Novartis Pharma AG
Global Discovery Chemistry

Prediction of small molecule developability using large-scale in silico ADMET models

Sheffield Conference on Cheminformatics
20th June 2023

Maximilian Beckers, Noé Sturm, Nikolas Fechner and Nikolaus Stiefl

NOVARTIS | Reimagining Medicine

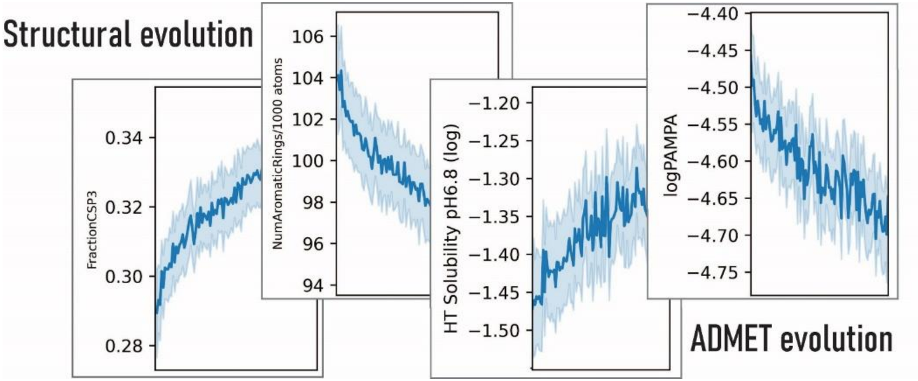
Leveraging historical MedChem optimization data

What can we learn from the past?

Patterns and trends?
Limiting factors?
New insights for early decision making?

Previously …

- Reconstruction of Novartis chemical series
- Tracing compounds during optimization
- Analysis of property evolution over time



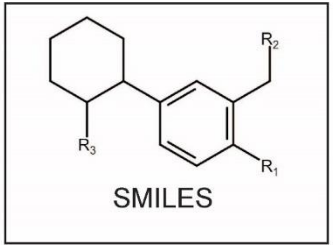
25 Years of Small-Molecule Optimization at Novartis: A Retrospective Analysis of Chemical Series Evolution
Maximilian Beckers, Nikolas Fechner and Nikolaus Stiefl - *Journal of Chemical Information and Modelling* (2022)

Today

Utilize the data to get prospective tools for compound and series evaluation

Annotation of terminal milestones for each compound
***in vitro* ADMET → *in vivo* PK → CSP → DC → Clinic**

Scoring compounds based on *in-silico* predictions



↓ MELLODDY

Global Internal Models

Predicted **ADMET**, ***in-vivo* PK** and **SAFETY** profile

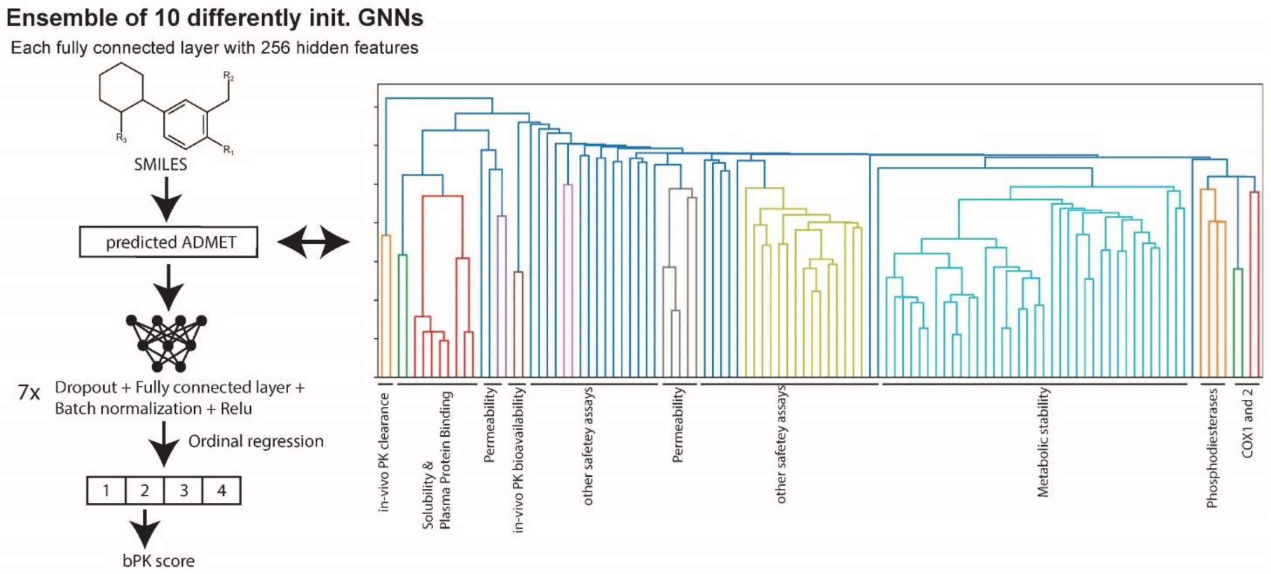
No measured data!
No target activities!

