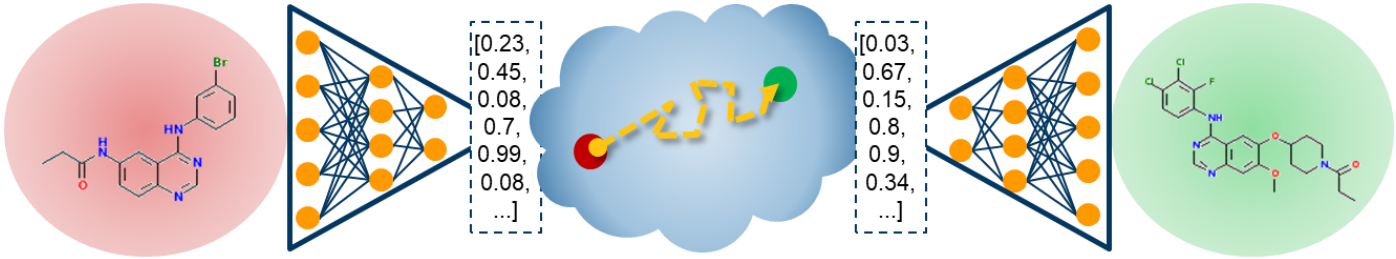


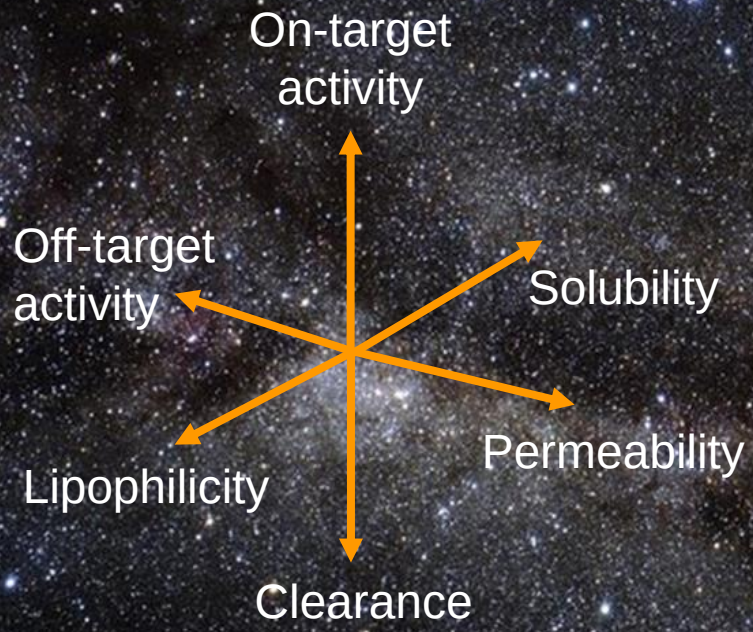
Augmenting Drug Hunters with Generative Chemistry Models

Nik Stiefl, NIBR

Summer School on Cheminformatics, Strasbourg, 2022

NOVARTIS | Reimagining Medicine

Drug hunting: How to identify the best candidate from 10^{60}/10^{23} potential molecules?



Low data availability

Biased sampling

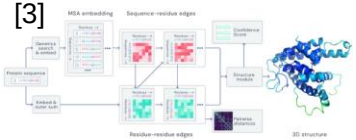
Discontinuous optimization landscape

NOVARTIS | Reimagining Medicine

Leverage ML methods to augment medchem teams

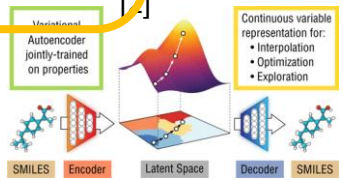
Protein structure prediction

[3]



Integrated de novo design

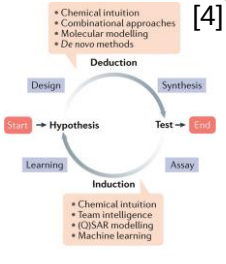
[2]



Automation:

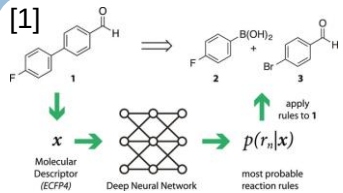
ML-driven closing the loop (DMTA cycle)

[4]



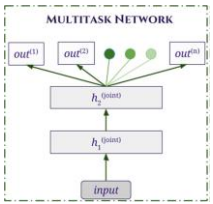
Retro-synthesis & reactivity prediction

[1]



Activity & off-target prediction

[5]



[1] Segler, M. H., & Waller, M. P. (2017). Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chemistry–A European Journal, 23(25), 5966-5971.
[2] Gómez-Bombarelli, R., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS central science, 4(2), 268-276.
[3] Jumper J. et al. High Accuracy Protein Structure Prediction Using Deep Learning. AlphaFold2, DeepMInd, Google
[4] Schneider G (2018) Automating drug discovery. Nat Rev Drug Discov 17:97–113
[5] Ramsundar, B., et al (2017). Is multitask deep learning practical for pharma?.Journal of chemical information and modeling, 57(8), 2068-2076.

NOVARTIS | Reimagining Medicine

Shiny headlines & big hopes

https://www.311institute.com/an-ai-designed-30000-drugs-in-21-days-and-came-up-with-winners/

AI accelerates drug discovery time from 3 years to 21 days

https://www.longevity.technology/ai-platform-accelerates-drug-discovery-time-from-3-years-to-21-days/

AI MODEL YIELDS NEW DRUG TO OVERCOME ANTIBIOTIC RESISTANCE

https://www.healthcareitnews.com/ai-powered-healthcare/ai-model-yields-new-drug-overcome-antibiotic-resistance

NEWS / 02.20.20

Artificial intelligence yields new antibiotic

By Anne Trafton, MIT News Office

https://www.broadinstitute.org/news/artificial-intelligence-yields-new-antibiotic

Combining generative artificial intelligence and on-chip synthesis for de novo drug design [1]

Beam Search for Automated Design and Scoring of Novel ROR Ligands with Machine Intelligence [2]

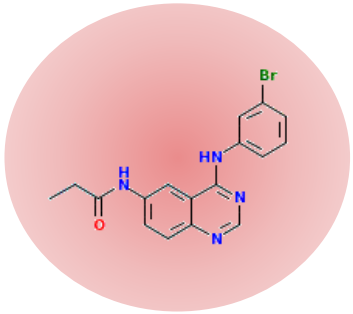
JAEGER – Hunting for Antimalarials with Generative Chemistry [3]

[1] Grisoni, F., Huisman, B. J., Button, A. L., Moret, M., Atz, K., Merk, D., & Schneider, G. (2021). Combining generative artificial intelligence and on-chip synthesis for de novo drug design. Science advances, 7(24), eabg3338.
[2] Moret, M., Helmstädter, M., Grisoni, F., Schneider, G., & Merk, D. (2021). Beam Search for Automated Design and Scoring of Novel ROR Ligands with Machine Intelligence. *Angewandte Chemie International Edition*.
[3] Godinez, W., Ma, E., Chao, A., Pei, L., Skewes-Cox, P., Canham, S., ... & Guiguemde, A. (2021). JAEGER–Hunting for Antimalarials with Generative Chemistry. Chemrxiv, DOI 10.33774/chemrxiv-2021-5t5xx

