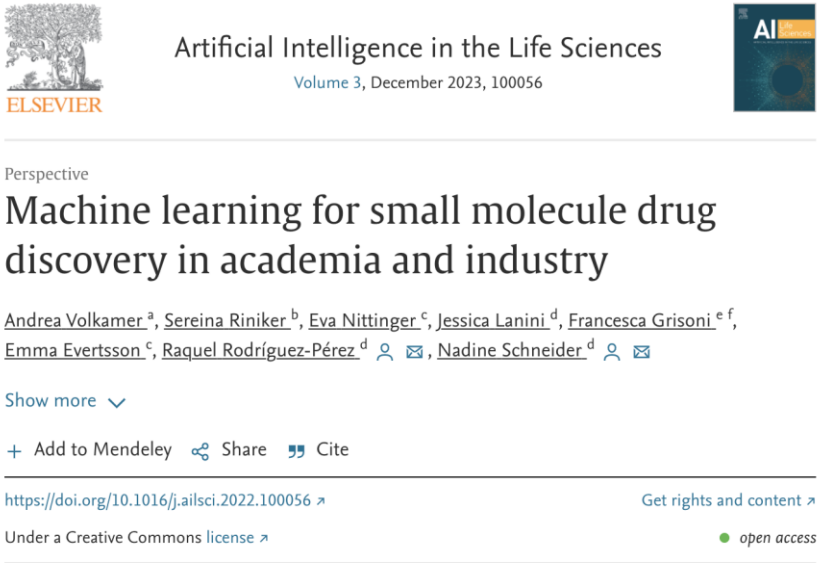
AI advancing drug discovery research in pharma industry and academia

Dr. Jessica Lanini
ChemTalks, 25 September 2024

NOVARTIS | Reimagining Medicine

Introduction

- We published a perspective with colleagues from academia and industry
  - Raquel Rodriguez Perez (Novartis)
  - Andrea Volkamer (Saarland University)
  - Sereina Riniker (ETH Zurich)
  - Eva Nittinger (AstraZeneca)
  - Francesca Grisoni (Eindhoven University of Technology)
  - Emma Evertsson (AstraZeneca)
  - Nadine Schneider (Novartis)
- In this talk, we will summarize the main commonalities, differences, and opportunities for collaboration between industry and academia
- Examples will **mostly focus** on industry and, more specifically, our work at Novartis



ELSEVIER

Artificial Intelligence in the Life Sciences

Volume 3, December 2023, 100056

Perspective

Machine learning for small molecule drug discovery in academia and industry

Andrea Volkamer [a], Sereina Riniker [b], Eva Nittinger [c], Jessica Lanini [d], Francesca Grisoni [e f], Emma Evertsson [c], Raquel Rodríguez-Pérez [d], Nadine Schneider [d]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.ailsci.2022.100056

Get rights and content

Under a Creative Commons license

open access

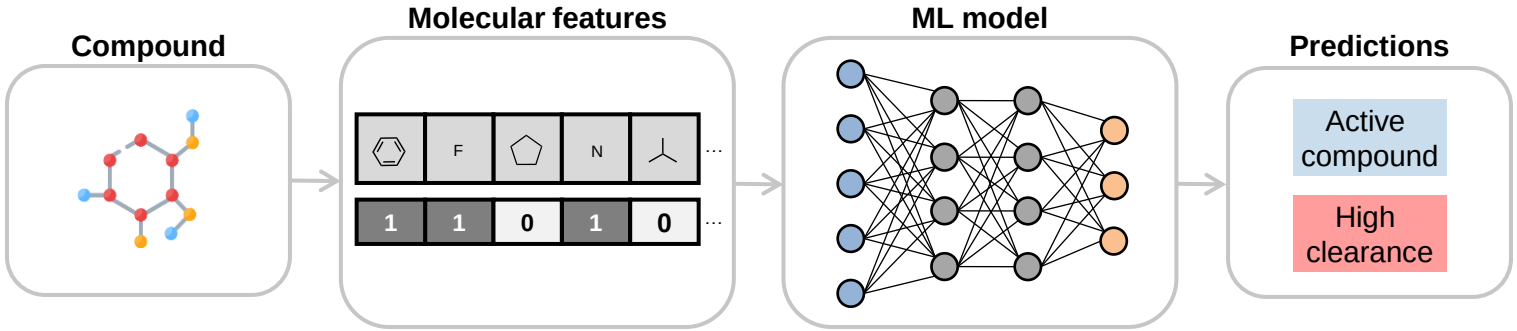
NOVARTIS | Reimagining Medicine

Machine learning in drug discovery

- **Machine learning (ML) is applied at different stages of drug discovery**
- Exemplary ML applications:
  - Predict the suitability of targets for drug discovery
  - Virtual screening of compounds for hit identification
  - Discovery of cell-specific biomarkers
  - Analyze gene signatures for patient sub-groups
  - Pathology image processing