↓ **Neural network trained on compounds with annoted milestones**

Estimated potential to go beyond PK

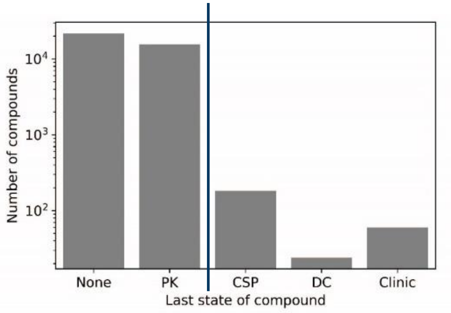
bPK score

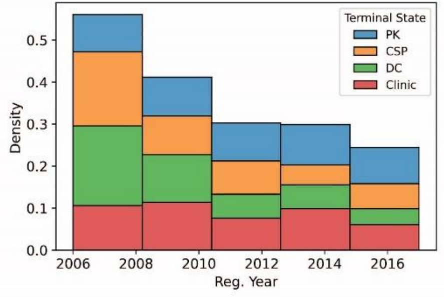
Scoring compounds based on *in-silico* predictions



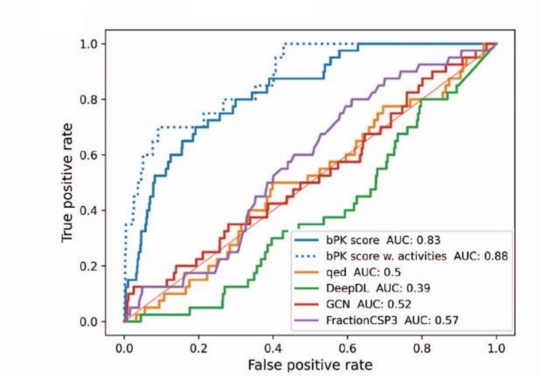
ADMET + PK + safety