NOVARTIS | Reimagining Medicine

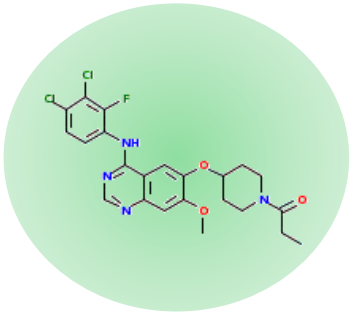
Generative chemistry

The principle



Non-optimal compound

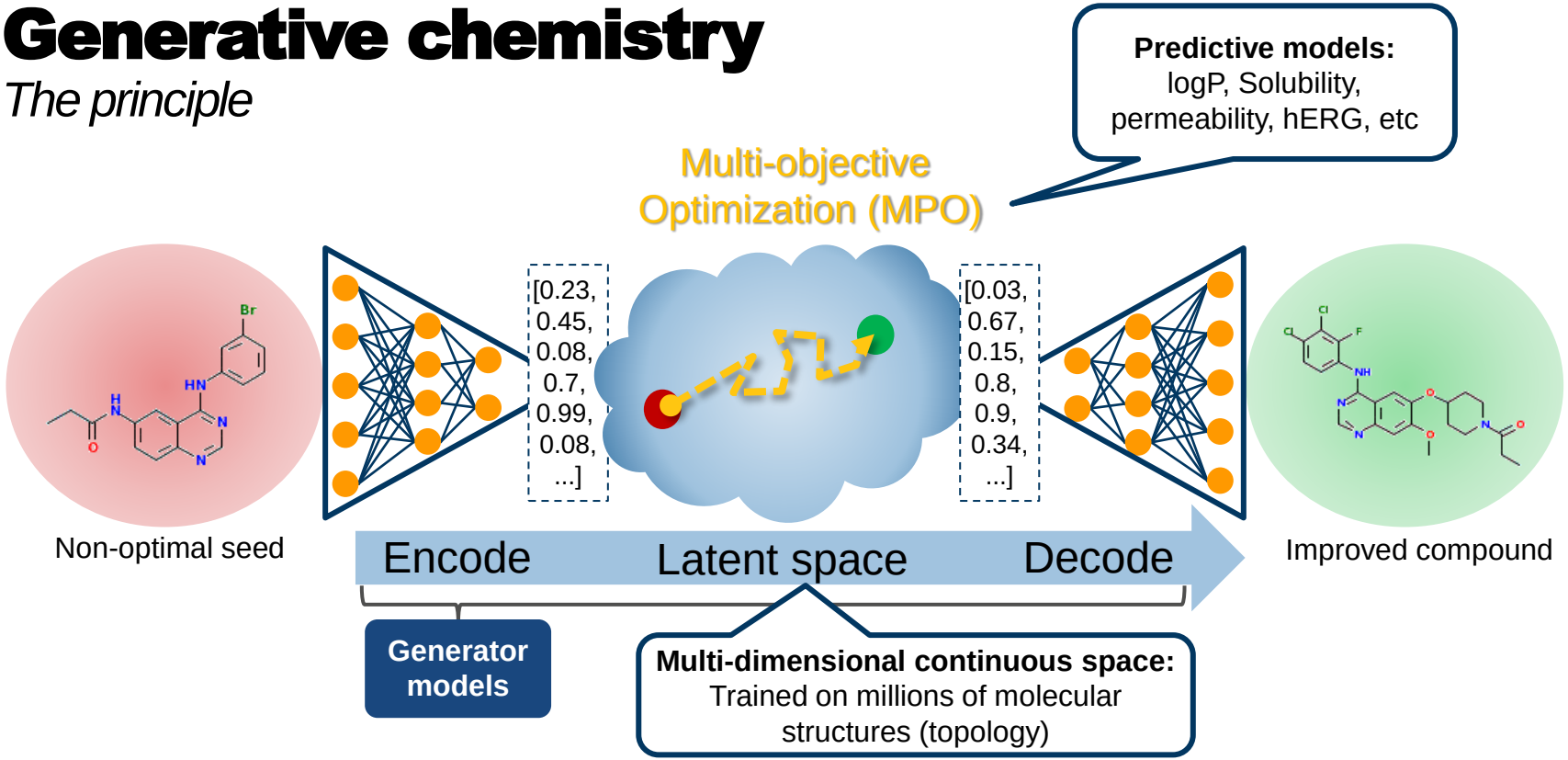
MedChem optimization



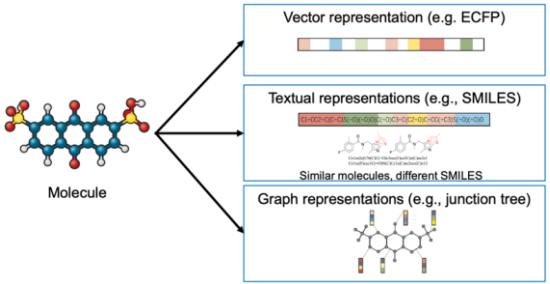
Improved compound

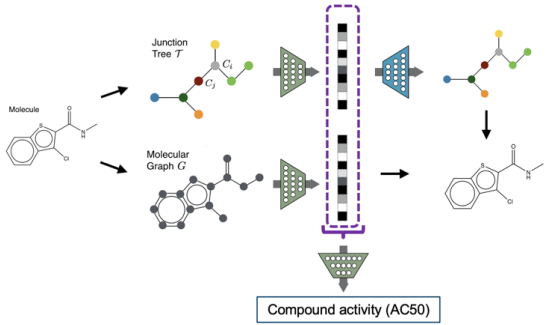
Generative chemistry

The principle



First, computers need to learn chemistry!





Method	Reconstruction	Validity
CVAE	44.6%	0.7%
GVAE	53.7%	7.2%
SD-VAE[2]	76.2%	43.5%
GraphVAE	-	13.5%
JT-VAE	**76.7%**	**100.0%**

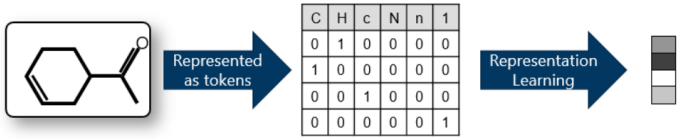
Junction tree variational autoencoder for molecular graph generation. International Conference on Machine Learning, pp. 2323-2332. PMLR, 2018.

Why is this important?

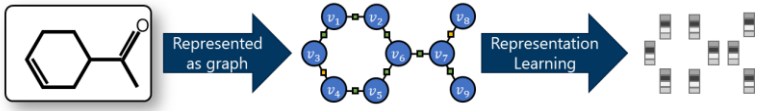
- Generative chemistry is futile without generation of **valid** molecules.
- Enables constraints, such as keeping scaffold fixed.

Example generator models

String-based methods (e.g. CDDD[1])



Graph-based methods (e.g. CGVAE[2], MoLeR[3])



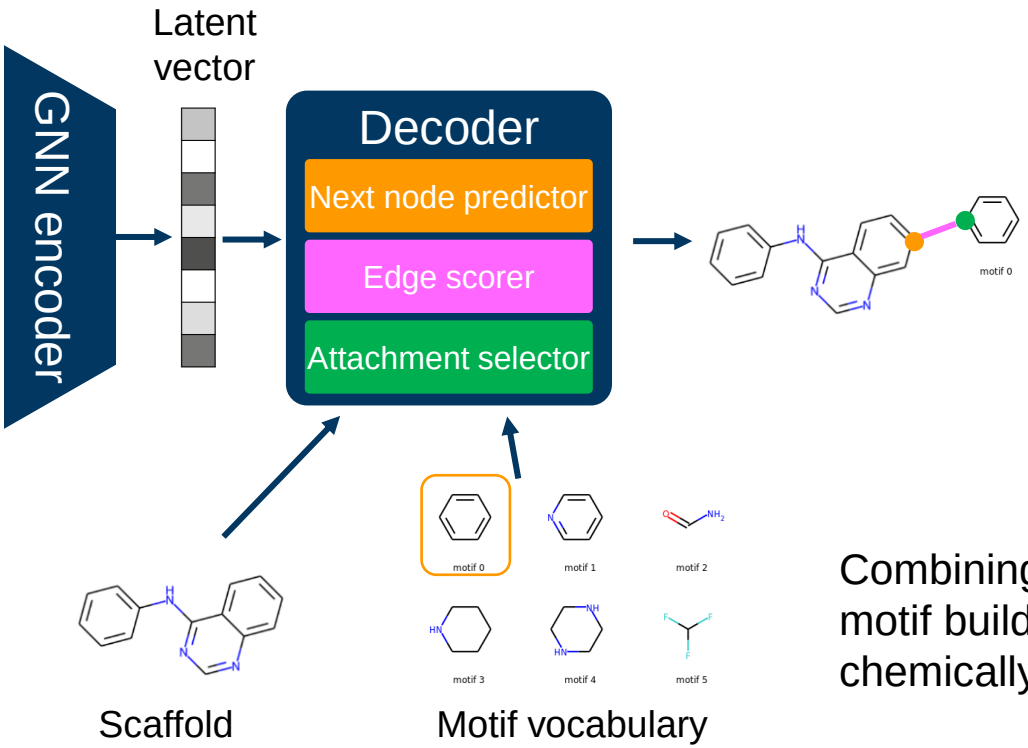
Conditional generation using [signatures, profiles, sequences] (e.g. pqsar2cpd[4])

Many other approaches exist, and new ones appear very frequently (often w/ open-source). Major application approaches:

- **Exploration: distribution learning** (reproduce sets of molecules)
- **Exploitation: goal-directed generation** (search latent space without full sampling)

[1] Winter, R. *et al. Chem. Sci.* **10**, 1692–1701 (2019)
[2] Jin, W. *et al. arXiv* (2019). https://arxiv.org/pdf/1802.04364.pdf
[3] Maziarz, K. *et al. arXiv* (2021) https://arxiv.org/pdf/2103.03864.pdf
[4] Pikusa M, *et al. bioarXiv* (2022) https://biorxiv.org/content/10.1101/2021.12.10.472084v1

MoLeR: a scaffold aware generator



Method	Guacamol Score	Guacamol Quality	Scaffolds Score	Scaffolds Quality
Best of dataset [9]	0.61	0.77	0.17	0.94
SMILES LSTM [9]	0.87	0.77	0.24	0.80
SMILES GA [9]	0.72	0.36	0.45	0.22
GRAPH MCTS [9]	0.45	0.22	0.20	0.64
GRAPH GA [9]	0.90	0.40	0.79	0.64
CDDD + MSO [47]	0.90	0.58	0.92	0.59
MNCE-RL [48]	0.92	0.54	-	-
MoLeR + MSO	0.82	0.75	0.93	0.61

Maziarz, K. *et al. arXiv* (2021)
https://arxiv.org/pdf/2103.03864.pdf

Brown, N. et al. (2019). GuacaMol: benchmarking models for de novo molecular design. *JCIM 59*(3), 1096-1108.