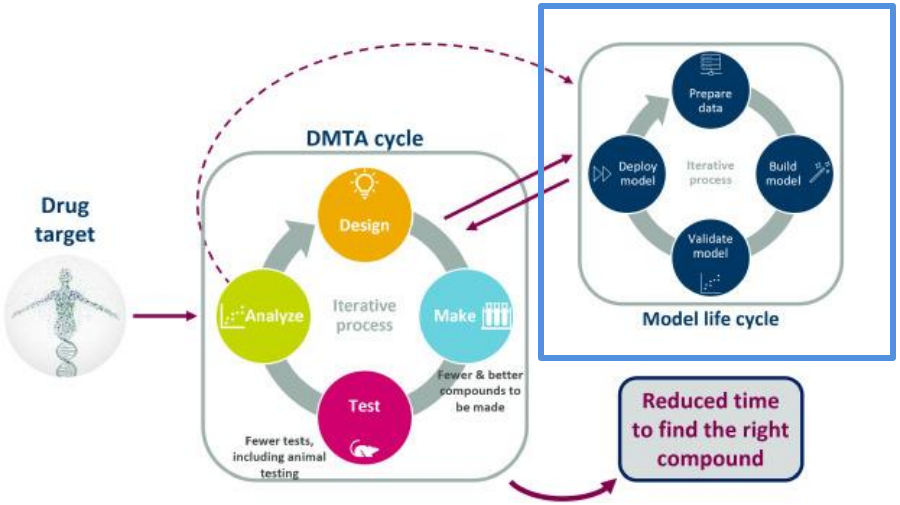
ML for compound property predictions

- ML models can relate molecular structures to compound properties
  - **Quantitative structure-property relationship (QSPR) modeling**
- Compounds can be represented numerically, e.g. descriptors, fingerprints, graph neural networks
- ML algorithms can be used to find structural or chemical patterns that correlate with specific compound properties, e.g. activity against the target of interest, clearance



QSPR models in drug discovery

- **Quantitative structure-property relationship (QSPR) modeling** is a necessary component of numerous drug discovery projects
- QSPR models are usually implemented in a result-oriented fashion to **find the most promising drug candidates**
- ML is used to make better decisions faster and to accelerate the **design-make-test-analyze (DMTA) cycle** of novel molecular entities



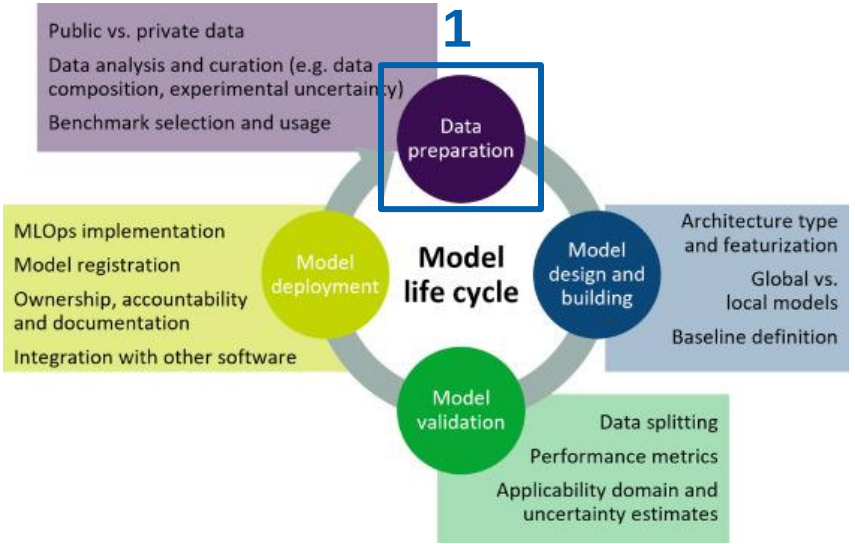
Volkamer et al. 2023 AILSCI, 3, 100056

NOVARTIS | Reimagining Medicine

Model life cycle

- Realizing actionable predictive models in drug discovery can be summarized as an iterative cycle of 4 steps:

1. **Training data preparation**: requires understanding the experimental data for reliable curation
2. **Model design and building**: including steps such as architecture definition and hyperparameter tuning
3. **Model validation**: performance estimation, which is crucial for prospective model use
4. **Model deployment**: the model is made available to users