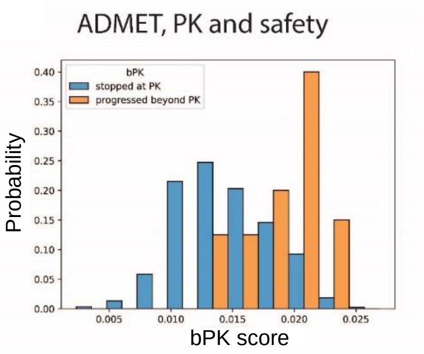
Training dataset

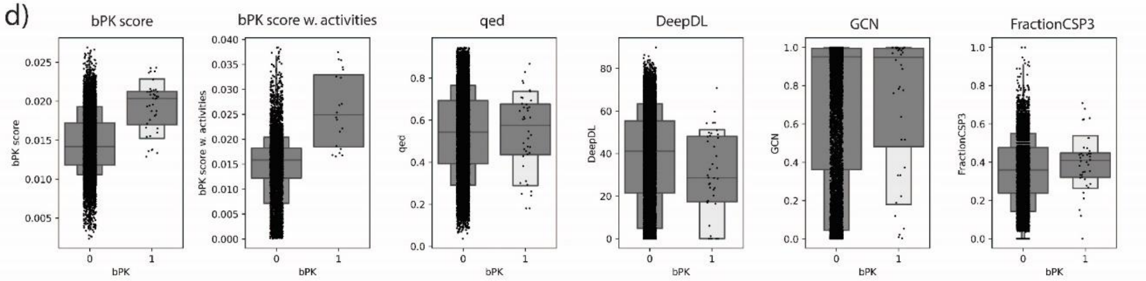




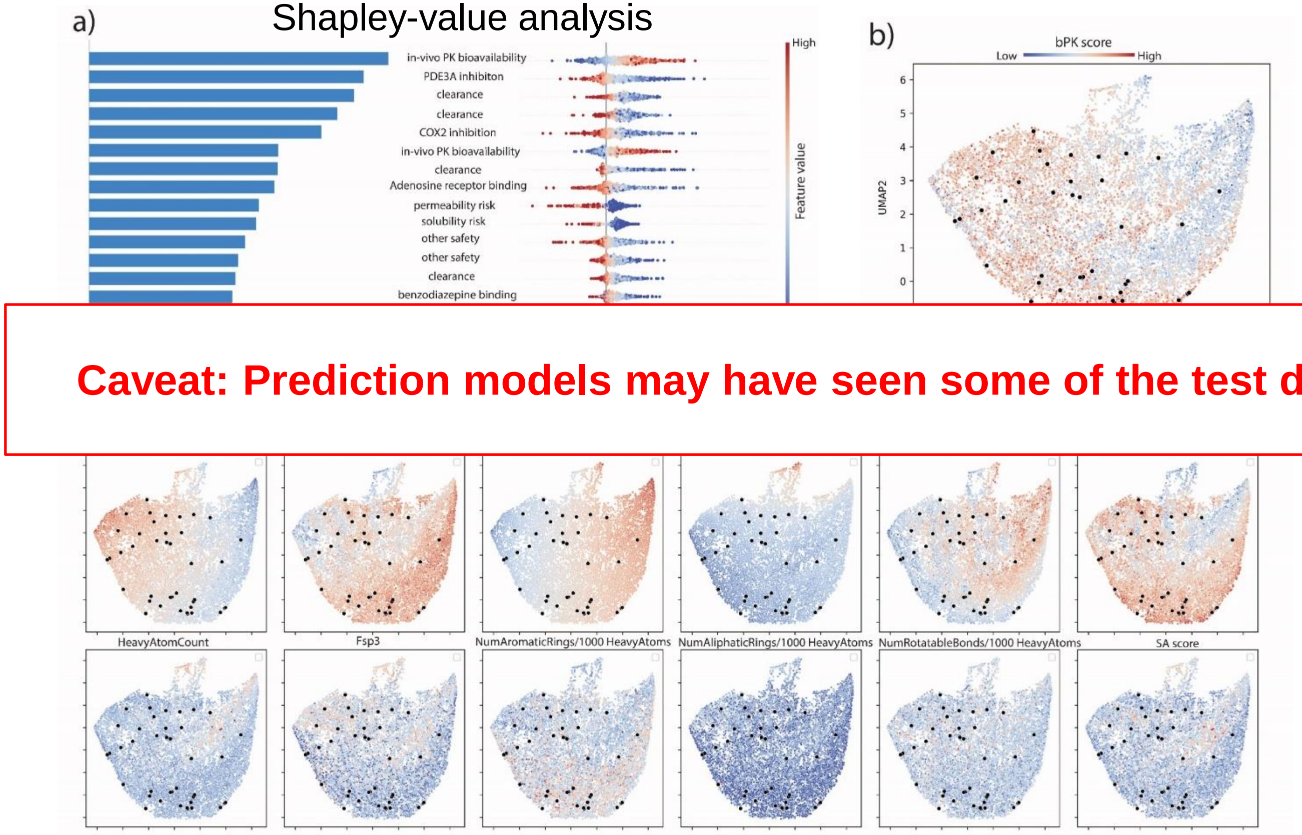
Application to Novartis internal data 2017-today







Explaining bPK scores



Shapley-value analysis

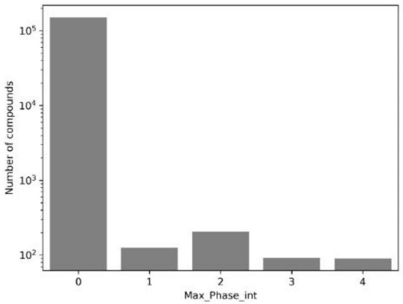
Caveat: Prediction models may have seen some of the test data

