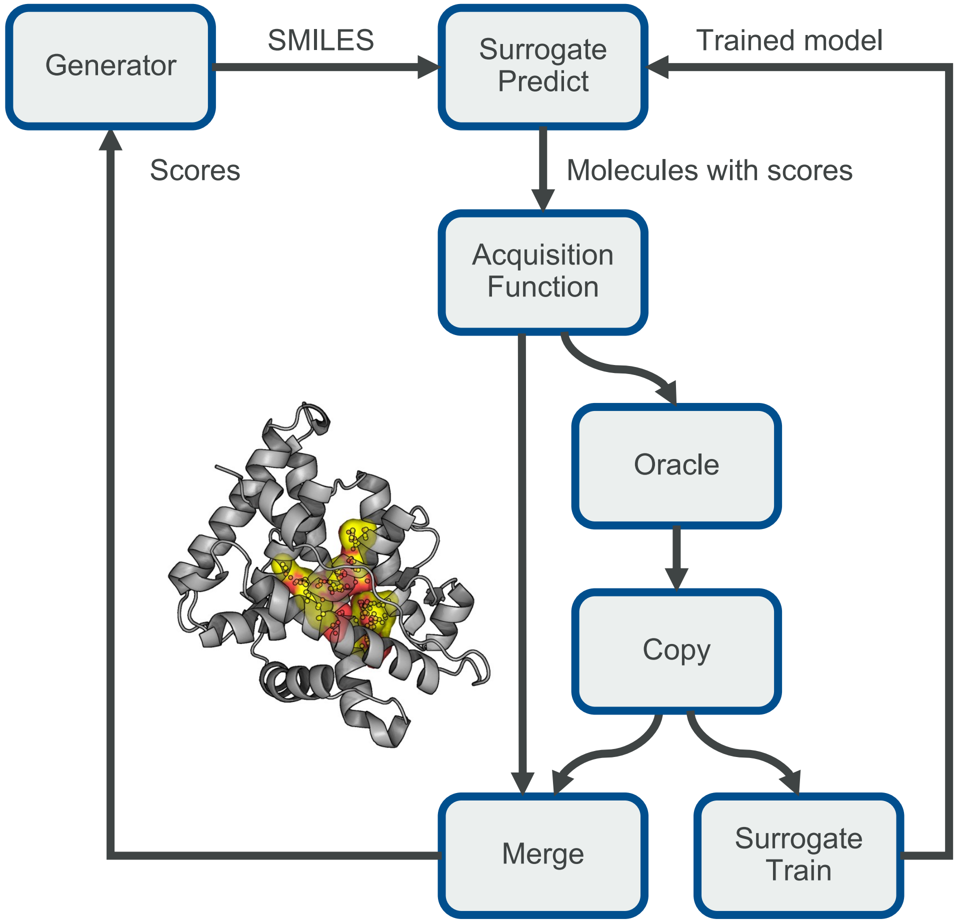
Computational chemistry workflows with Maize

Or: the importance of engineering in applied science



Thomas Löhr

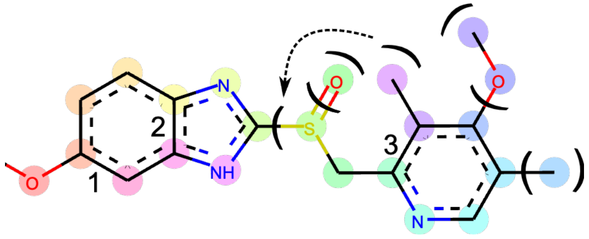
Senior Scientist @ Molecular AI, AstraZeneca

AiChemist School Berlin

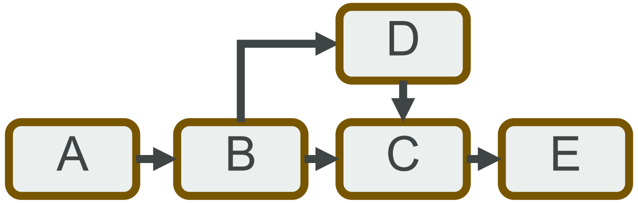
04/03/24

Overview

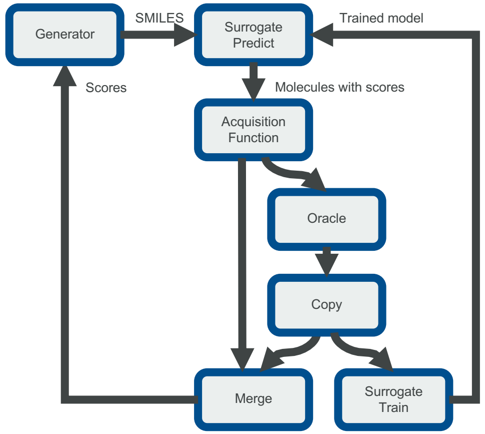
1. Early-stage drug discovery with generative models



2. The need for an advanced workflow manager



3. Applying Maize to active learning and other projects



Machine learning techniques are poised to impact pharmaceutical development industry from beginning to end

Early discovery

Target identification

Drug Design

Imaging/analysis

Scale-up/process

Clinical

Late/clinical development

Molecular AI

Ola Engkvist

Synthesis prediction

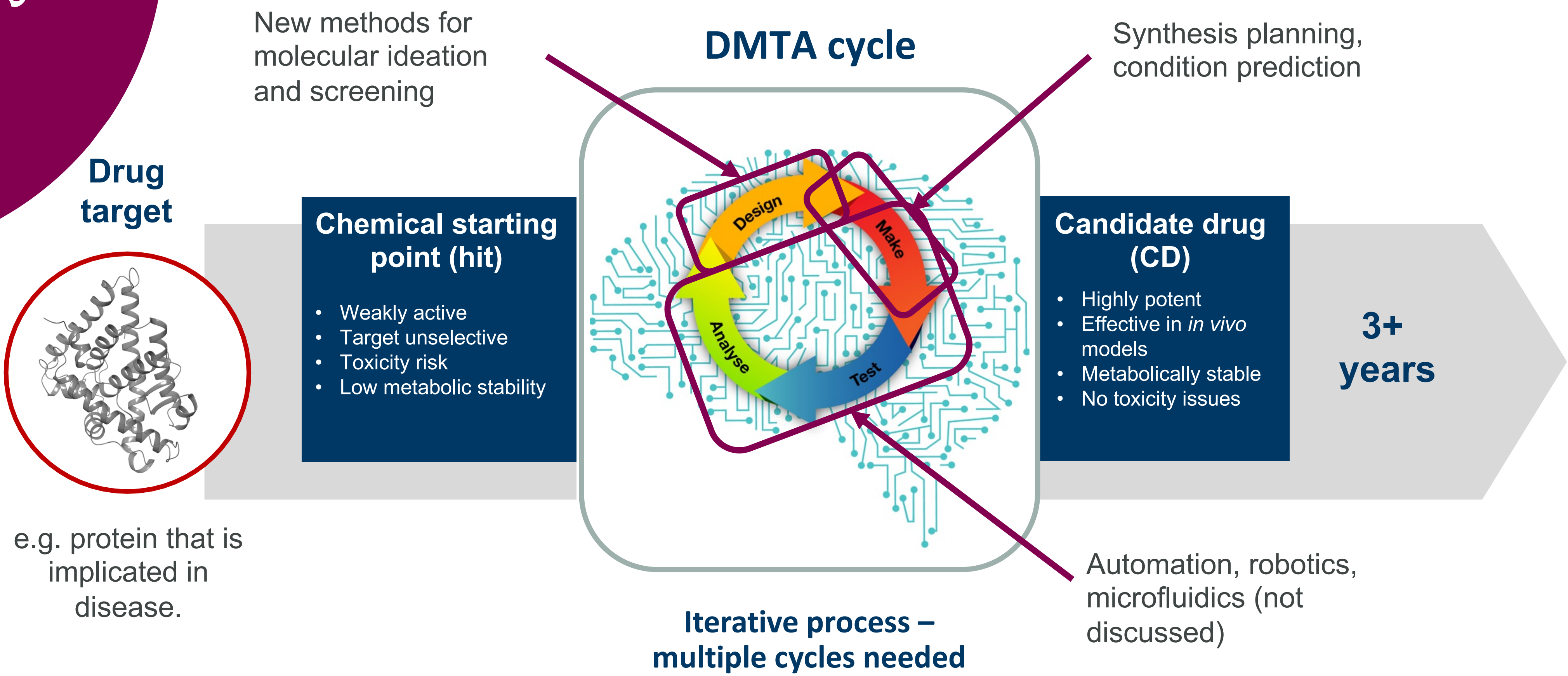
de novo molecular design

physics-informed molecular screening

The Drug Discovery Process

At the heart of the drug design process is the Design, Make, Test and Analyze (DMTA) cycle, which is a core concept for iterative, hypothesis driven design.