Combining an atom-by-atom with a motif-by-motif build-up enables high exploration and chemically valid molecules

De-novo generation as reverse phenotypic profiling

Predict compound activity

Design the compound

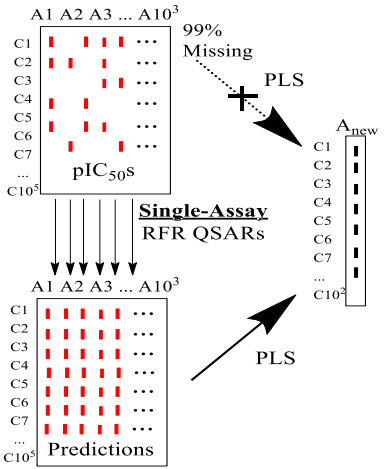
pQSAR as reliable predicted activity profiles

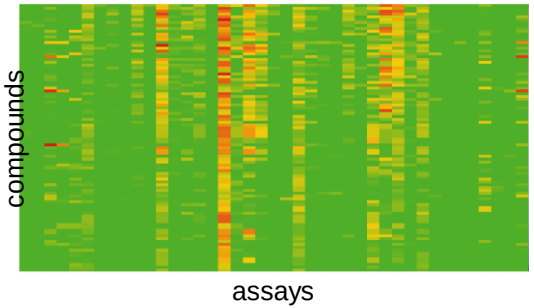
All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration IC_{50}s for 8558 Novartis Assays

Eric J. Martin,[*,†] Valery R. Polyakov,[†,§] Xiang-Wei Zhu,[†] Li Tian,[†,‡,∥] Prasenjit Mukherjee,[†,⊥] and Xin Liu[†,‡,#]

[†]Novartis Institute for Biomedical Research, 5300 Chiron Way, Emeryville, California 94608-2916, United States
[‡]China Novartis Institutes for BioMedical Research Company, Limited, 2F, Building 4, Novartis Campus, No. 4218 Jinke Road, Zhangjiang, Pudong, Shanghai 201203, China

J Chem Inf Model **59,** 4450–4459 (2019).

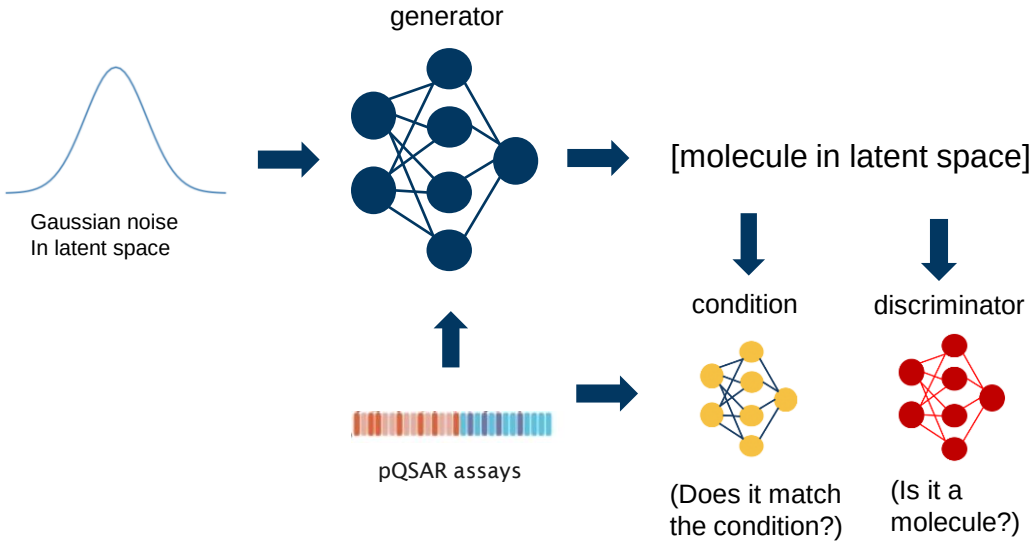




NOVARTIS | Reimagining Medicine

pqsar2cpd – zero-sum game

- Conditional Generative Adversarial Networks (Goodfellow et al. 2014) → co-training of chemistry and profile (per project specificity)
- **Generator tries to deceive the discriminator by creating samples that are hard to distinguish from real data**

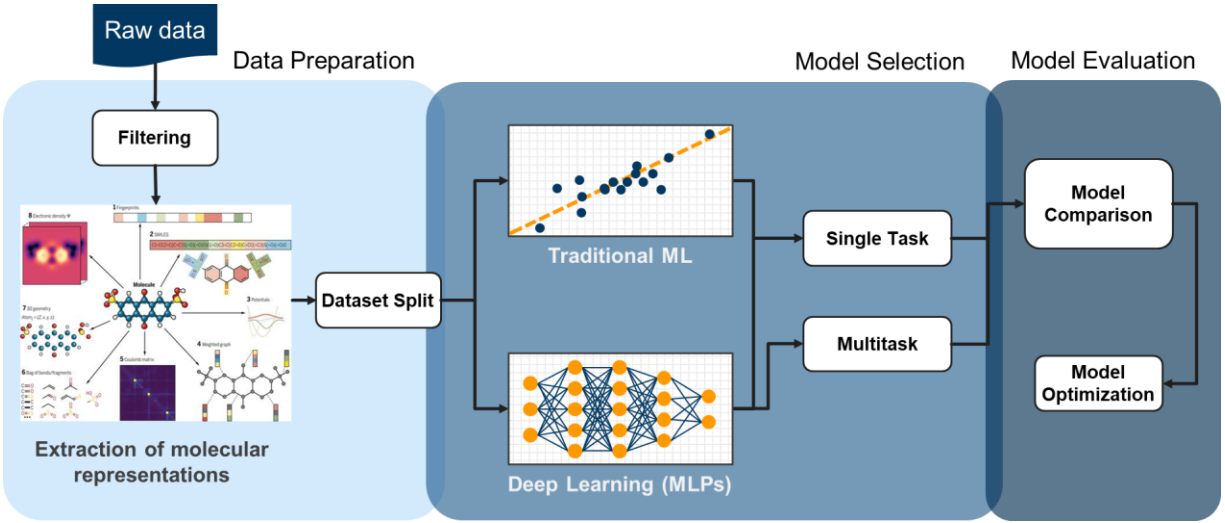


Predictive models for MPO

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

Simplification of:

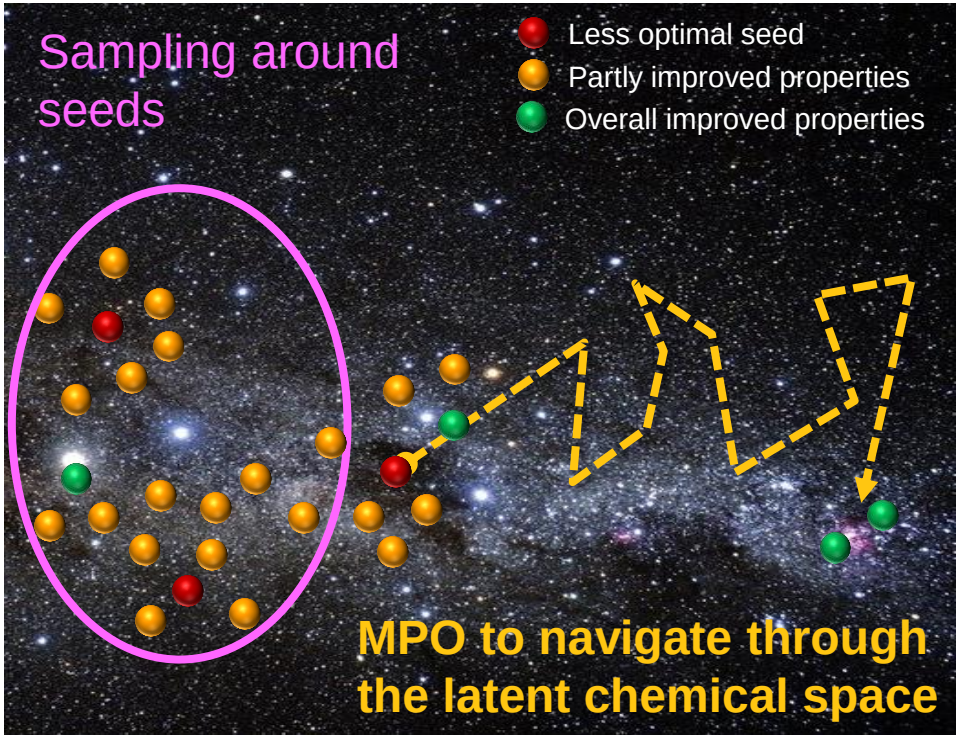
- Data preparation
- Comparison of molecular representations
- Comparison of ML models
- Evaluation and optimization of final models



Molecular representation image source: Sanchez-Lengeling, B., & Aspuru-Guzik, A. *Science*, **361(6400)**, 360-365 (2018).