Curation of a public dataset that resembles in house compound archives

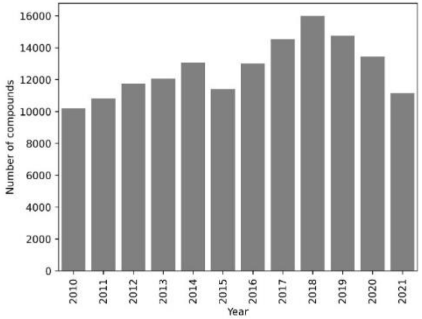
Extracted from ChEMBL

- All compounds with annotated clinical phases (“Development Candidates”)
- All other compounds from the original publication of the clinical compounds (“Series”)
- All other compounds in JMedChem Papers (“Unsuccessful series”)
- <u>Additional Restrictions</u>: max. 50 compounds per paper, registered no earlier than 2010

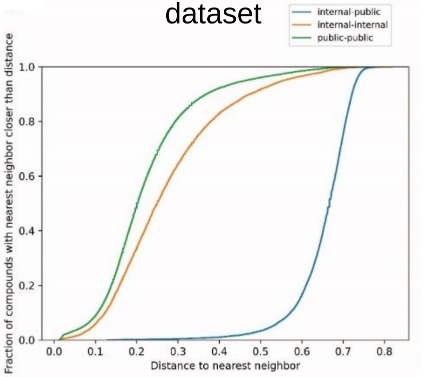
Number of compounds per clinical phase



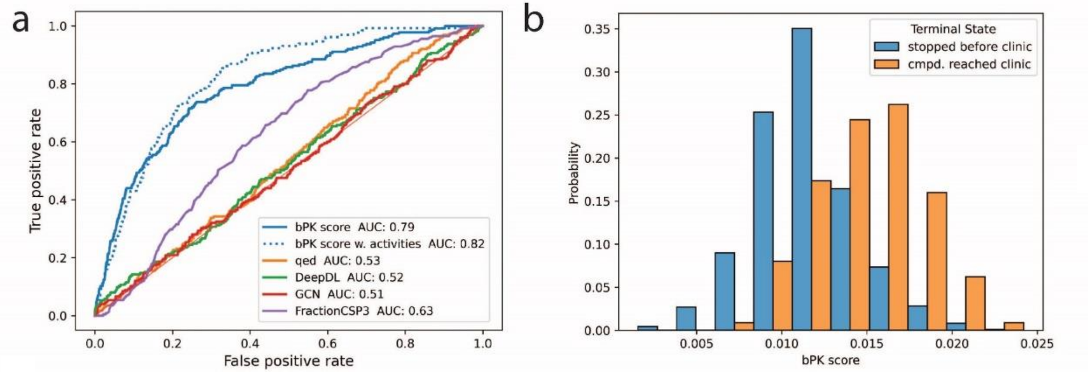
Number of compounds per year



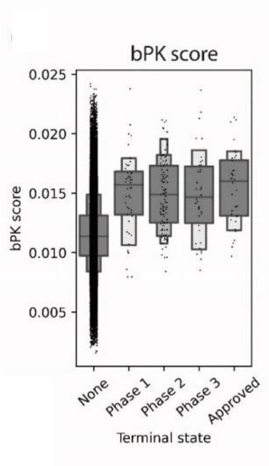
Similarity to internal test dataset



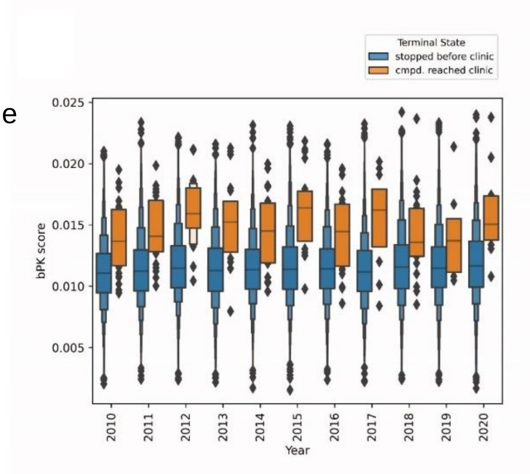
ChEMBL dataset



No differences for different clinical phases apparent