How can we accelerate this process?



New methods for molecular ideation and screening

DMTA cycle

Synthesis planning, condition prediction

Drug target

e.g. protein that is implicated in disease.

Chemical starting point (hit)

- Weakly active
- Target unselective
- Toxicity risk
- Low metabolic stability

Candidate drug (CD)

- Highly potent
- Effective in *in vivo* models
- Metabolically stable
- No toxicity issues

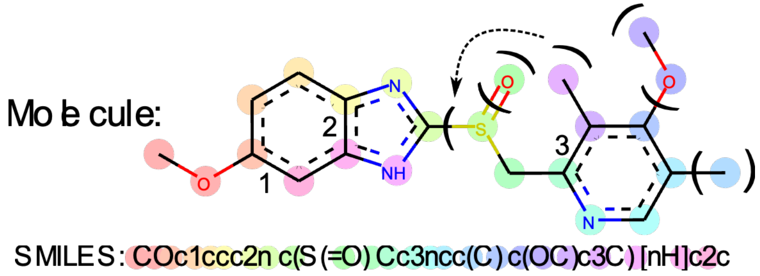
3+ years

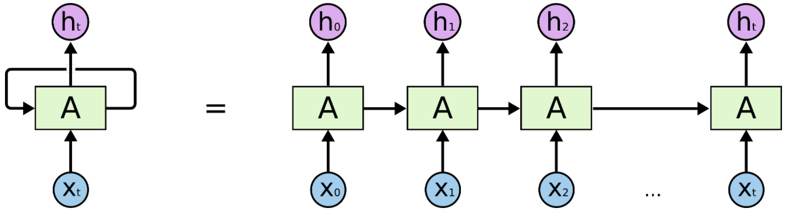
Iterative process – multiple cycles needed

Automation, robotics, microfluidics (not discussed)

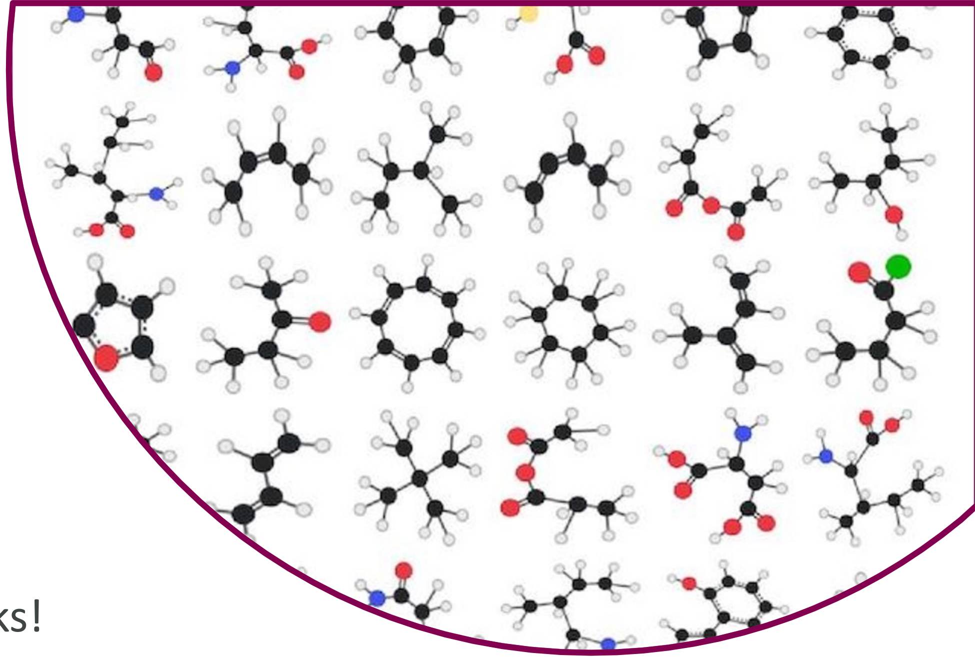
Chemical language models are central to much of our work

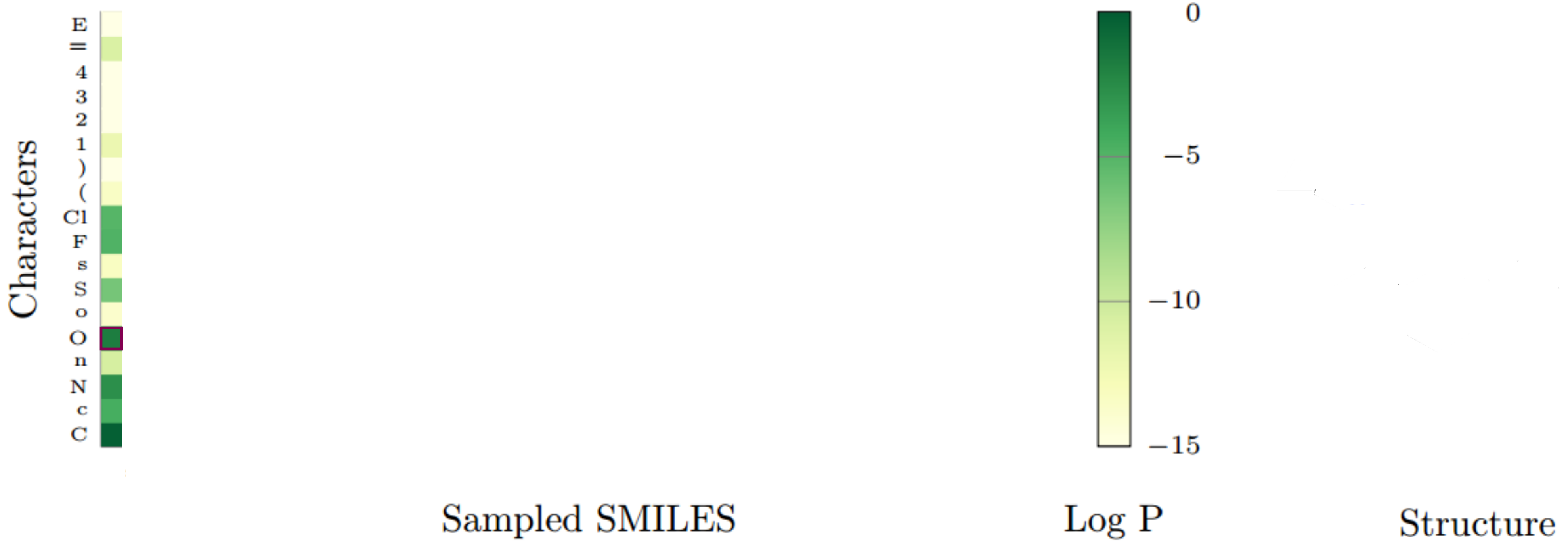
Molecules can be described in the language of SMILES…