[1] Lanini, J. *et al. Manuscript in preparation*

Multi-objective optimization (MPO)



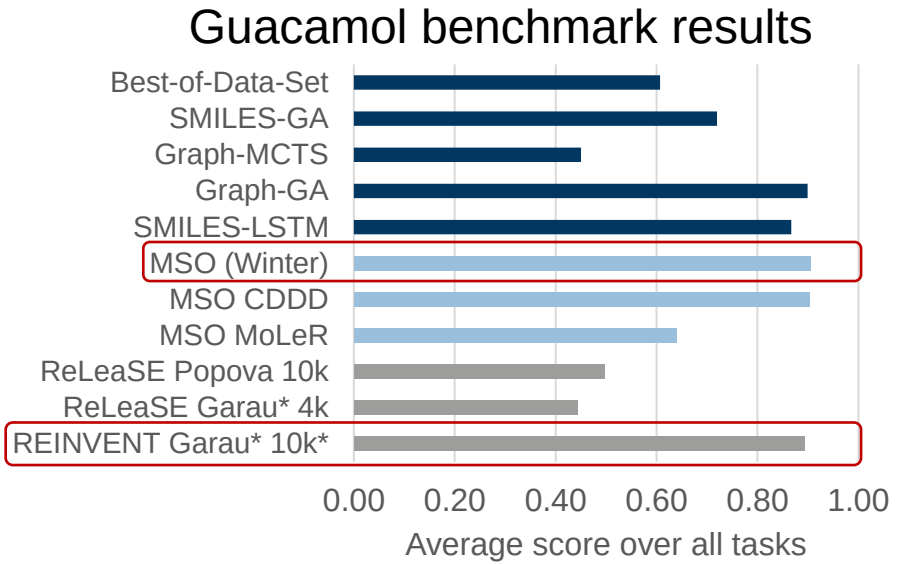
- **Pre-defined target property profile** guides the search in the latent space
- **Predictive models** used to determine properties of a new point in the latent space
- **Example methods:** evolutionary algorithm (MSO), reinforcement learning, etc.
- **<u>Challenges</u>**:
  - Reliable predictive models
  - Contradicting properties
  - Combination of optimization algorithm and generator model

Multi-Objective Optimization Strategies

Using different strategies

Different strategies are needed to explore different optimization task:

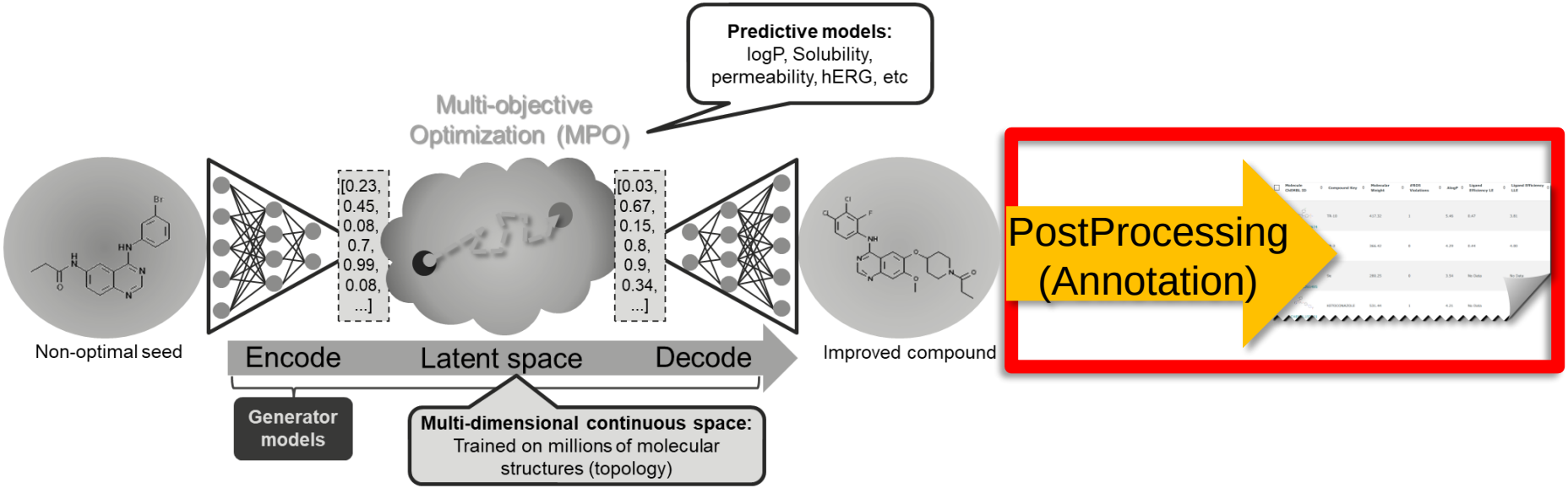
- Fast, interactive and local optimization:
  - MSO (molecule swarm optimization)
- Global optimization with strong project-specific focus:
  - Reinforcement Learning-based methods (REINVENT, ReLeaSE)
  - Genetic algorithms (GA)
  - MonteCarlo Tree Search (MCTS)



Baselines taken from the Guacamol benchmark: Brown, N., et al.. **59(3)**, 1096-1108. (2019)

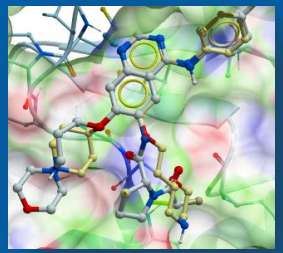
Generative chemistry in NIBR

Enriching GenChem output with MedChem relevant information



Postprocessing Workflow - Annotation

Structure-based or Ligand-based



Latent Space Property Models

	EGFR predictions	solubility predictions	logD predictions
1	0.89	0.9901	1.8
2	0.8402	0.9936	1.9
3	0.8918	0.993	1.9
4	0.8815	0.9945	1.7

Substructure Alerts

	SubstructureMatches	Min N O filter	Frac N O	Covalent	SpecialMoi	SeverityScore
4	no match	no match	0.2	0	0	0
5	Screeningdeck_2019_halogen_aromatic_count_3_min(3) | Screeningdeck_2019_polyhalogenated_aromate_min(1)	no match	0.2	0	0	1
6	Screeningdeck_2019_phenol_ester_min(1)	no match	0.25	0	0	0

Global Property Models

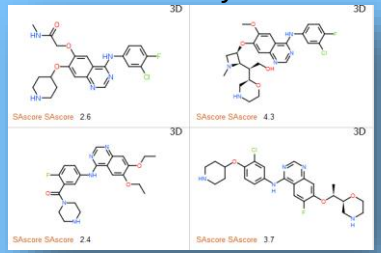
	NIBR-logP nibr:logP	NIBR-logD nibr:logD(pH=7.4)	Solubility (HT-eq,pH6.8) pH6.8 HT Solubility Classification	BSEP ([3H]taurocholate uptake) BSEP Class ([3H]taurocholate uptake)
1	2.8	1.3	>100 uM	inconclusive
2	3.7	2.3	>100 uM	<= 30 uM
3	3.0	0.7	>100 uM	inconclusive
4	3.5	2.0	>100 uM	inconclusive
5	3.6	1.6	>100 uM	inconclusive
6	3.9	2.1	>100 uM	inconclusive

2D based descriptors

	TPSA	ExactMW	NumLipinskiHBA	NumLipinskiHBD	NumRotatableBonds	FractionCSP3	NumHeavyAtoms
1	97	459.1	8	3	7	0.3	32
2	101	517.2	9	3	8	0.4	36
3	89	439.2	8	2	7	0.3	32
4	90	501.2	8	3	7	0.4	35
5	78	460.2	7	2	8	0.4	32

Synthetisability evaluation

SA score & AiZynthfinder[1]



[1] Genheden, S. *et al. Journal of Cheminformatics* **12**, 70 (2020).