Volkamer et al. 2023 AILSCI, 3, 100056

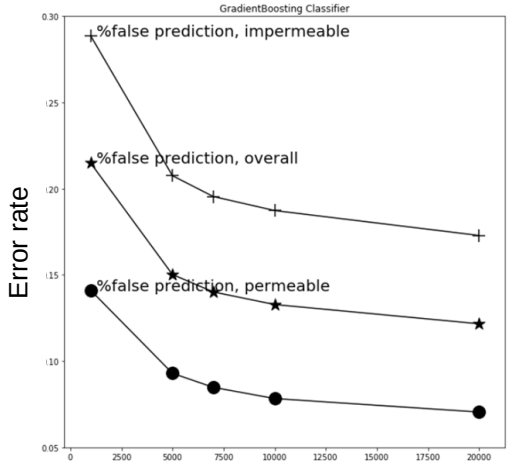
NOVARTIS | Reimagining Medicine

1. Training data

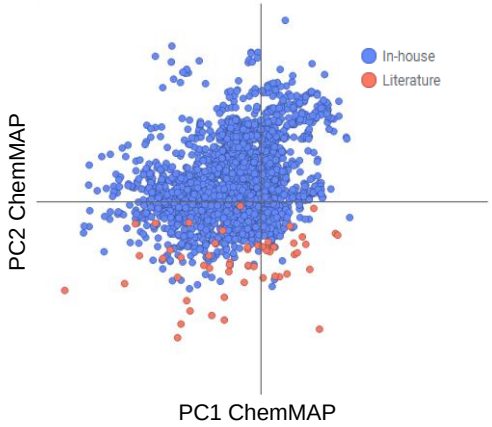
- Model quality depends on the **experimental data used for training**, e.g.
  - **Size:**
    How many compounds (cpds)?
  - **Chemical space:**
    What is the coverage?
  - **Property space:**
    What is the dynamic range?
  - **Diversity:**
    How biased is the data set?
  - **Errors:**
    How noisy is the data set?

Example: Permeability model (LE-MDCK)

In-house models were built with **different sizes**, and models with <7k cpds showed low performance



Chemical space of the literature-based model* did not cover in-house data



* doi:10.1021/js9803205
ChemMAP: doi:10.1016/j.drudis.2011.07.003

Training data: Academia vs. Industry

- Data availability in the public domain has dramatically increased thanks to major databases such as ChEMBL or PubChem
- However, in-house (proprietary) data sets are typically larger

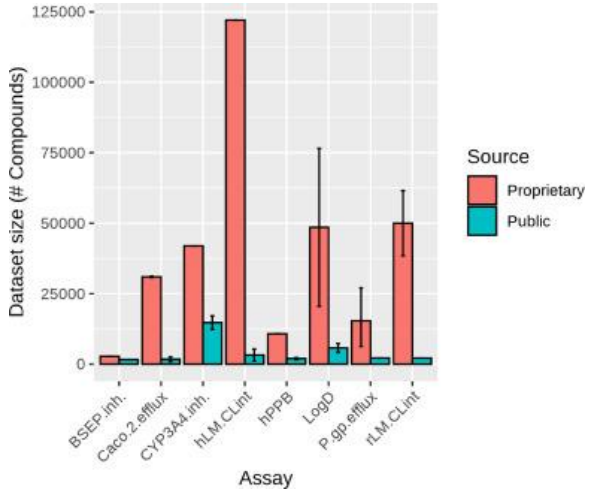
Academia

- More heterogeneous, imbalanced, and smaller data sets
- Mostly single measurement per compound
- Data are usually available for free sharing and re-utilization

Industry

- Larger, more homogenous data
- Access to assay protocols, which sometimes change over time
- Often multiple repeated measurements available
- Project-specific data sets may be small and biased

Size of ADME data sets used for modelling from in-house (Bayer, Novartis, Merck or Boehringer Ingelheim) and public sources



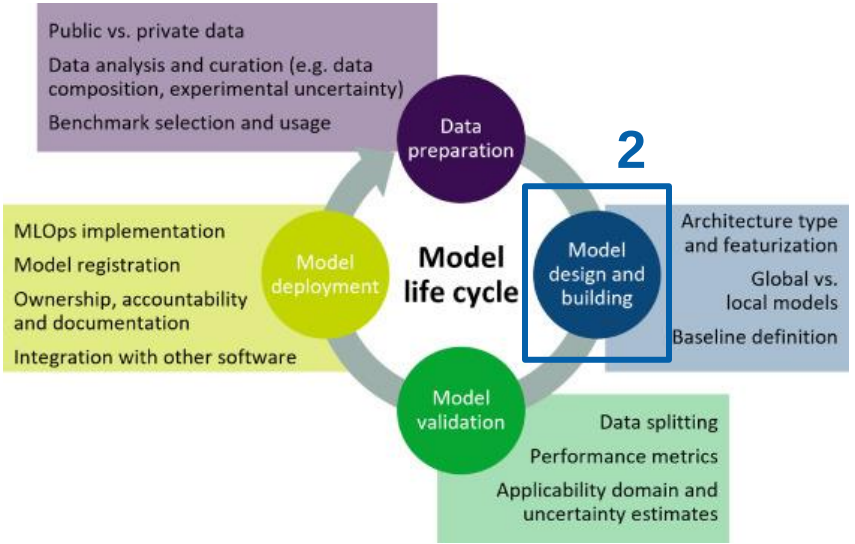
Volkamer et al. 2023 AILSCI, 3, 100056

NOVARTIS | Reimagining Medicine

Model life cycle

- Realizing actionable predictive models in drug discovery can be summarized as an iterative cycle of 4 steps:

1. **Training data preparation**: requires understanding the experimental data for reliable curation
2. **Model design and building**: including steps such as architecture definition and hyperparameter tuning
3. **Model validation**: performance estimation, which is crucial for prospective model use
4. **Model deployment**: the model is made available to users