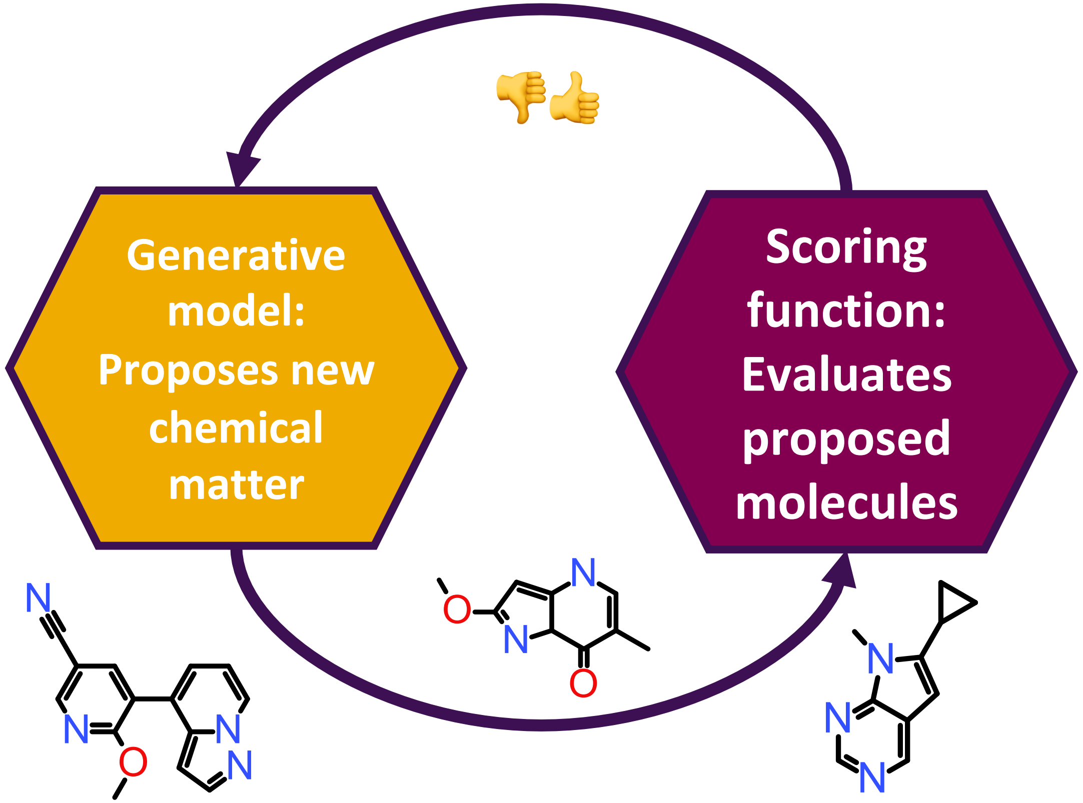
and fed into (recurrent) neural networks!





REINVENT – designing new molecules with AI

REINVENT is the in-house developed de novo molecular design tool, using generative reinforcement learning to solve *in silico* molecular design tasks


Generative model: Proposes new chemical matter

Scoring function: Evaluates proposed molecules

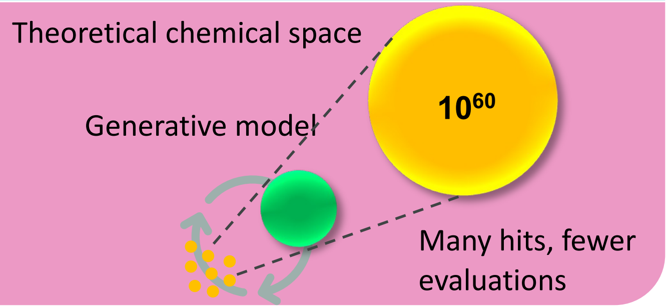
Traditional approaches rely on searching a large database for a small number of suitable hits

10^9

Searching for a needle in a modest haystack

Few hits, many evaluations

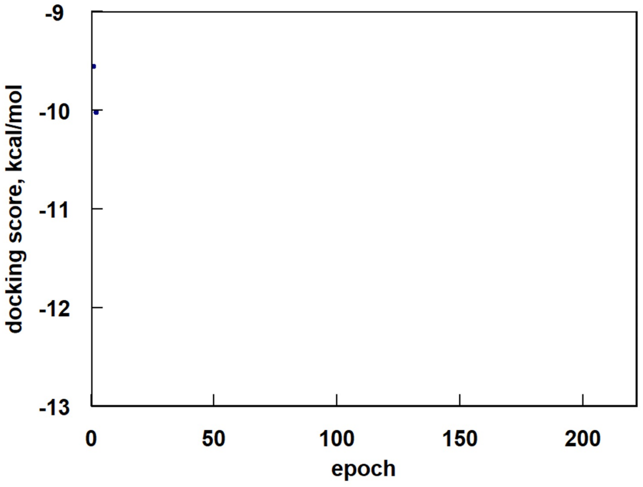
Generative models encode practically unlimited chemical space probabilistically

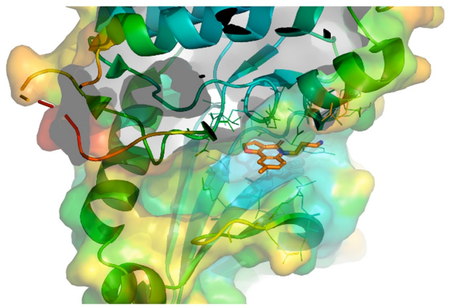

Theoretical chemical space

10^{60}

Generative model

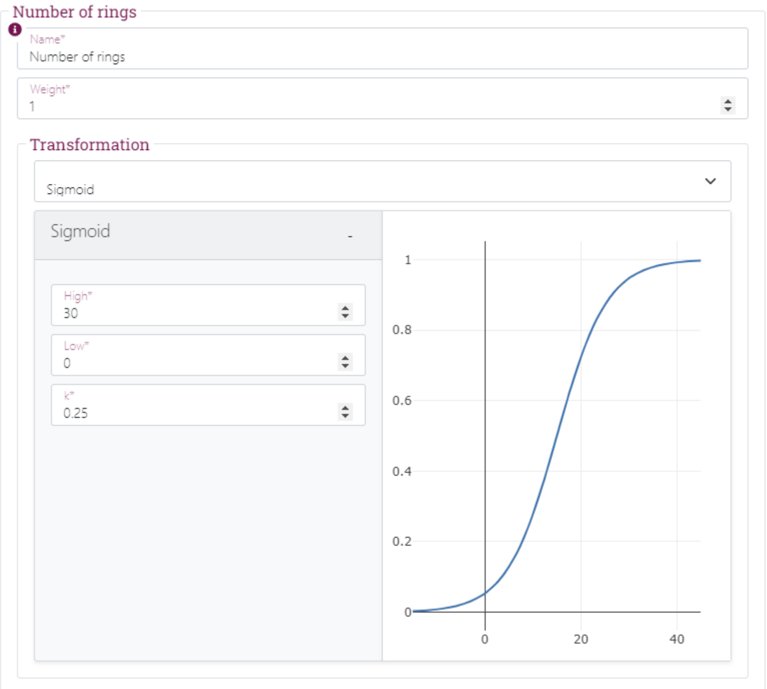
Many hits, fewer evaluations





Superpowered molecular optimization engines

Reinvent agents exhibit remarkable plasticity wrt prior and retain adaptability after 100s of epochs. E.g. spend ~800 epochs learning to make as many rings as possible…