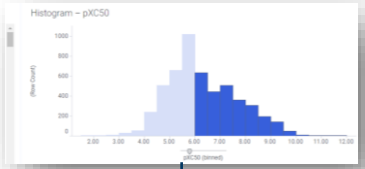
Similarity searches in public and internal compound DBs*

Enamine Hit Smiles	Enamine Hit Similarity	Enamine Hit IDs	Mcule Hit Smiles	Mcule Hit Similarity	Mcule Hit IDs	ChEMBL25 Hit Smiles	ChEMBL25 Hit Similarity	ChEMBL25
smiles	0.56	Z1765419275	smiles	0.50	DXIIZFGPCYZEQW-UHFFFAOYSA-N	smiles	0.84	CHEMBL21
smiles	0.58	PV-002369348(	smiles	0.51	XSRYZILJXXHJBH-UHFFFAOYSA-N	smiles	0.83	CHEMBL56

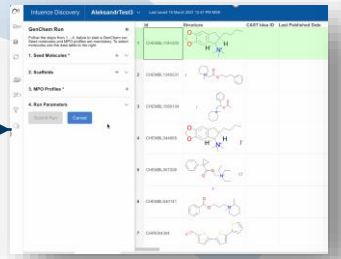
* Internal DBs (not shown): Novartis Corporate Archive & CAST ideas

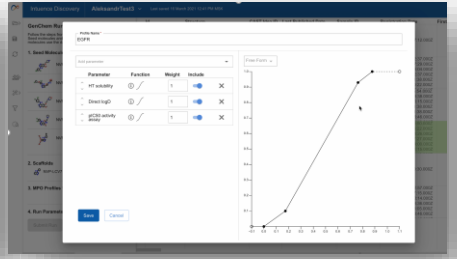
GenChem in Action

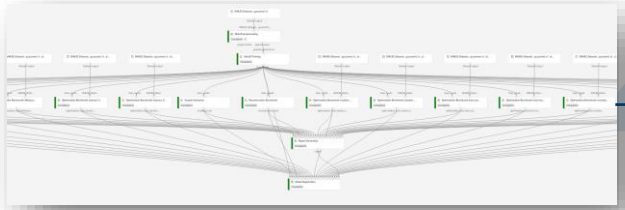
Project data curation & local model building (DS/CADD)



Define profile & start generation







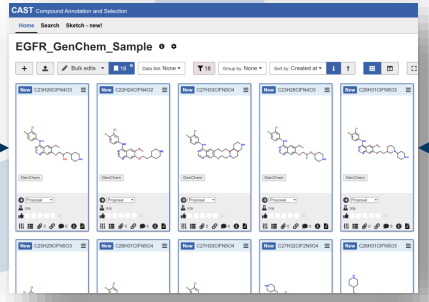
Global generator &
model training by GenChem team
~ 1-2x / year

Model store

Molecule generation & MPO

Postprocess

Output for **discussion & selection**



Synthesis & testing

NOVARTIS | Reimagining Medicine

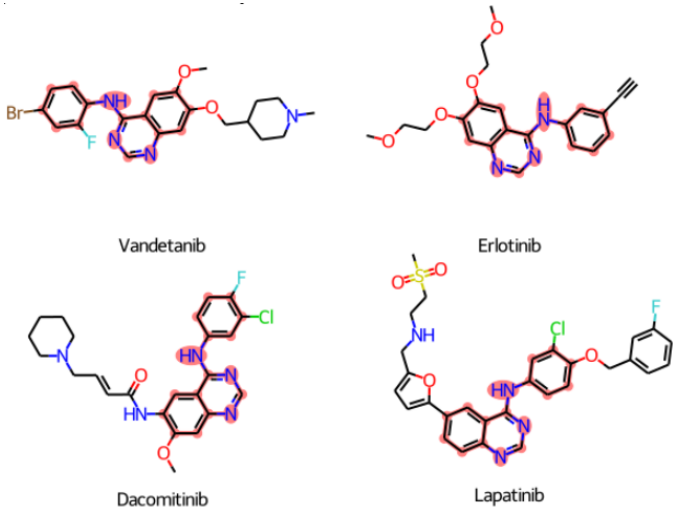
Real-world applications

- Public example – EGFR
- Observations from *in house* data sets

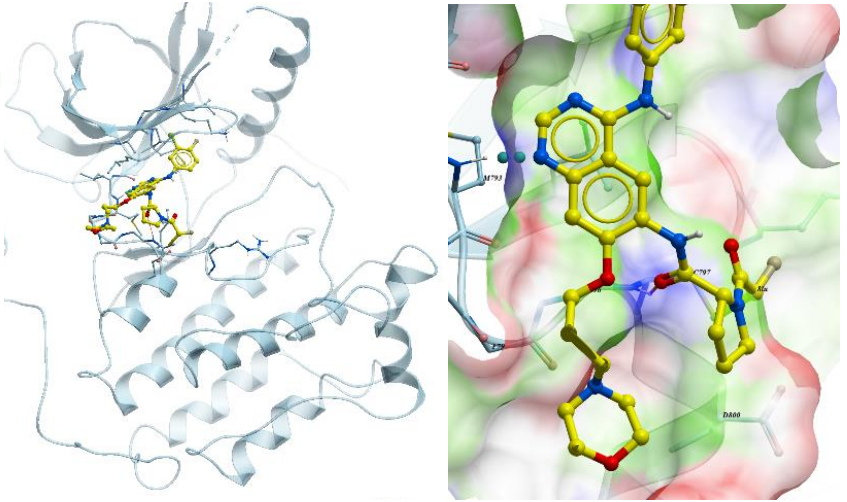
* Sun et al. *JCIM*, **2017**, *9*, 17
* Prykhodko et al. *JCIM*, **2019**, *11*, 74

GenChem application example: EGFR

EGFR is a tyrosine-kinase targeted in Non-Small Cell Lung Cancer



Drugs targeting EGFR with amino-quinazoline core



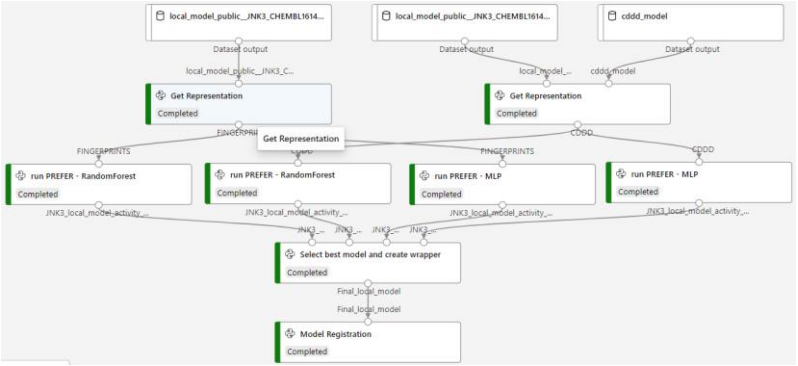
Crystal structure of an amino-quinazoline compound bound to EGFR (PDB: 5y25)

GenChem application example: EGFR

Set-up

Step 1: local model building

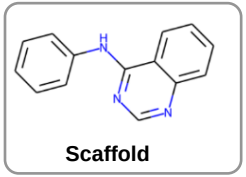
- EGFR model (ExCape dataset[1], #5204 cmpds)
- JNK3 model (ChEMBL dataset, #362 cmpds)

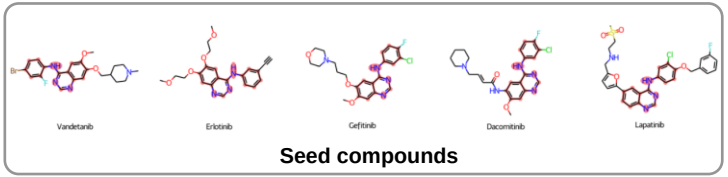


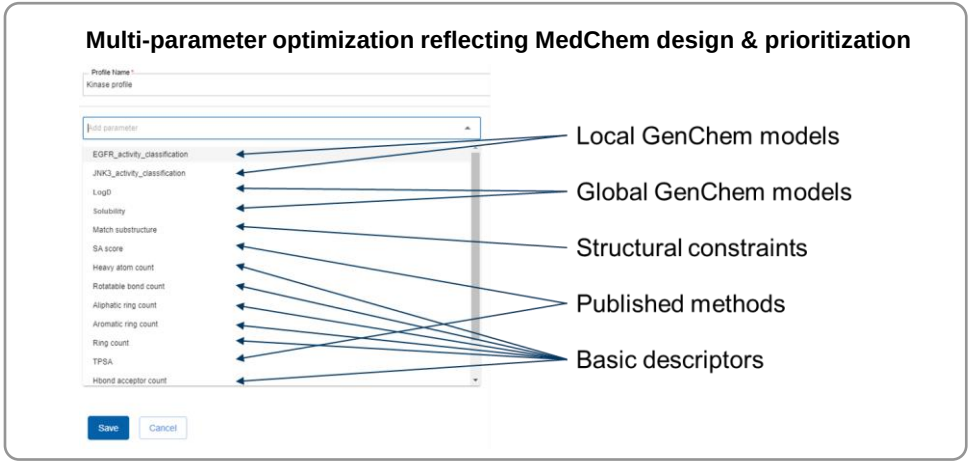
- High-quality EGFR model (AUC 0.92), medium quality JNK3 model (AUC 0.66)

[1] Sun, J., et al. (2017). ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *Journal of cheminformatics*, *9*(1), 1-9.