Volkamer et al. 2023 AILSCI, 3, 100056

2. Model design and building

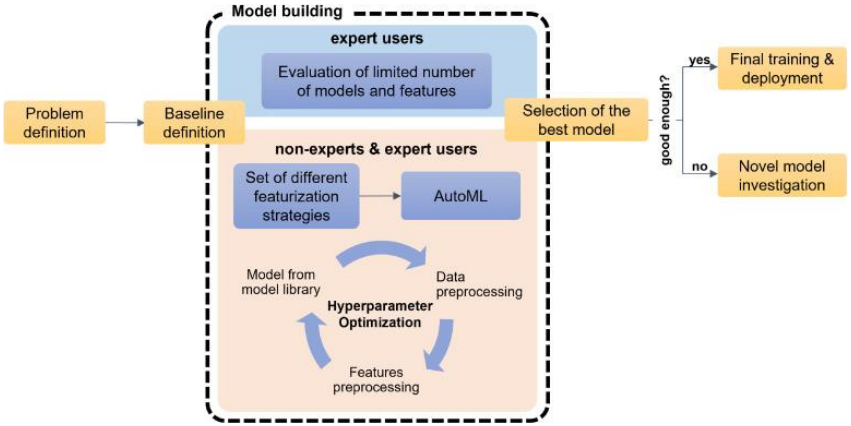
- Model design starts with the **definition of the problem and prediction task**
- Model design is greatly affected by the applications, which are different in academia and industry even for the same QSAR/QSPR field

Academia

- Towards theory and method development (improved performance, new strategies or applications)
- Develop new tools or improve understanding

Industry

- Prospective usage to assist in drug design, e.g. experiment selection, compound prioritization
- Practical applications need to be considered during model design

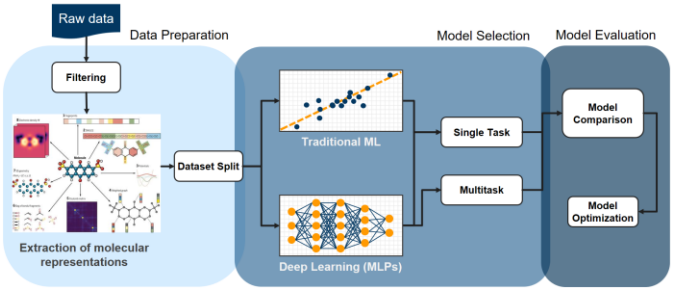


Volkamer et al. 2023 AILSCI, 3, 100056

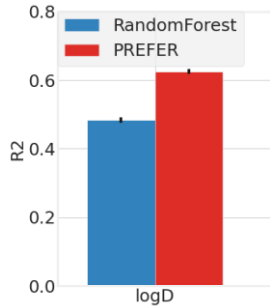
NOVARTIS | Reimagining Medicine

Standardizing model building and benchmarking – the PREFER framework[1]

- The PREFER framework is written in Python and based on AutoSklearn[2]
- **Extensive comparison between different molecular representations and ML models**
  - Features include fingerprints, 2D descriptors, and data-driven representations (CDDD[3], MOLER[4])
  - Models include single-task and multi task learning with hyperparameters' optimization
- **PREFER models automatically give you an ensemble of models that provide strong baselines**