Number of rings

Name* Number of rings

Weight* 1

Transformation

Sigmoid

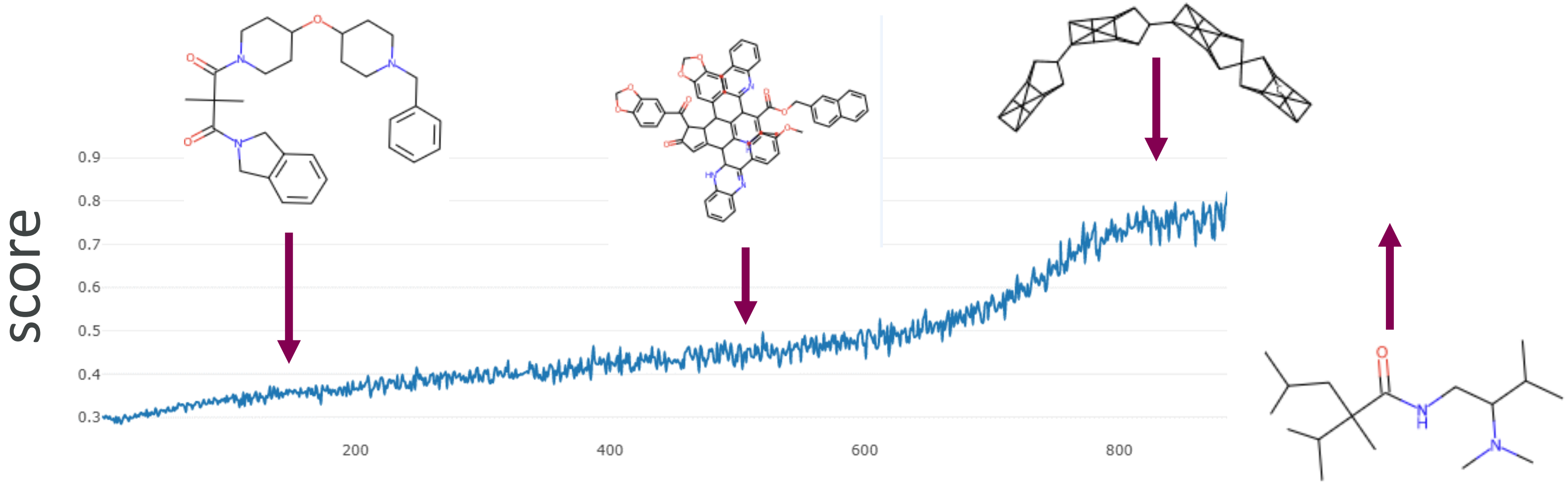
Sigmoid -

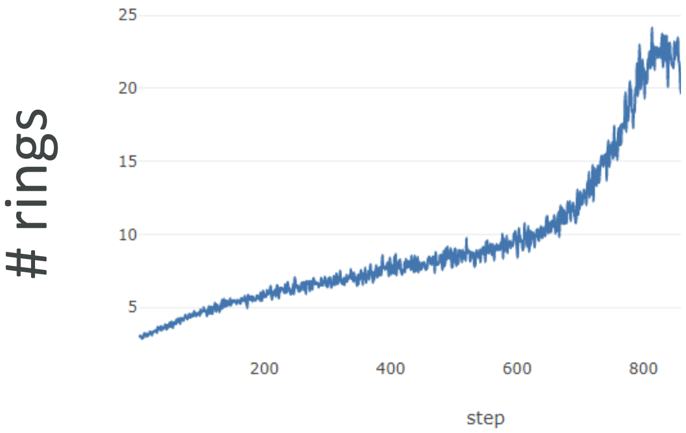
High* 30

Low* 0

k* 0.25

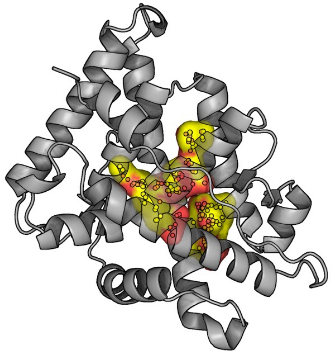
Then reverse score transform!



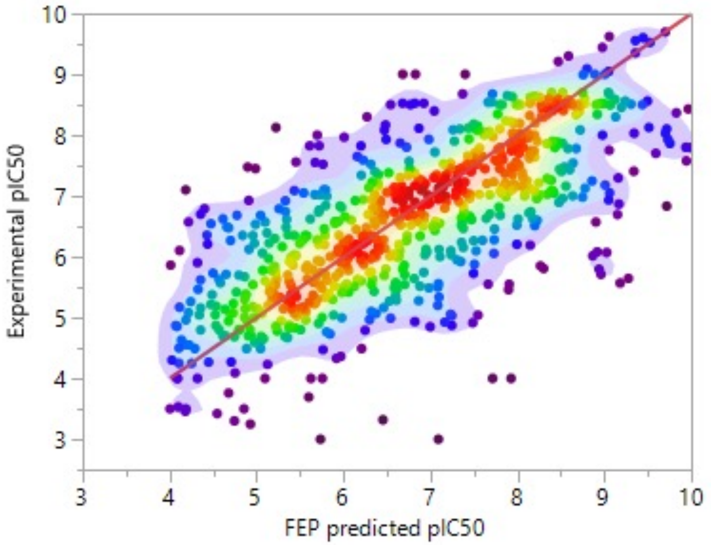


Combining generative models & state-of-the-art simulation

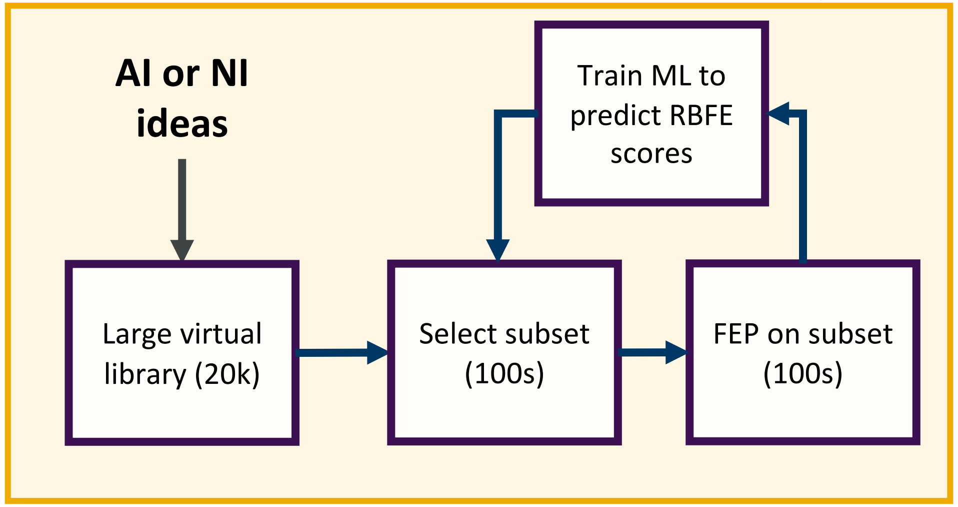
Proteins are dynamic. This is important for ligand binding.



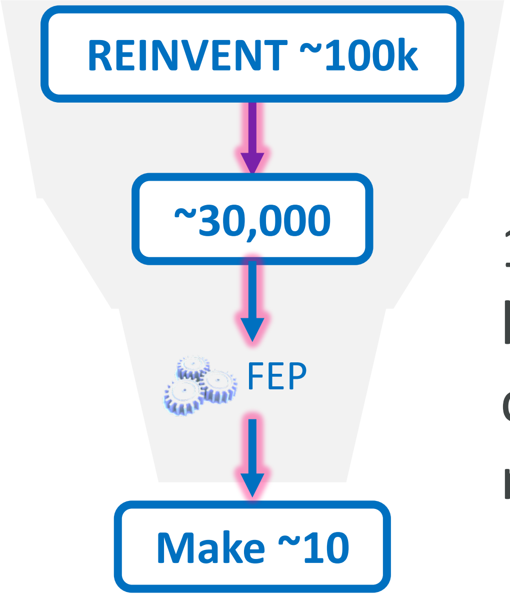
Relative binding free energy (RBFE) calculations are an advanced, computationally expensive but accurate way to predict potency of new compounds using molecular dynamics . Validated over 16 targets, 15k compounds at AZ over 3+ years.



Active learning RBFE



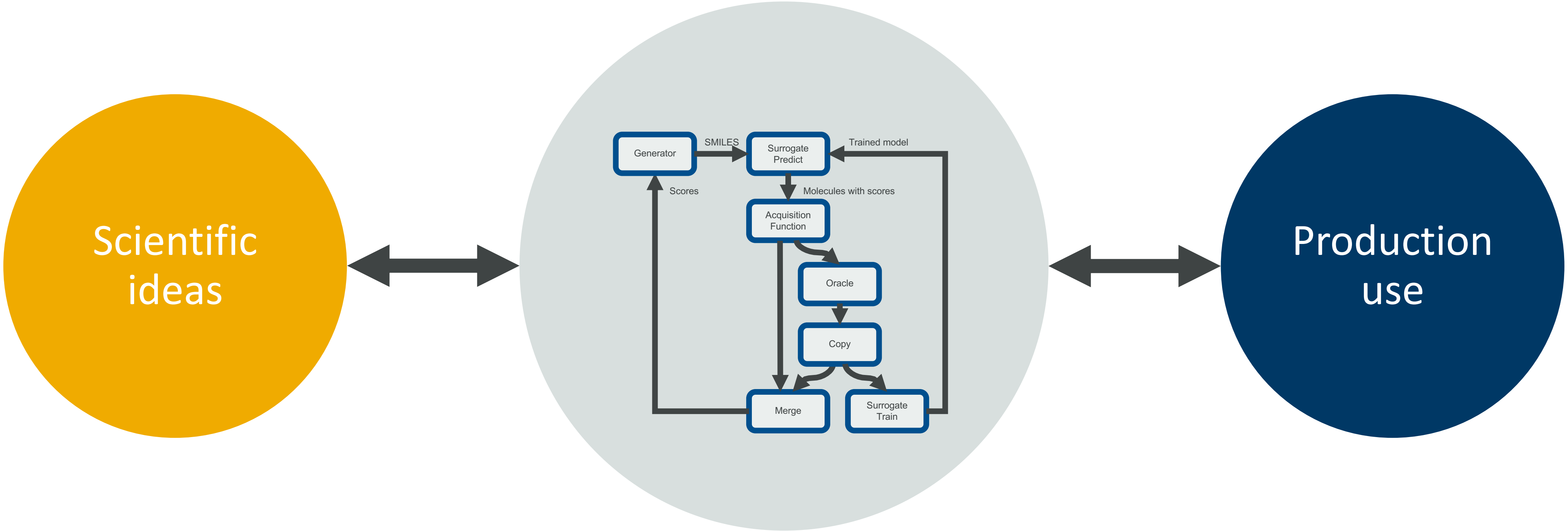
We combine RBFE with active learning to accelerate to larger scales



100s of GPU-hours per compound made!