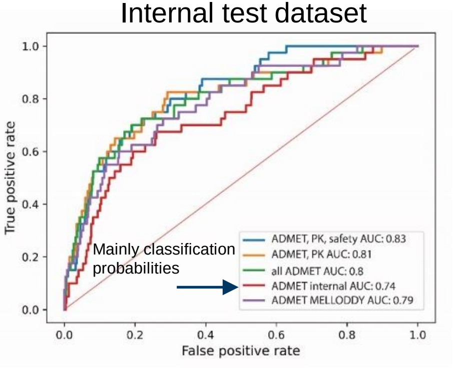
No trends over time apparent



Exploration of alternative ML approaches

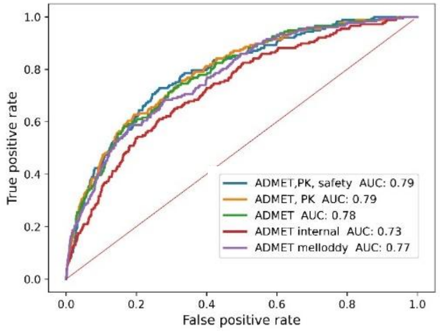
Different predicted assay endpoints

Internal test dataset

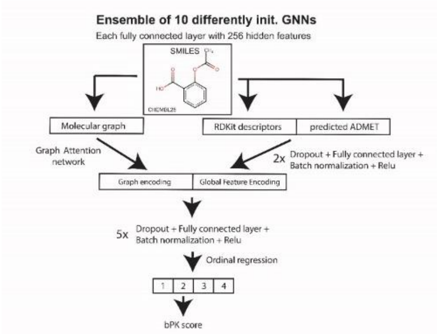


Mainly classification probabilities

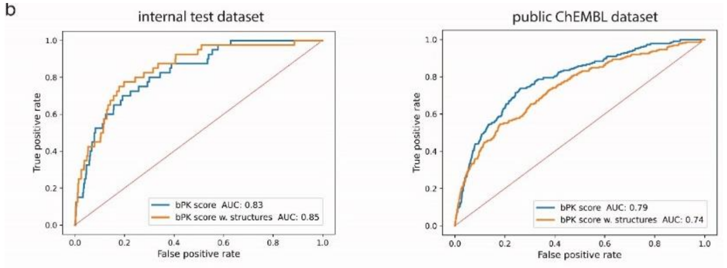
Public ChEMBL dataset



Graph neural network

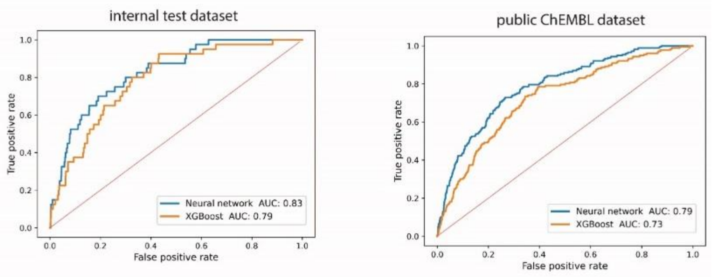


Ensemble of 10 differently init. GNNs
Each fully connected layer with 256 hidden features
SMILES
CHEMBL25
Molecular graph
RDKit descriptors
predicted ADMET
Graph Attention network
2x
Dropout + Fully connected layer + Batch normalization + Relu
Graph encoding
Global Feature Encoding
5x
Dropout + Fully connected layer + Batch normalization + Relu
Ordinal regression
1
2
3
4
bPK score