***LogD prediction**:*
R2 for baseline random forest (RF + Fingerprints) and best PREFER model



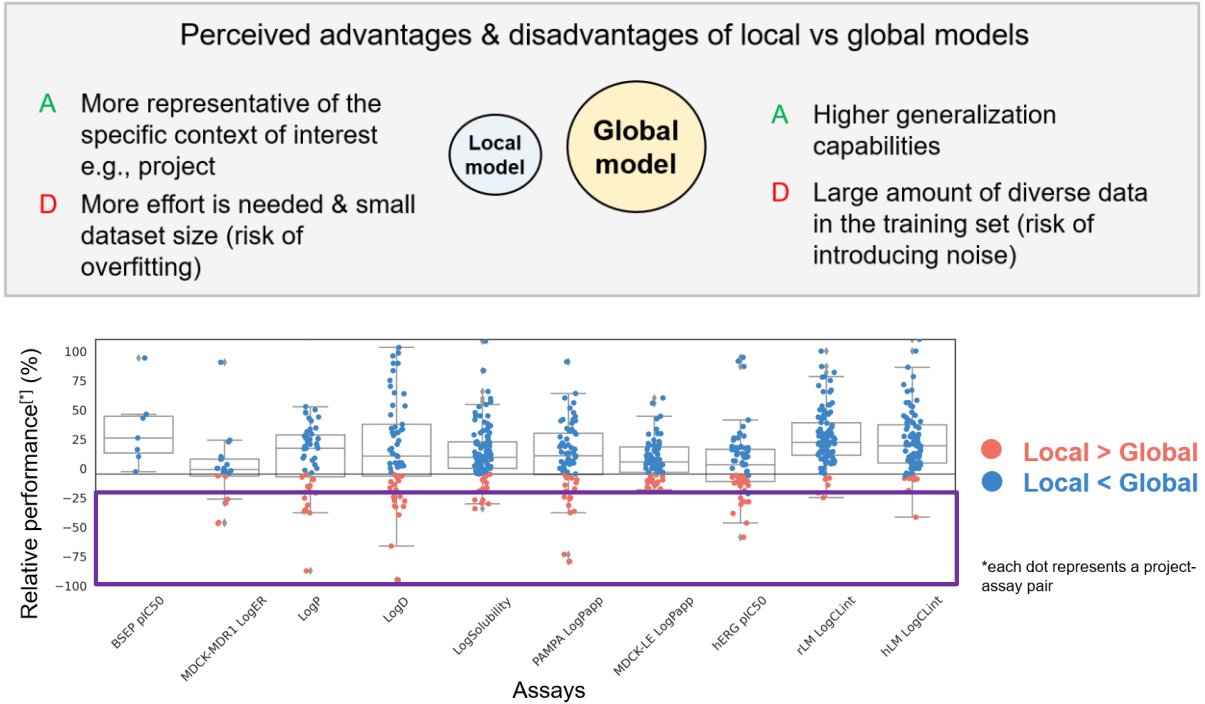
https://github.com/rdkit/PREFER

[1] Lanini, J. et al. *Submitted, 2023*
[2] Feurer, Matthias, et al. Advances in neural information processing systems, 2015
[3] Winter R et al. Chemical science, 2019
[4] Maziarz, Krzysztof, et al. International Conference on Machine Learning. 2021.

NOVARTIS | Reimagining Medicine

Local vs. Global models: What data shall we use for modeling?

- **Global models perform better than the local models**
- Only **7%** of the local models present an improvement > 20% over the global models



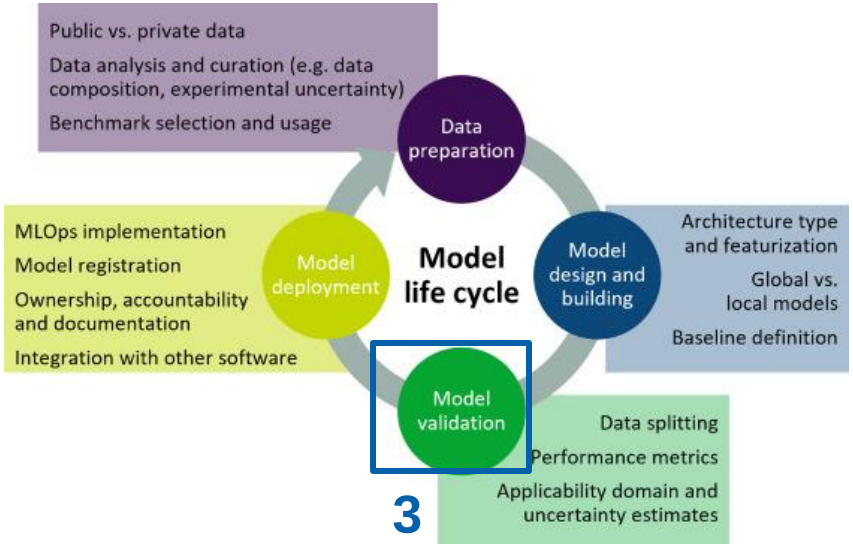
Di Lascio et al. 2023, Mol Pharm, 20, 3, 1758–1767

NOVARTIS | Reimagining Medicine

Model life cycle

- Realizing actionable predictive models in drug discovery can be summarized as an iterative cycle of 4 steps:

1. **Training data preparation**: requires understanding the experimental data for reliable curation
2. **Model design and building**: including steps such as architecture definition and hyperparameter tuning
3. **Model validation**: performance estimation, which is crucial for prospective model use
4. **Model deployment**: the model is made available to users



Volkamer et al. 2023 AILSCI, 3, 100056

3. Model validation

- Model validation is essential and focuses on different aspects in academia and industry
- Model generalizability can only be estimated with **proper data splitting procedures and metrics**

Academia

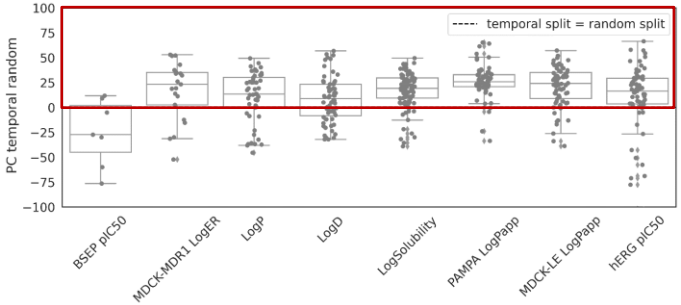
- Random or cluster-based splits are used in the absence of temporal information
- Standardized benchmarking platforms

Industry

- Time-split possible and should be the gold standard
- Application of the model dictates the required level of quality and robustness

Example: Time-split vs. Random split

Percentage change (PC) of prediction error between temporal & random splits: More optimistic results were obtained with random split