How do we put this together?

Workflow managers as a tool for abstraction

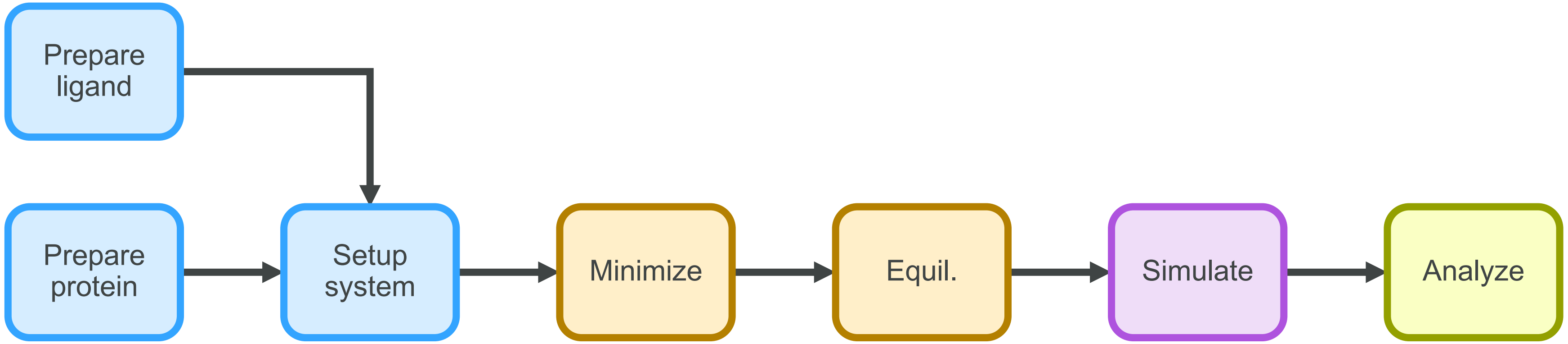


Scientific ideas

Production use

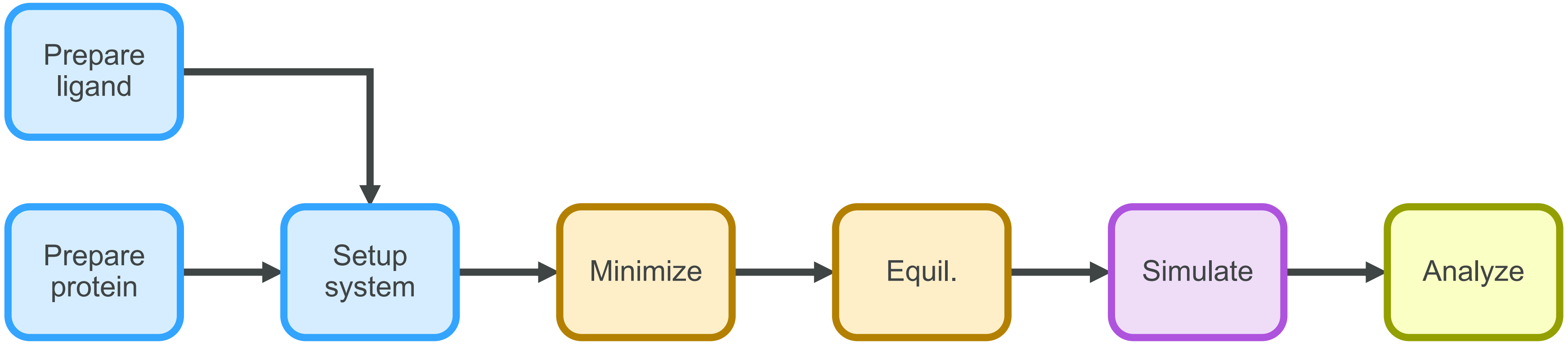
Why use a workflow manager at all?

- **Reproducibility** – not just for others, but for yourself too!
- **Configuration** – no searching through shell history for used parameters
- **Modularization** – easily exchange components to make experiments easier
- **Automation** – easier to integrate into routine systems
- **Abstraction** – components can be thought of as black boxes
- **Performance** – some workflows allow parallelization