Step 2: MPO and GenChem run set-up





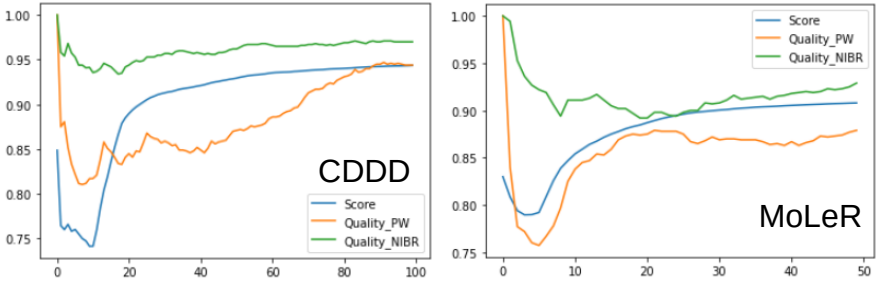


NOVARTIS | Reimagining Medicine

GenChem application example: EGFR

Quantitative results

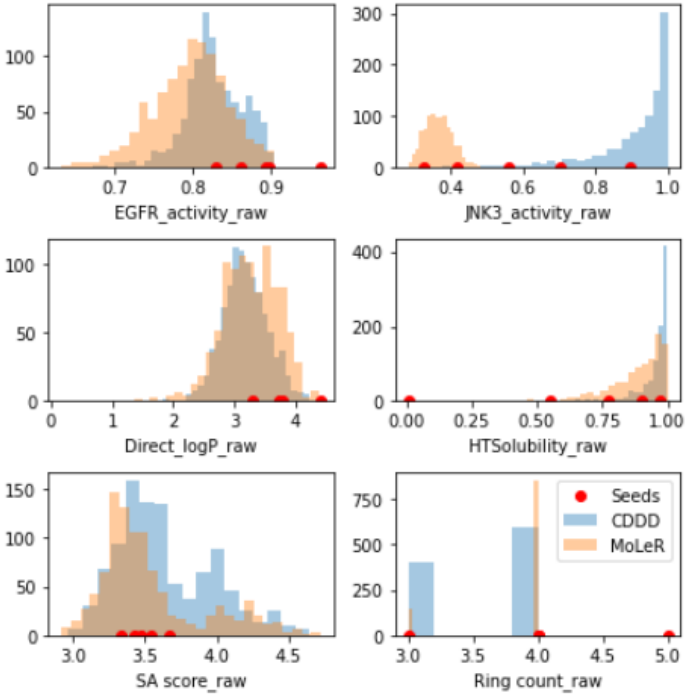
Score and quality evolution



Optimization steps

Optimization steps

Property distribution of 2000 best cmpds



Runtime:

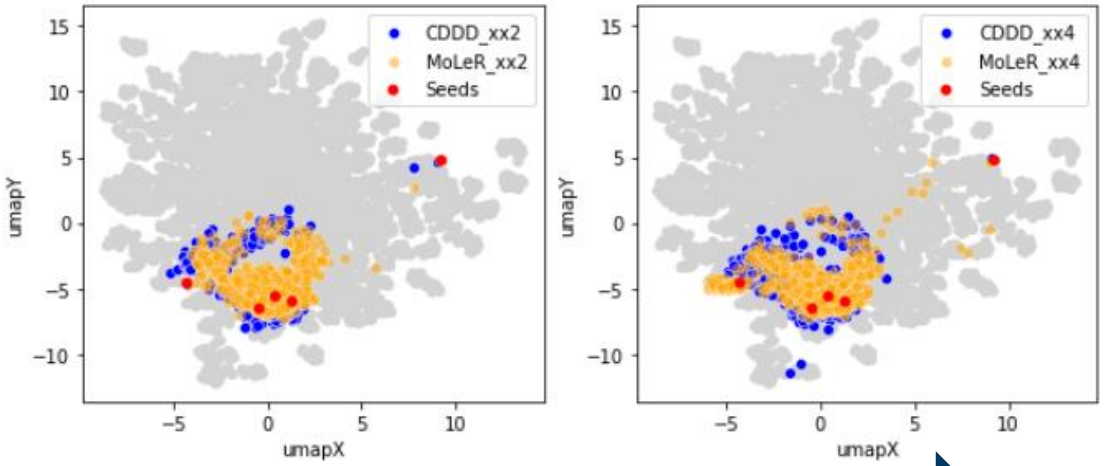
- 5 seeds, 40 particles, 50-100 steps, overall 30’000 new molecules generated in <1h
- 2000 best molecules were kept (1000 per generative model)

NOVARTIS | Reimagining Medicine

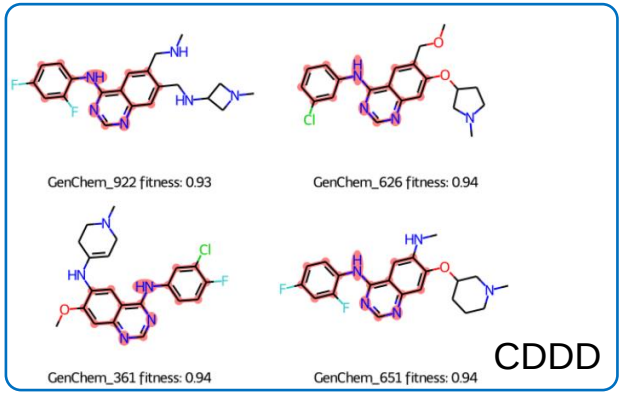
GenChem application example: EGFR

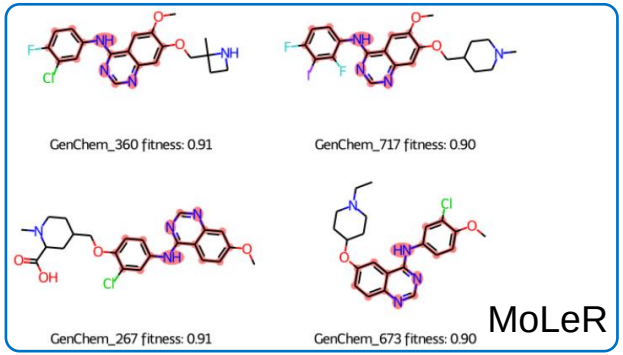
Qualitative results

EGFR chemical space (in gray) [1]