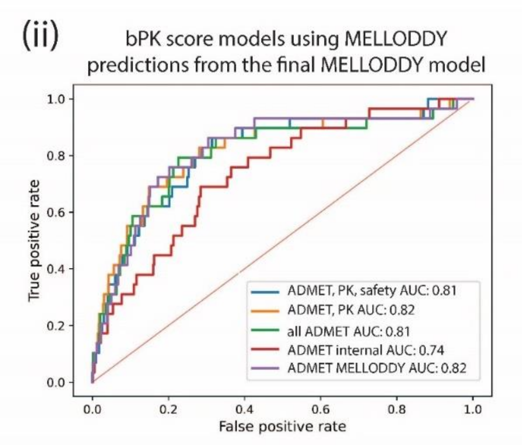
XGBoost

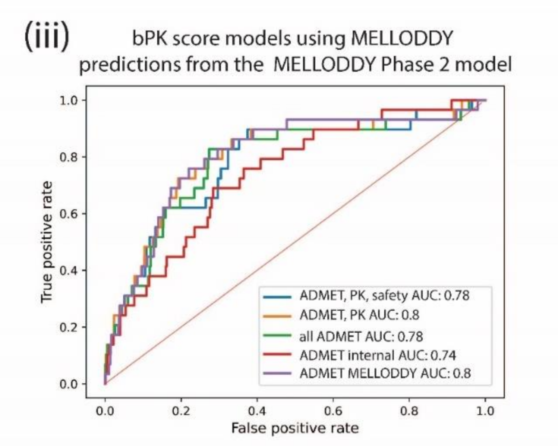
- 'colsample_bylevel': 0.56
- 'eta': 0.40, 'gamma': 0.09
- **'max_depth’: 1**
- **'num_rounds': 80**
- 'scale_pos_weight': 3.10
- 'objective': 'binary:logistic'



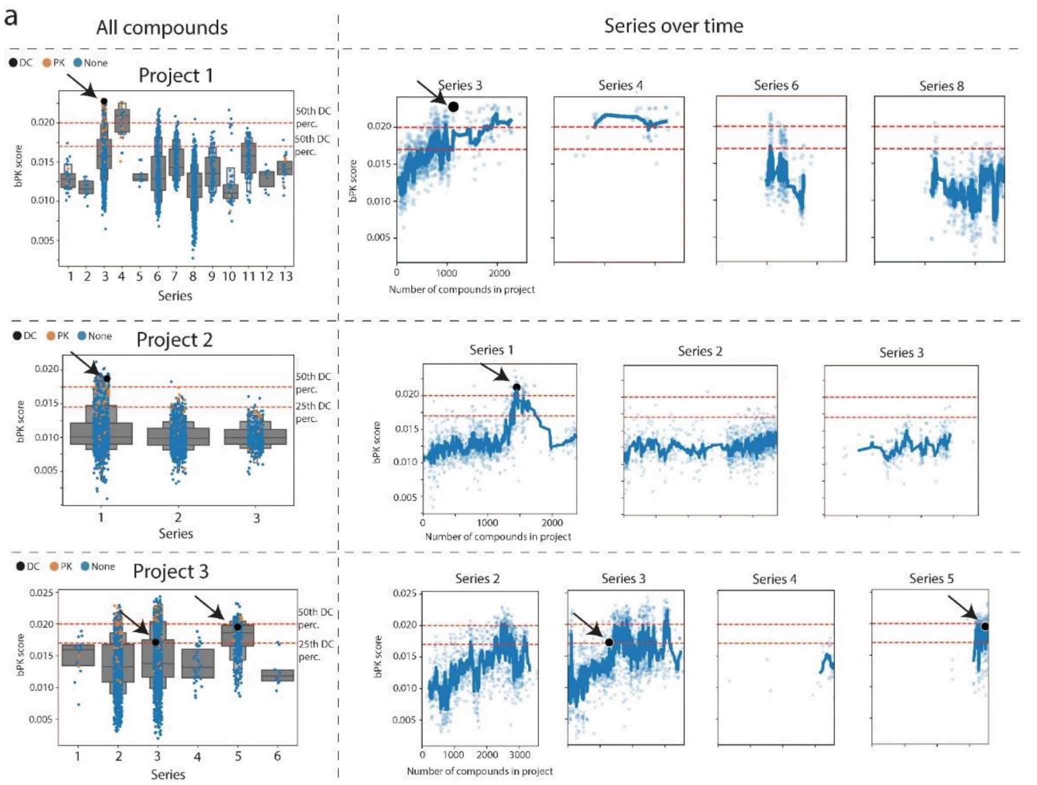
Last source of train-test leakage: Exploiting MELLODDY test-folds

- MELLODDY was trained using data from other companies, which could be in our public dataset
- Scaffold-based train-test splitting strategy was employed for MELLODDY
  → create subset of our public dataset not seen by MELLODDY Phase 2 models