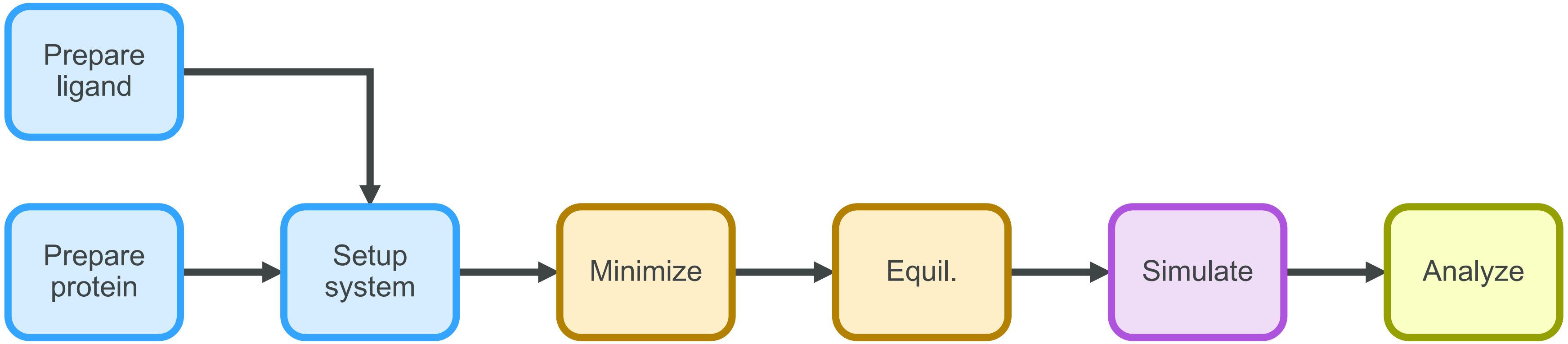
Workflows in computational chemistry

Typical protein + ligand simulation workflow

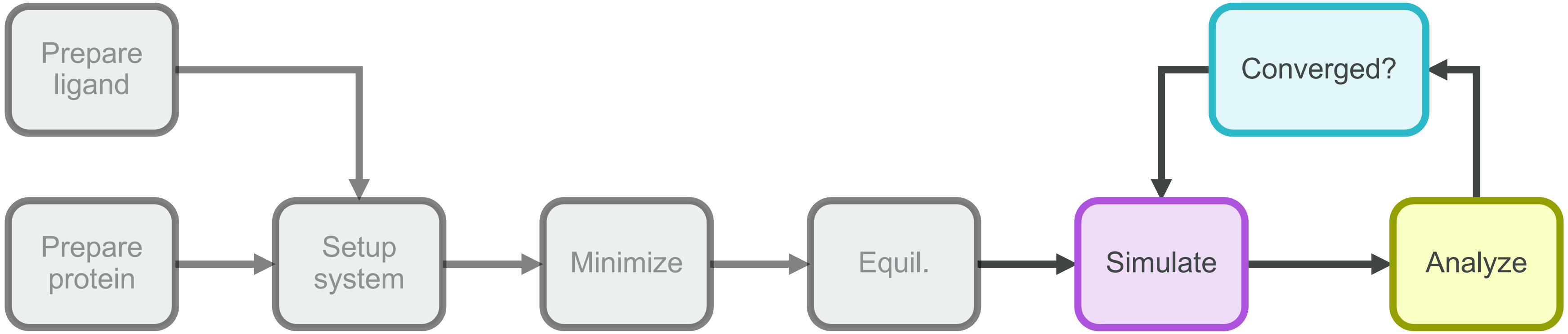


Workflows in computational chemistry

Typical protein + ligand simulation workflow

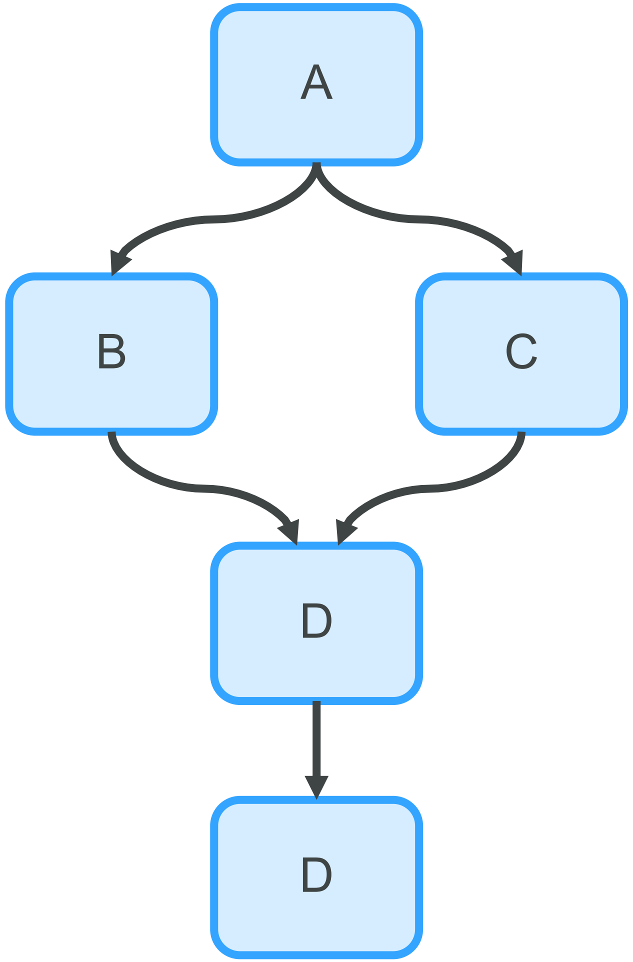


Problem: we actually want to run in a cycle!

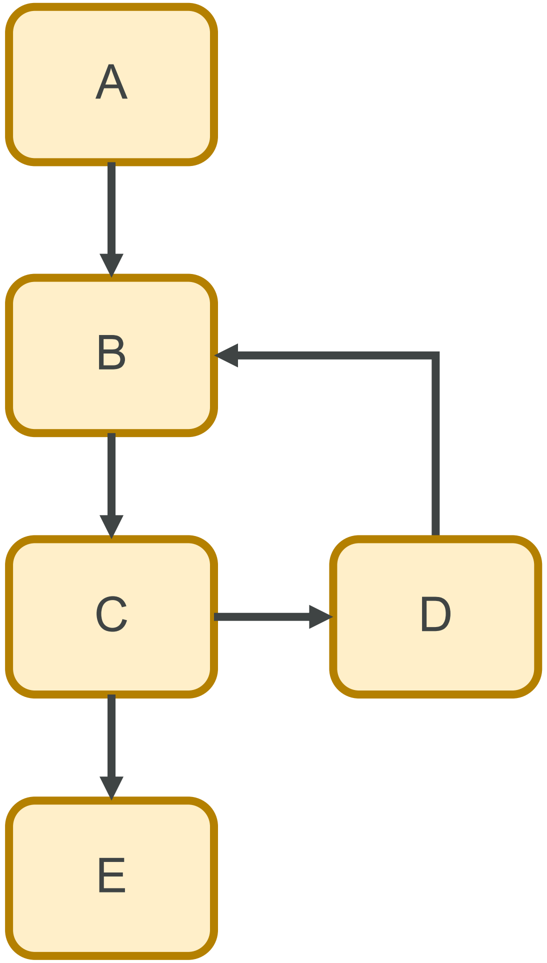


Complex workflows: cycles and conditionals

Directed Acyclic Graph (DAG)



Directed Cyclic Graph (DCG)



DAGs are the basis for most workflow engines:

- Apache Airflow
- Luigi
- DAGster
- …

DCGs allow iteration, control flow, but have limited support:

- Akka (Scala)
- NoFlo (JavaScript)

Workflow manager wishlist

🐍 **Reproducibility** – Simple and portable workflow definitions in Python

🔧 **Configuration** – Separation of system and workflow configuration

🧱 **Modularization** – Workflow nodes with well-defined I/O, easy to share

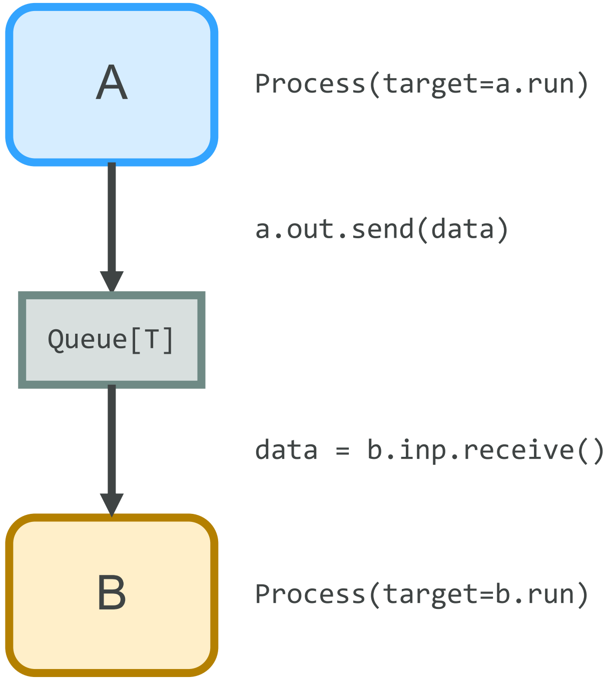
🚗 **Automation** – Flexible execution: conditionals, cycles, use in *Jupyter*

🎨 **Abstraction** – Grouping of nodes into subgraphs

📈 **Performance** – Parallelization by default

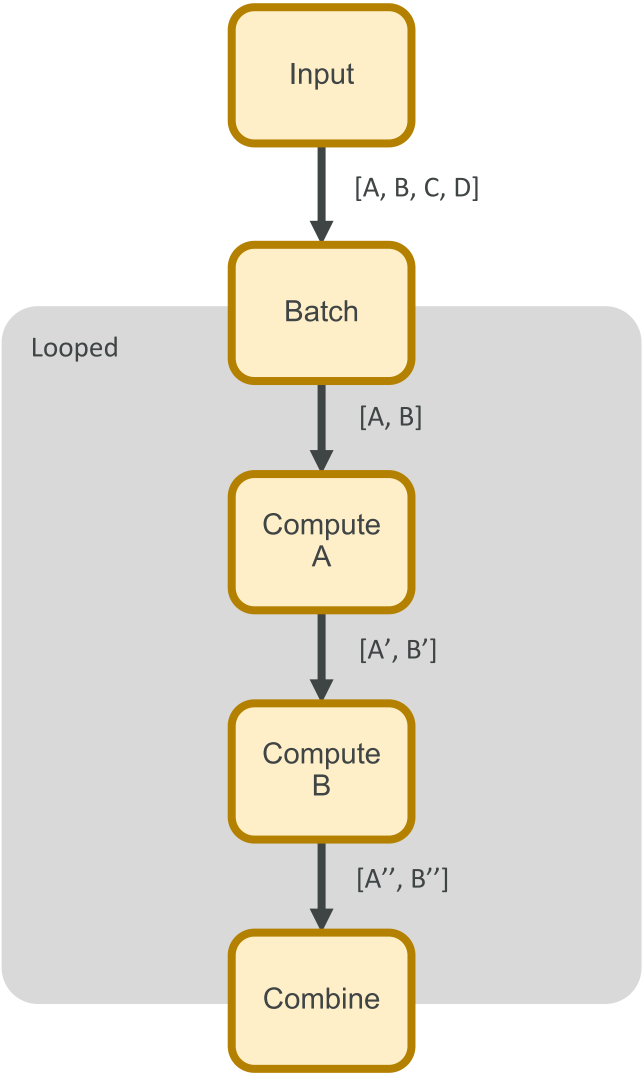
Flow-based programming with Maize

- Written in Python with concepts of flow-based programming
- Each node is isolated and has type-safe, well-defined I/O and its own environment
- Each node runs in a separate process and can receive / send data at any time
- Parallelization, batch processing, load-balancing for free!