Increasing exploration level



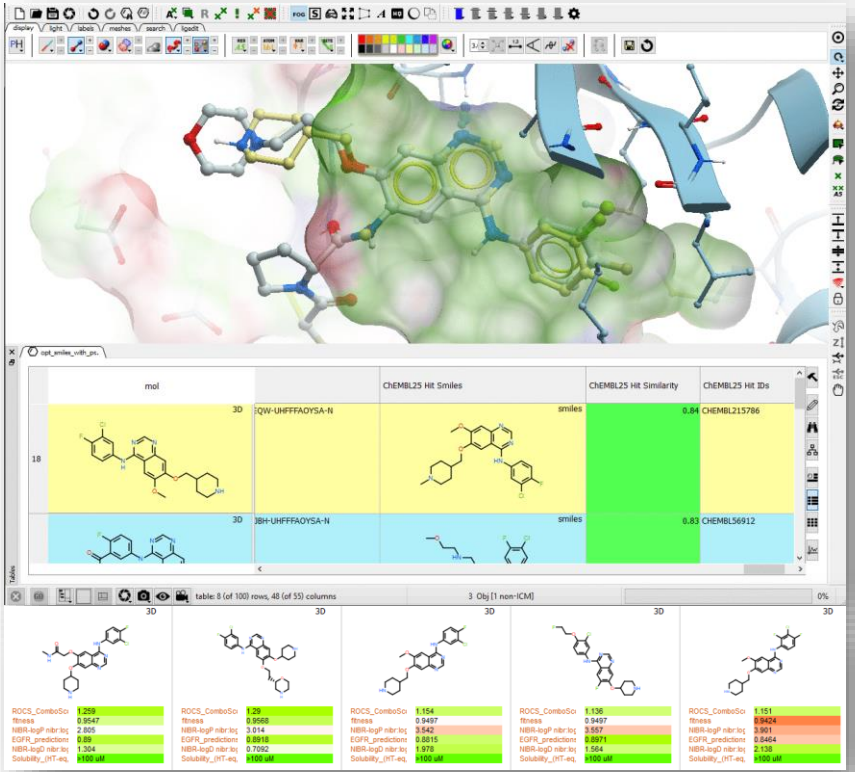


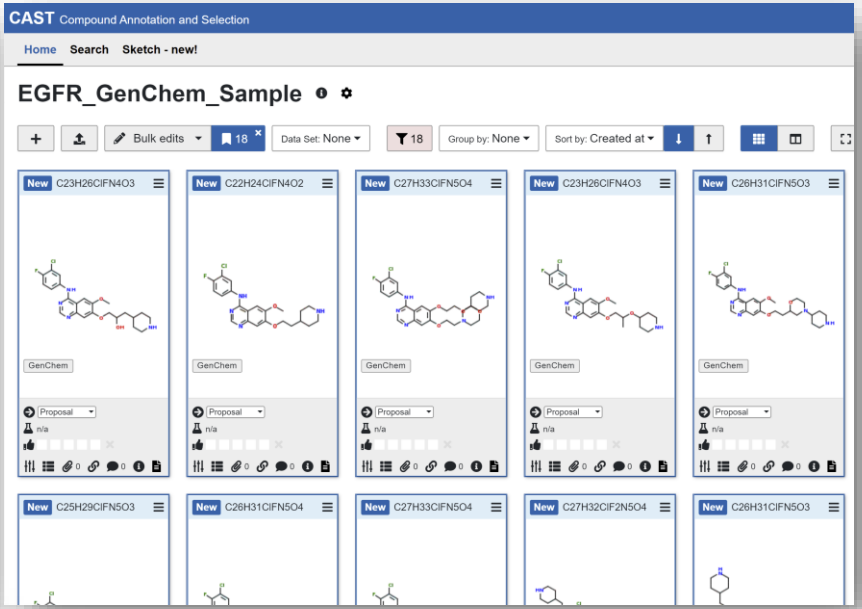
[1] ExCape dataset, #5204 cmpds
Sun, J. et al. (2017). ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *Journal of cheminformatics*, *9*(1), 1-9.

NOVARTIS | Reimagining Medicine

GenChem application example: EGFR

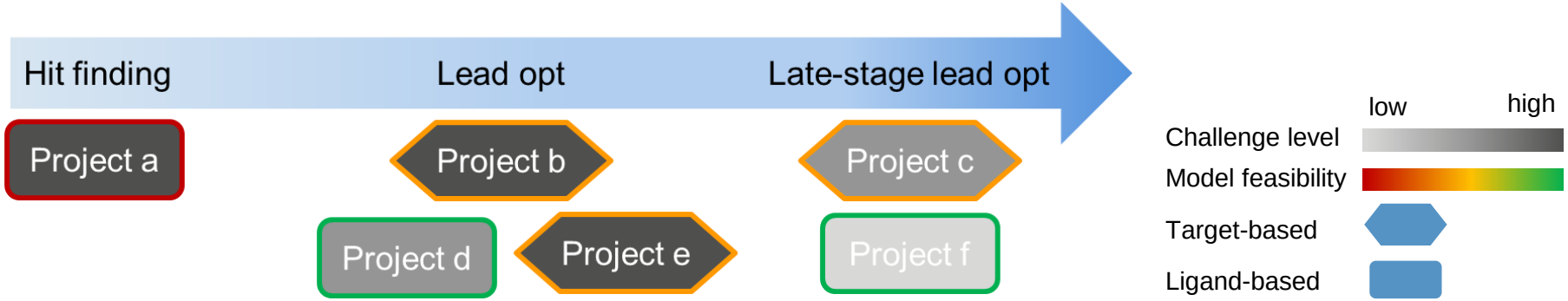
Post-processing outputs multiple ways to analyse results





NOVARTIS | Reimagining Medicine

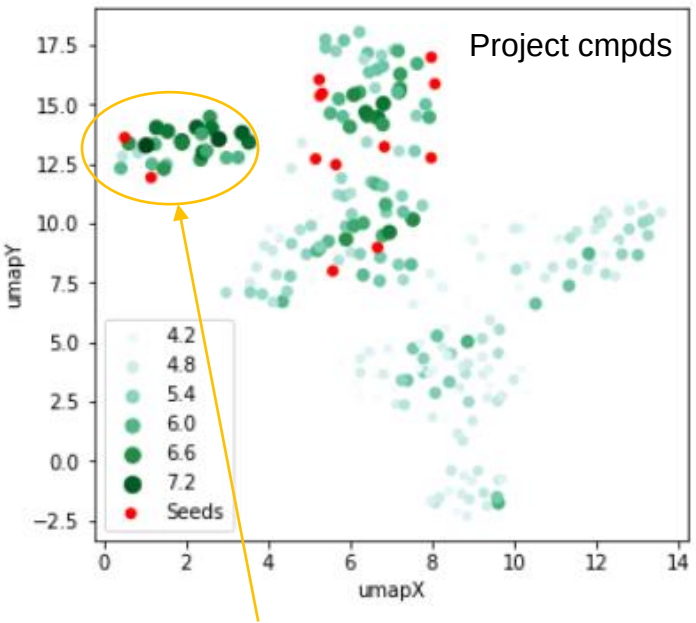
MedChem project selection



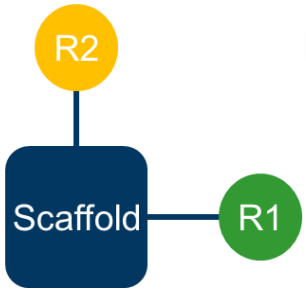
- **Selection criteria**: broad coverage of project stages, challenge level, data availability, MedChem team commitment,...
- **Goal**: Increase the benefit for both - MedChem project teams and GenChem enhancements

- **Other challenging aspects**:
  - Scientific
  - Social
  - Strategic

Project 1: setup



Cluster of highly active cmpds (including lead series)



Exploration of 2 exit vectors:

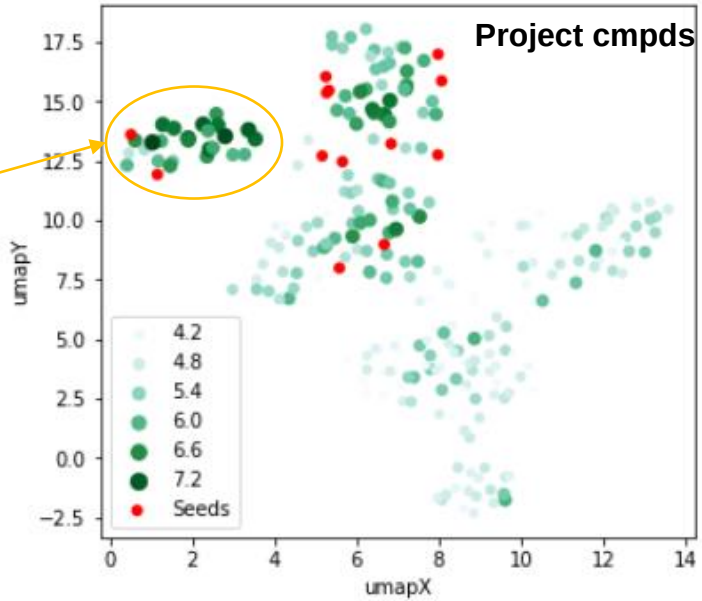
- R2: very sharp SAR, team already identified good vectors here, covered by selecting a diverse set of seeds
- R1: main interest of the team

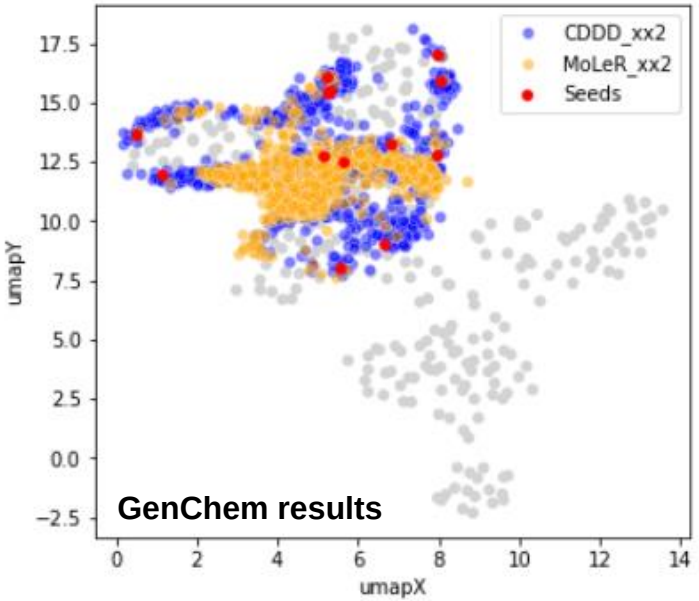
MPO definition: improve activity and lipophilicity