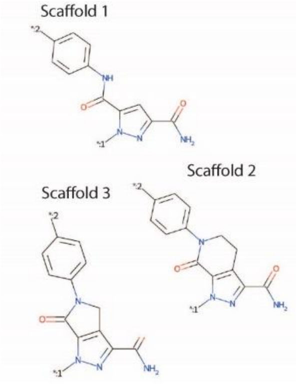


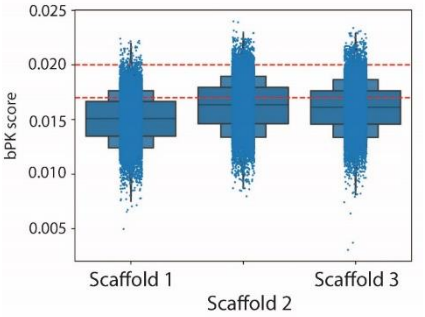
Application to three in-house projects

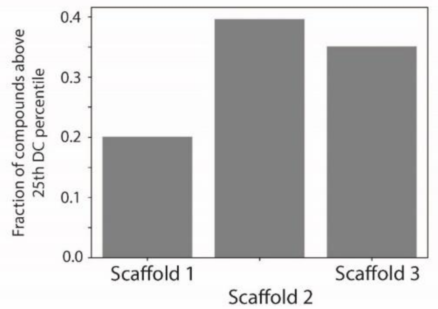


Application to in silico generated virtual compounds

- Scaffold prioritization



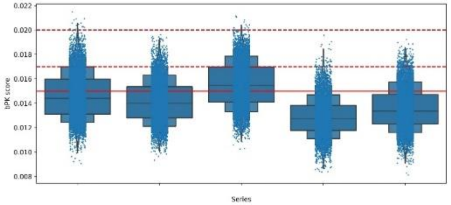


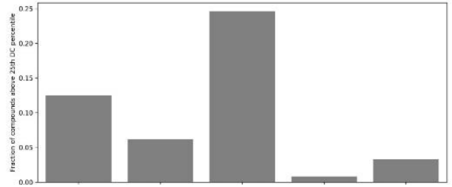


Pinot *et al.* Discovery of 1-(4-Methoxyphenyl)-7-oxo-6-(4-(2-oxopiperidin-1-yl)phenyl)-4,5,6,7-tetrahydro- 1*H*-pyrazolo[3,4-*c*]pyridine-3-carboxamide (Apixaban, BMS-562247), a Highly Potent, Selective, Efficacious, and Orally Bioavailable Inhibitor of Blood Coagulation Factor Xa, *J. Med. Chem. 50, 22 (2007)*

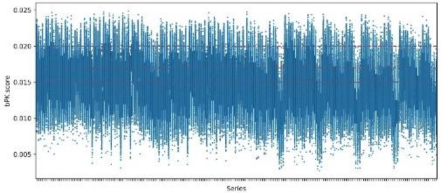
In-house projects

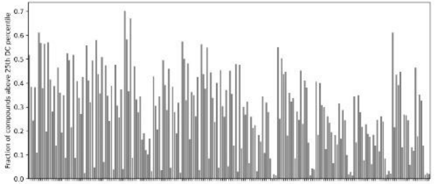
Project A



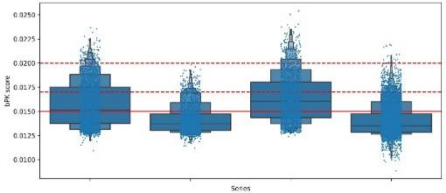


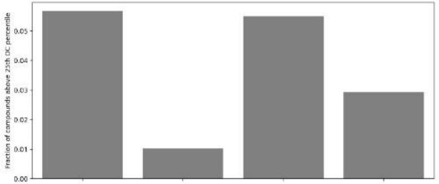
Project B





Project C





Challenges

- Project specific information
  - Mode of action
  - Formulation
- Applicability domain
  - New modalities
- False negatives make training hard
  - Only one out of many possible other compounds is selected as DC
  - Strategic and operational reasons complicate labelling

Outlook

- Further application to de-novo generation of molecules
- bPK scores for ultra-large enumerated libraries (e.g. for virtual screening)
- Screening follow ups, prioritization of new scaffolds
- Monitoring progress of optimization projects
- DC identification

Acknowledgements

PostDoc Mentors
Nik Stiefl (GDC)
Nikolas Fechner (NX)

GDC
Nadine Schneider
Jessica Lanini
Paolo Tosco
Oh-hyeon Choung
Yves Auberson
Claudia Betschart
Rainer Machauer
Trixi Brandl
Henrik Moebitz
David Carcache
Nina Gomermann
Thomas Knoepfel
Christian Markert