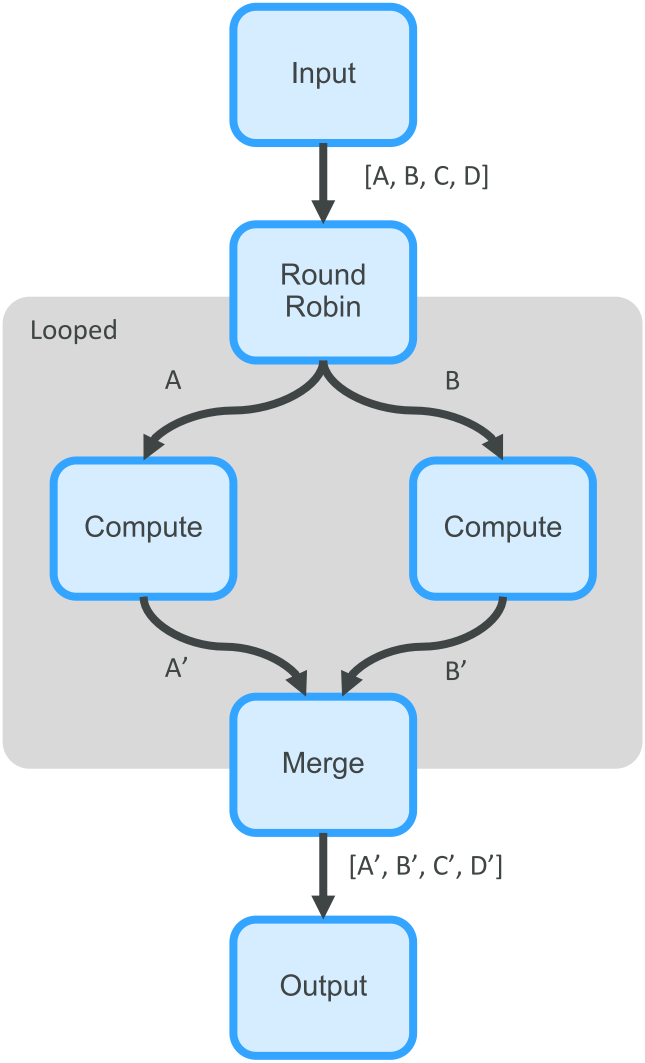


Common workflow patterns

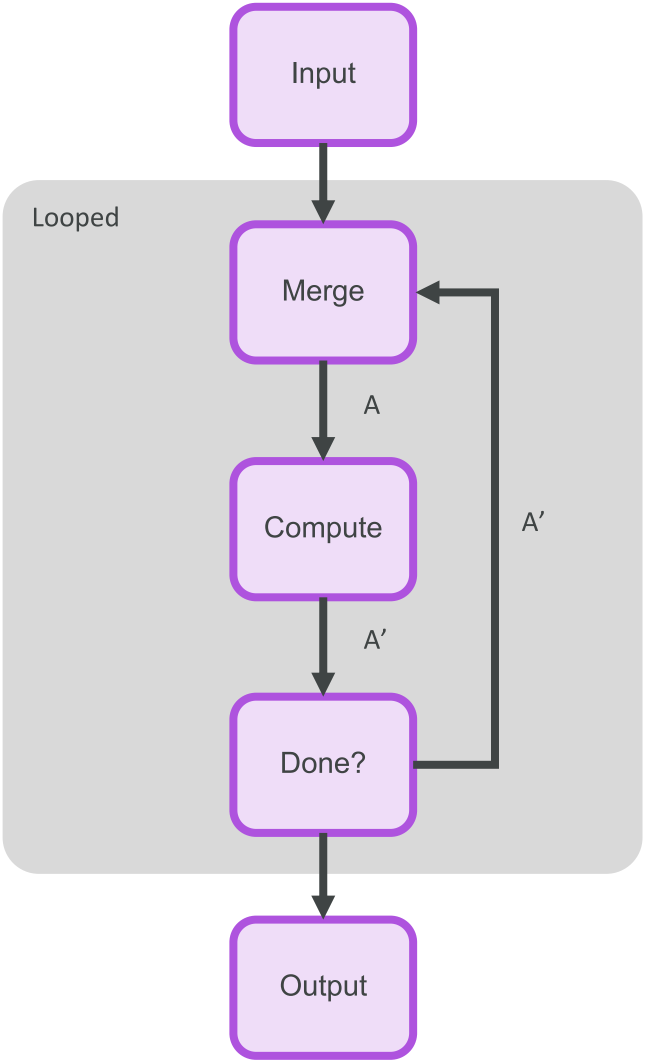
Batch processing



Parallelization



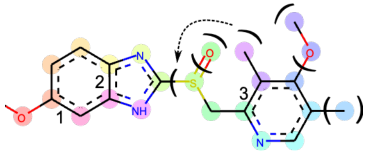
Iteration



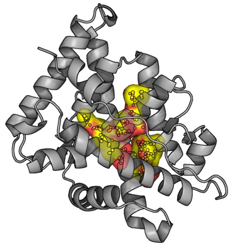
Implemented software & scope

Reinforcement learning

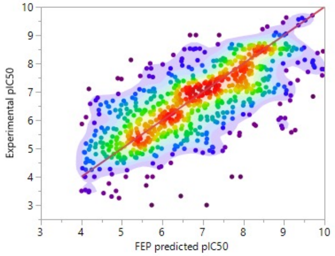
Small molecule generation



REINVENT



Small molecule docking:
AutoDockGPU, **AutoDock Vina, Schrödinger Glide**



Binding free energy:
OpenFE RBFE

Utilities:

- Molecule I/O
- Filtering
- Embedding
- Conformer generation
- Docking grid preparation
- Shape matching
- RMSD
- Active learning

Post-processing

- Molecular dynamics: **Gromacs MM/PBSA**
- Semi-empirical: **Crest, XTB**
- Quantum: **Gaussian**

Coming soon:

- **Absolute BFEs**
- **GNINA**
- **And more!**

Many more options:

- Data filters
- Switches
- Caching
- Data merging / splitting
- Lambdas
- File IO

Application to active scientific projects at AZ

Molecular dynamics simulations (MDs) in Maize

Lili Cao

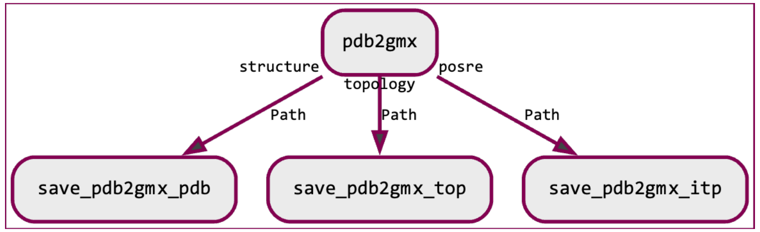
GROMACS

- Topology and Force Field
  - gmx **pdb2gmx**
    gmx **acpype**
- Solvation and Ionization
  - gmx **editconf**
    gmx **solvate**
    gmx **genion**
- Energy minimization
  Equilibration (NVT, NPT)
  Production
  - gmx **grompp**
    gmx **mdrun**
- Analysis