Higher errors with time-split

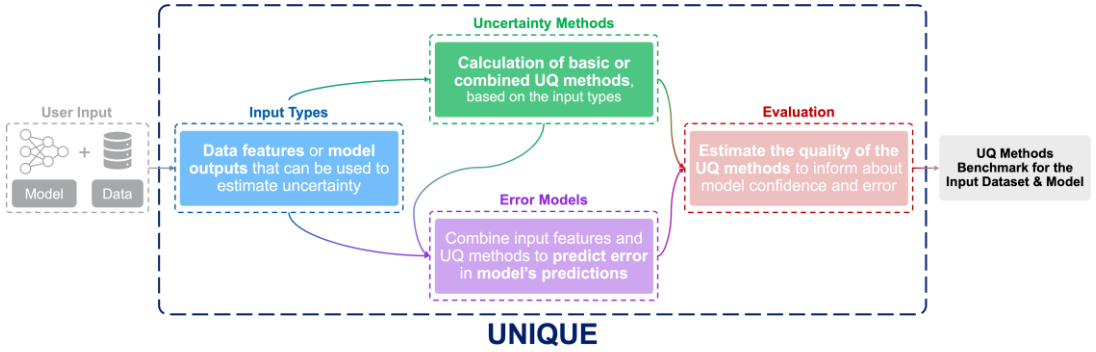
Di Lascio et al. 2023, Mol Pharm, 20, 3, 1758–1767

NOVARTIS | Reimagining Medicine

https://github.com/Novartis/UNIQUE/

Uncertainty Quantification: UNIQUE

- Quantify uncertainty of Machine Learning (ML) models
- Combines and **benchmark** multiple uncertainty quantification (UQ) methods
- **Example** on public LogD data*:



Input Dataframe

	canonical_smiles	molecule_chembl_id
	COc1cc(F)ccc1-c1cncc(CNCC2CC2)n1	CHEMBL1778865
	COc1cc(F)ccc1-c1cncc(CNCC2CC2)n1	CHEMBL1778865
	CN(C(=O)Cc1ccc(-n2cnnn2)cc1)[C@@H]1CCN(Cc2ccc(...	CHEMBL2010849

standard_value	fingerprints	which_set	predictions
1.80	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	TRAIN	1.889113
1.80	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	TRAIN	1.889113
1.80	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	CALIBRATION	2.633078

Evaluation Metrics

(a) Ranking-based Evaluation Benchmark

	UQ Method	Subset	AUC Difference: UQ vs. True Error	Spearman Correlation	Decreasing Coefficient	Increasing Coefficient	Performance Drop: High UQ vs. Low UQ (3-Bins)	Performance Drop: All vs. Low UQ (3-Bins)	Performance Drop: All vs. Low UQ (10-Bins)	Performance Drop: High UQ vs. Low UQ (10-Bins)
0	TanimotoDistance[fingerprints]	TEST	0.204	0.322	**4.452**	0.153	2.001	1.443	1.597	2.694
1	EnsembleVariance[variance]	TEST	0.194	0.345	4.229	0.086	2.040	1.540	1.824	2.805
2	Diff5NN[TanimotoDistance[fingerprints], EnsembleVariance[variance]]	TEST	0.256	0.217	4.240	0.960	1.526	1.185	1.171	1.712
3	Diff5NN[TanimotoDistance[fingerprints], prediction]	TEST	0.324	0.017	2.643	**1.966**	1.096	1.030	0.991	0.972
4	Dist2Var[TanimotoDistance[fingerprints]]	TEST	0.204	0.322	**4.452**	0.153	2.001	1.443	1.597	2.694
5	SumOfVariances[Dist2Var[TanimotoDistance[fingerprints]]]	TEST	0.186	0.360	4.406	0.036	2.130	1.533	1.822	2.919
6	UniqueRandomForestRegressor[fingerprints+UQmetrics+predictions](I1)	TEST	**0.185**	**0.375**	4.350	0.147	**2.214**	**1.616**	1.775	2.607
7	UniqueRandomForestRegressor[UQmetrics+predictions](I1)	TEST	0.189	0.356	4.421	0.083	2.126	1.547	**1.840**	2.952
8	UniqueRandomForestRegressor[transformedUQmetrics+predictions]	TEST	0.187	0.365	4.411	0.162	2.199	1.559	1.831	**3.094**

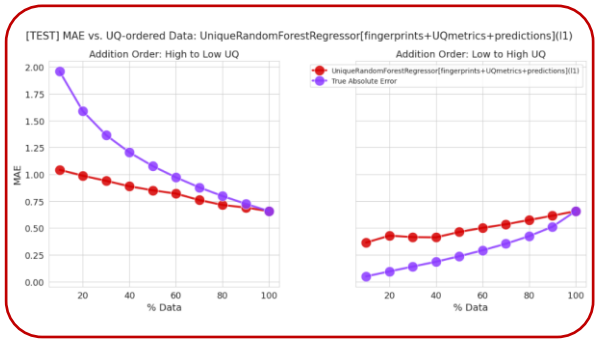
(b) Calibration-based Evaluation Benchmark

	UQ Method	Subset	MACE	RMSCE
0	EnsembleVariance[variance]	TEST	0.145	0.180
1	Dist2Var[TanimotoDistance[fingerprints]]	TEST	0.138	0.173
2	SumOfVariances[Dist2Var[TanimotoDistance[fingerprints]]]	TEST	**0.072**	**0.088**
3	UniqueRandomForestRegressor[fingerprints+UQmetrics+predictions](I1)	TEST	0.164	0.202
4	UniqueRandomForestRegressor[UQmetrics+predictions](I1)	TEST	0.166	0.204
5	UniqueRandomForestRegressor[transformedUQmetrics+predictions](I1)	TEST	0.163	0.201