NIBR Postdoctoral Program
Anne Granger

NIBR Informatics (NX)
Holger Hoefling
Nicholas Holway

CBT
Noe Sturm
Ansgar Schuffenhauer

PKS
Gregori Gerebtzoff
Raquel Rodriguez-Perez
Jimmy Kromann

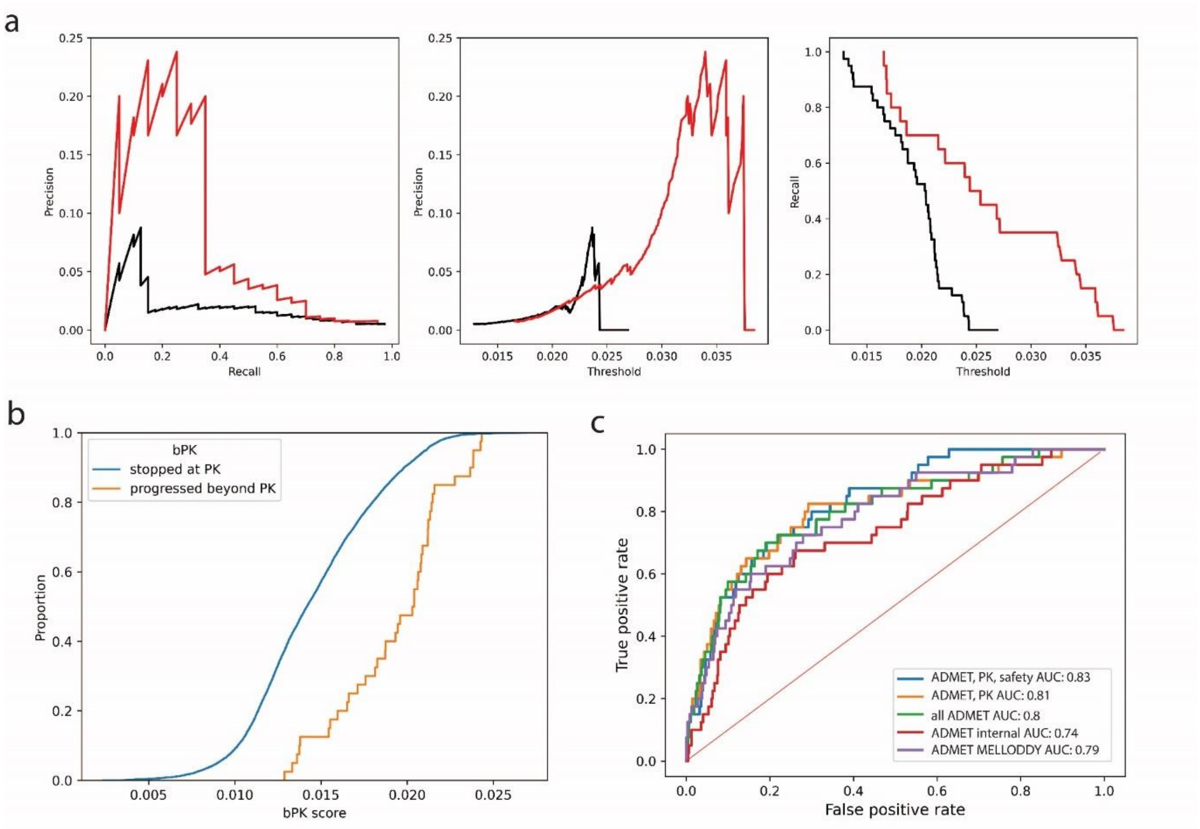
ETH Zürich
Gregory Landrum
Sereina Riniker

And everyone at Novartis who participated and commented in internal presentations

Thank you

NOVARTIS | Reimagining Medicine

Appendix


a
Precision
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Recall
Precision
0.25
0.20
0.15
0.10
0.05
0.00
0.015
0.020
0.025
0.030
0.035
Threshold
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.015
0.020
0.025
0.030
0.035
Threshold
b
bPK
stopped at PK
progressed beyond PK
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.005
0.010
0.015
0.020
0.025
bPK score
c
True positive rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
ADMET, PK, safety AUC: 0.83
ADMET, PK AUC: 0.81
all ADMET AUC: 0.8
ADMET internal AUC: 0.74
ADMET MELLODDY AUC: 0.79