- **Activity**:
  - R1 activity regression model
  - Overall activity classification model
- **Phys-chem properties**:
  - lipophilicity regression model
  - permeability classification model
  - solubility model
- **Chemical attractiveness**: SA score, Heavy atom count, Rotatable bonds

Project 1: Chemical space exploration

Cluster of highly active cmpds (including lead series)

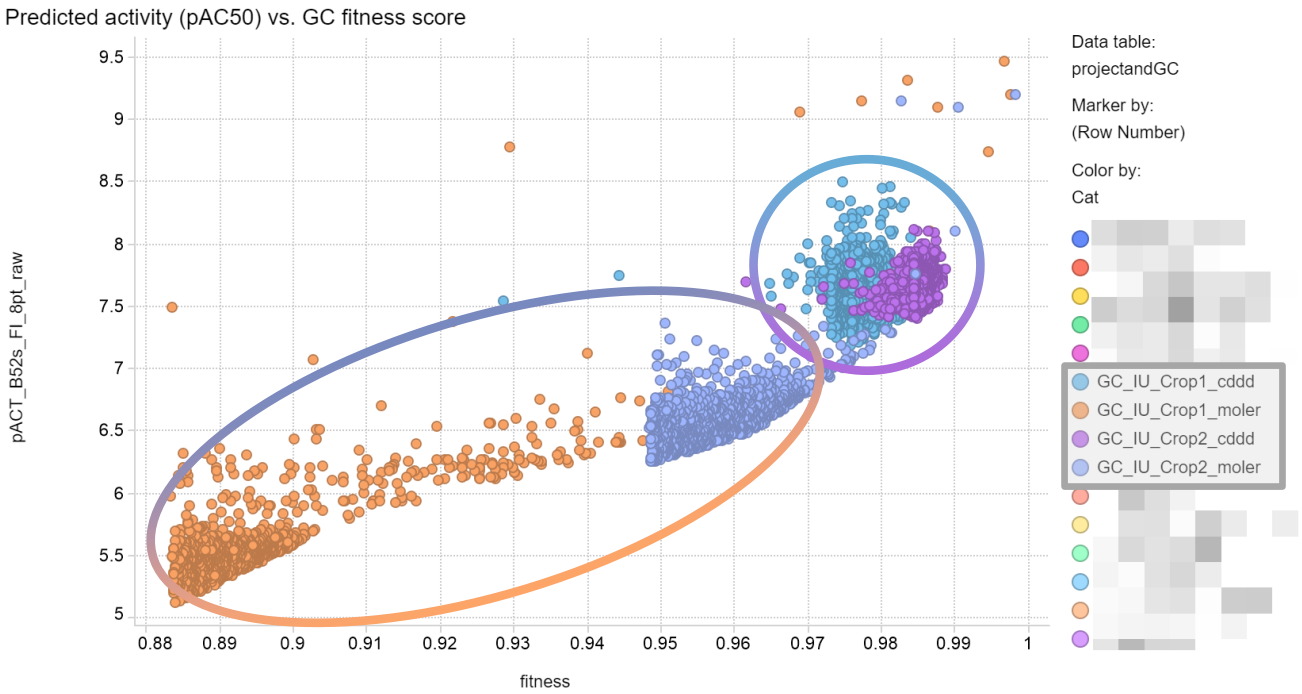




- Interesting exploitation of GenChem in areas between current series
- Different embeddings & settings provide different exploration profiles

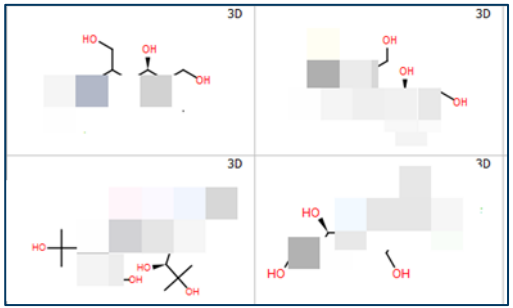
Project 2 – more observations

Higher fitness and activity score != more «realistic» molecules

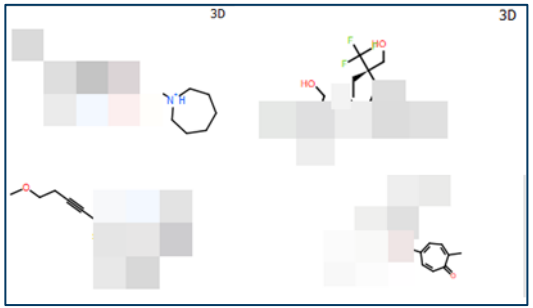


Project 3 – more observations

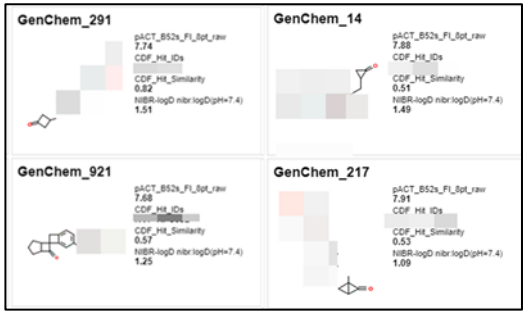
GenChem creates compounds (sometimes) outside our known chemistry space
→ 0 alert flags (internal SubStructure flags list) but still MedChemist «no-go's»



Repeating motifs (OH)



Gremlins (large rings, reactive, ugly, ...)



Very highly strained rings with ketone moiety

→ Iterative optimization of the GenChem workflow

GenChem - Take home messages

Manage expectations

Still in evolution

New scientific challenges

Good integration is key

- Do not expect magic but idea augmentation by ML methods
- Expecting surprises depends on the definition of your target property profile
- Generative methods are new and still in the evolution phase, we learn new ‘tricks’ with every project and every method
- Reliable property models are key in the MPO, uncertainty estimation is highly recommended to mitigate risk
- Diverse generators and optimization methods allow broader and complementary exploration of chemical space
- Seamless integration of GenChem in the daily project work will be key for success of the new method

Acknowledgements

Novartis

- Achim Plueckebaum
- **Aleksandr Kolodeev**
- Alexander Blaessle
- Alokesh Duttaroy
- Angela Mackay
- Ansgar Schuffenhauer
- Betrand Bodson
- **Bulent Kiziltan**
- Cara Brocklehurst
- Carlotta Caroli
- Chris Ball
- Chris Sarko
- Christian Kolter
- Clayton Springer
- Dave Barkan
- David Dunstan
- Dennis Koester
- Derek Lowe
- Dimitris Agrafiotis
- Eric Ma
- Eric Martin
- **Erik Anderson**
- **Finton Sirockin**
- Florentina Tofoleanu
- Florian Nigsch
- Gerardo Manfredi
- Gianluca Santarossa
- Gopichand Mukkapati
- Holger Hoefling
- **Hubert Misztela**
- **ID team**
- Imtiaz Hossain
- **Iya Khalil**
- Jacob Gora
- Jane Panteleev
- Jay Bradner
- Jeremy Jenkins
- Jeroen Verheijen
- **Jessica Lanini**
- Jose Duca
- Jürgen Wagner
- Karin Briner
- Kaspar Zimmermann
- Lee Sargeant
- Loic Giraud
- Lorenzo Romeo
- Marilisa Neri
- **Michal Pikusa**
- Natalie Dales
- **Nadine Schneider**
- **Nicholas Kelley**
- **Nikolas Fechner**
- **Nikolaus Stiefl**
- Noe Sturm
- Olivier Rene
- **Qurrat Ul Ain**
- Peter Ertl
- Peter Owotoki
- Peter Skewes-Cox
- Peter Speyer
- Rajesh Agrawal
- Riccardo Vianello
- **Richard Lewis**
- Richard Quinn
- **Rishi Gupta**
- Rosie Higgins
- Shahram Ebadollahi
- Simona Cotesta
- Srini Rao
- Steffen Renner
- Viktor Hornak
- **William Godinez**
- Xian Zhang
- Yen Liang Chen

Microsoft Research

- Alex Menezes
- Chris Bishop
- Henry Jackson-Flux
- Ian Kelly
- John Bronskill
- Junaid Bajwa
- **Krzysztof Maziarz**
- **Marc Brockschmidt**
- **Marwin Segler**
- Megan Stanley
- Nadzeya Paleyes
- **Pashmina Cameron**
- Sasa Juratovic

... and everyone else who supported us / contributed in our explorations

Thank you

NOVARTIS | Reimagining Medicine