(c) Proper Scoring Rules Evaluation Benchmark

	UQ Method	Subset	NLL	CheckScore	CRPS	IntervalScore
0	EnsembleVariance[variance]	TEST	1.294	0.237	0.469	2.359
1	Dist2Var[TanimotoDistance[fingerprints]]	TEST	1.228	0.238	0.472	2.367
2	SumOfVariances[Dist2Var[TanimotoDistance[fingerprints]]]	TEST	1.308	0.247	0.488	2.520
3	UniqueRandomForestRegressor[fingerprints+UQmetrics+predictions](I1)	TEST	1.226	0.237	0.469	2.364
4	UniqueRandomForestRegressor[UQmetrics+predictions](I1)	TEST	1.231	0.237	0.469	2.357
5	UniqueRandomForestRegressor[transformedUQmetrics+predictions](I1)	TEST	**1.213**	**0.236**	**0.468**	**2.345**

Visualizations



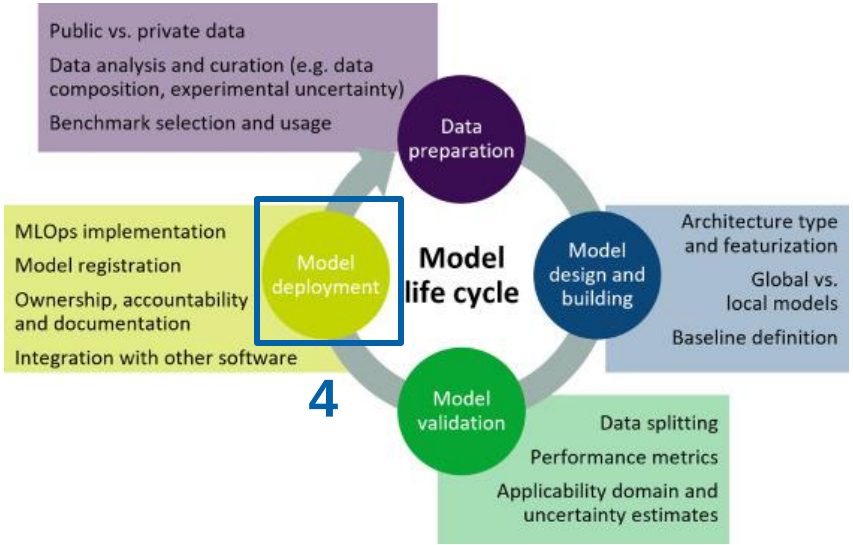
*ChEMBL (ebi.ac.uk)

NOVARTIS | Reimagining Medicine

Model life cycle

- Realizing actionable predictive models in drug discovery can be summarized as an iterative cycle of 4 steps:

1. **Training data preparation**: requires understanding the experimental data for reliable curation
2. **Model design and building**: including steps such as architecture definition and hyperparameter tuning
3. **Model validation**: performance estimation, which is crucial for prospective model use
4. **Model deployment**: the model is made available to users



Volkamer et al. 2023 AILSCI, 3, 100056

4. Model deployment

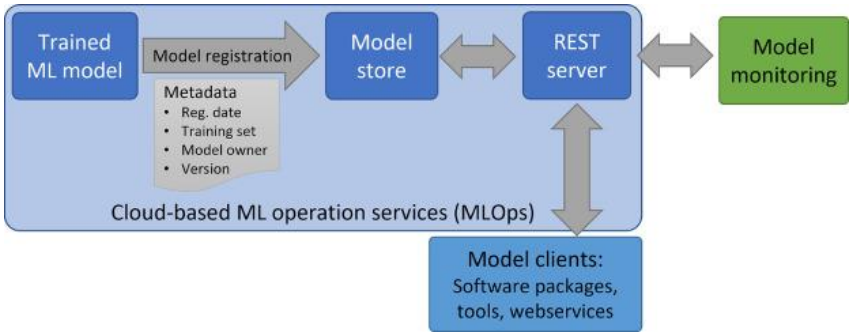
- The deployment of the trained ML model and includes several important tasks:
  - Model registration, documentation and guidelines, integration into existing tools and workflows, accessibility for non-data scientists, model ownership and accountability, monitoring, and model maintenance

Academia

- Models are less often deployed
- Difficulties to keep models accessible and updated
- Code and data are more often shared for reproducibility

Industry

- Model accessibility by users is crucial
- Ownership and accountability is considered



Volkamer et al. 2023 AILSCI, 3, 100056

NOVARTIS | Reimagining Medicine

How can industrial and academic partners work together?

COLLABORATION EXAMPLES

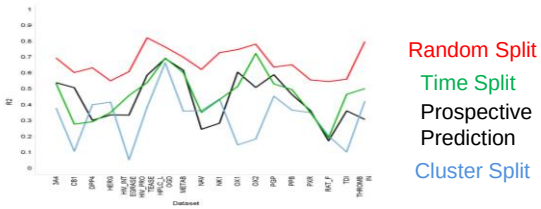
NOVARTIS | Reimagining Medicine

Beyond the challenges: collaboration among industries and academia

- **Different constraints** in industry and academia may lead to **difficulties** in the collaboration
- Collaborating or publishing may **require data sharing** thus limiting industry involvement
- Different **strategies** to overcome the problem
  - Use proprietary data to demonstrate methodology and public data for in-depth analysis
  - Federated learning (e.g. MELLODDY, FLuID)
  - Open-source tools to simplify collaboration

Example of industry-academia collaboration: SIMPD (SImulated Medicinal chemistry Project Data)

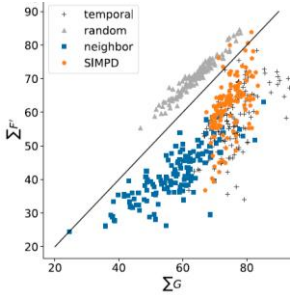
Temporal split is a better proxy for prospective validation[1]



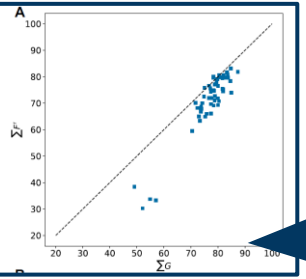
Temporal information is not always available in the public domain

Leverage academic research and proprietary data to provide a unique tool to replicate train/test splits similarly as in real project data

Use NIBR project data to learn the project **data descriptors** and **spatial statistics of test/train temporal splits**[2]



Use public data to replicate results



NOVARTIS | Reimagining Medicine

[1] Sheridan R., "Time-split cross-validation as a method for estimating the goodness of prospective prediction." Journal of chemical information and modeling, 2013
[2] Landrum, Gregory A., et al. "SIMPD: an Algorithm for Generating Simulated Time Splits for Validating Machine Learning Approaches." (2023).

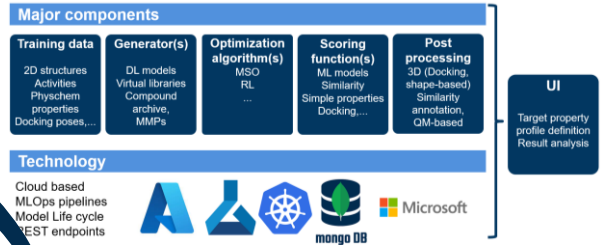
Example of industry-industry collaboration: GenChem

Collaboration between Microsoft and Novartis on generative chemistry (GenChem)

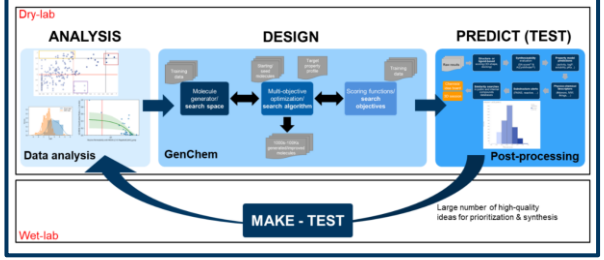
Microsoft NOVARTIS

Create a de novo design workflow to support drug discovery projects with novel and high-quality ideas

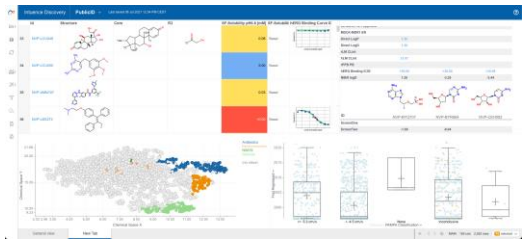
Collaborative development of the GenChem platform



Apply on internal projects and provide feedback



Integrate to internal data analysis tools



NOVARTIS | Reimagining Medicine

[1] Maziarz, Krzysztof, et al. "Learning to Extend Molecular Scaffolds with Structural Motifs." International Conference on Learning Representations.

Conclusions

- Academia and industry are both driving the field of molecular ML research
- **Models have permeated almost every step in the DMTA cycle, as we have shown with exemplary applications at Novartis**
- Due to the fast emergence of new ML algorithms, the field needs to adapt quickly, including changes in collaboration – sharing data, protocols, code, and models – and multidisciplinary scientists' education
- **More collaboration efforts between academia and industry to share data or code might lessen the gap between exploratory and applied research work**
- Some examples of private-public collaborations were mentioned showcasing constellations in which science can be advanced in real-world project set-ups while keeping sensitive data private

Thank you

Grégori Gerebtzoff
Nadine Schneider
Niko Fechner
Nik Stiefl
Elena Di Lascio
Seid Hamzic
Markus Trunzer
Bernard Faller
Sandrine Desrayaud
Richard Lewis
Finton Sirockin
Birgit Schoeberl
NIBR M&S Data Science
NIBR CADD
Intuence team
GenChem team
Other BR colleagues

Sereina Riniker
Andrea Volkamer
Eva Nittiger
Francesca Grisoni
Emma Evertsson
Greg Landrum
Jose Jimenez-Luna
Kenza Amara
Other MSR colleagues

NOVARTIS | Reimagining Medicine