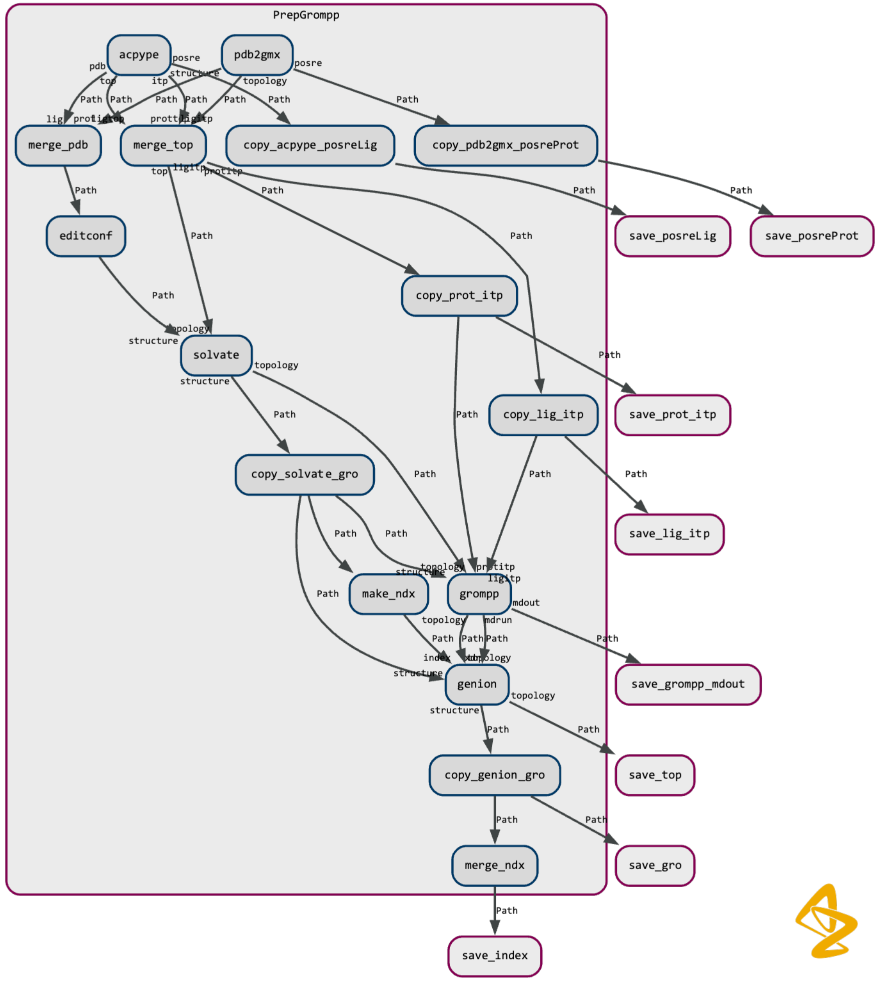
Maize

- **gmx commands** have been integrated into **Maize** as workflow **nodes**
- Example workflows
  - Each node
  - Connected nodes

Example: pdb2gmx



Example: Everything you need before grompp and mdrun

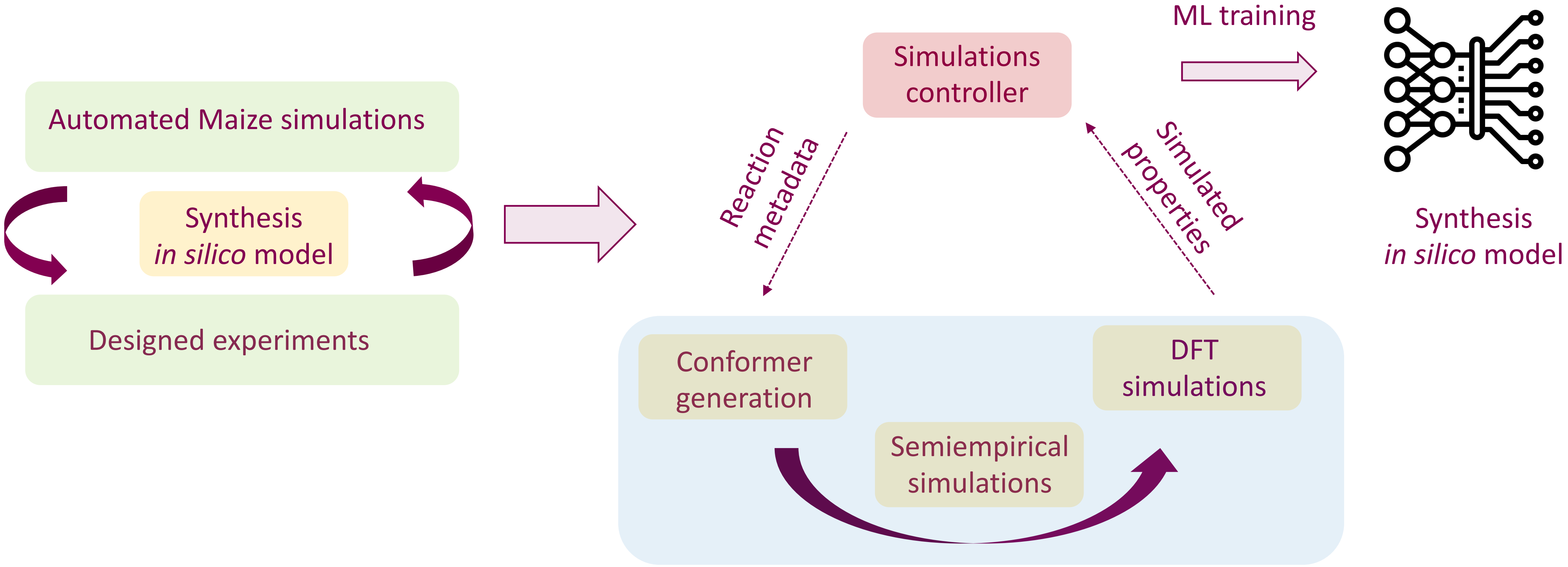


Building predictive synthesis models

Michele Assante

Mikhail Kabeshov

- Maize fully automated workflow allows building ML model from *in silico* generated QM features



Automated Maize simulations

Synthesis *in silico* model

Designed experiments

Simulations controller

Reaction metadata

Simulated properties

Conformer generation

Semiempirical simulations

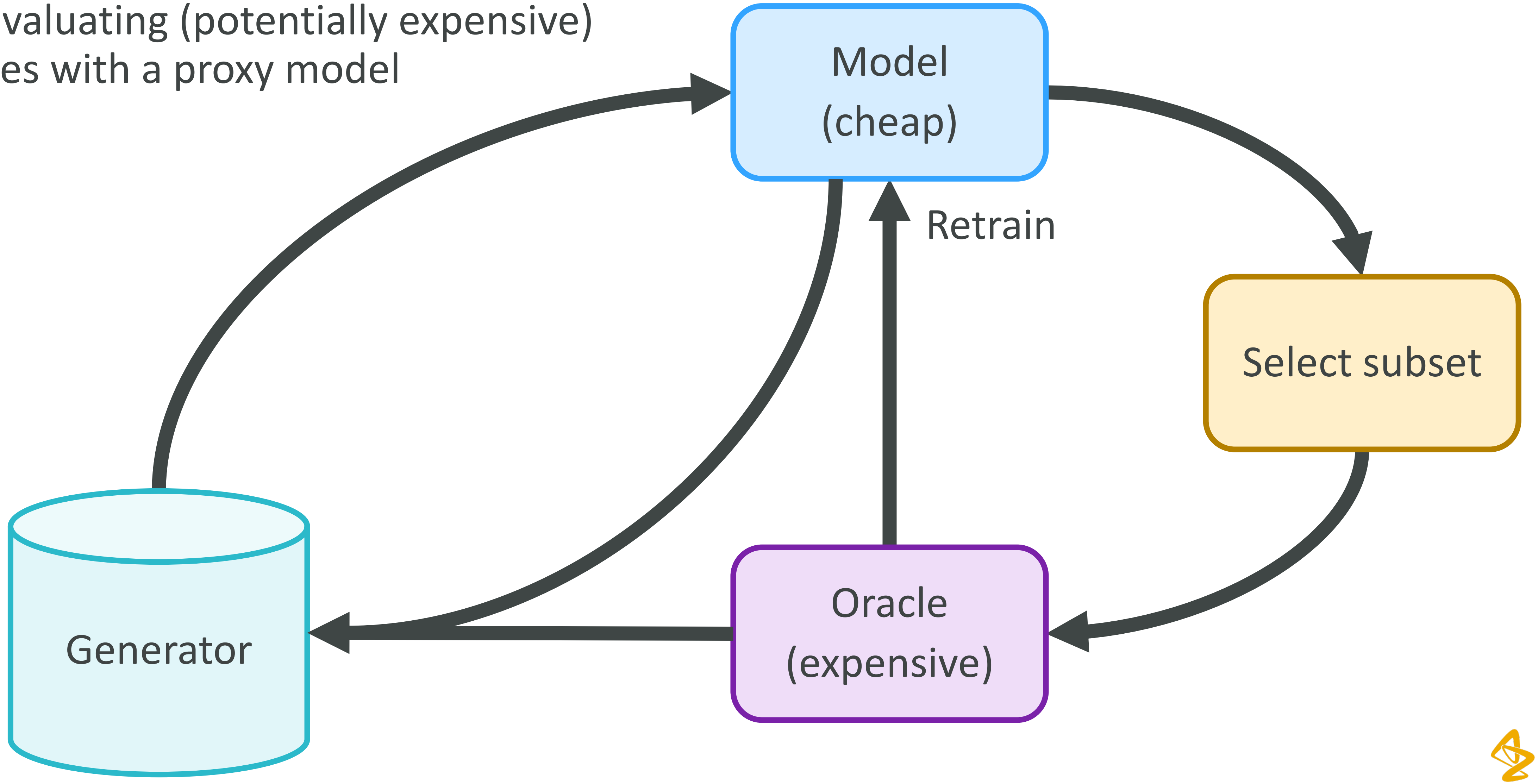
DFT simulations

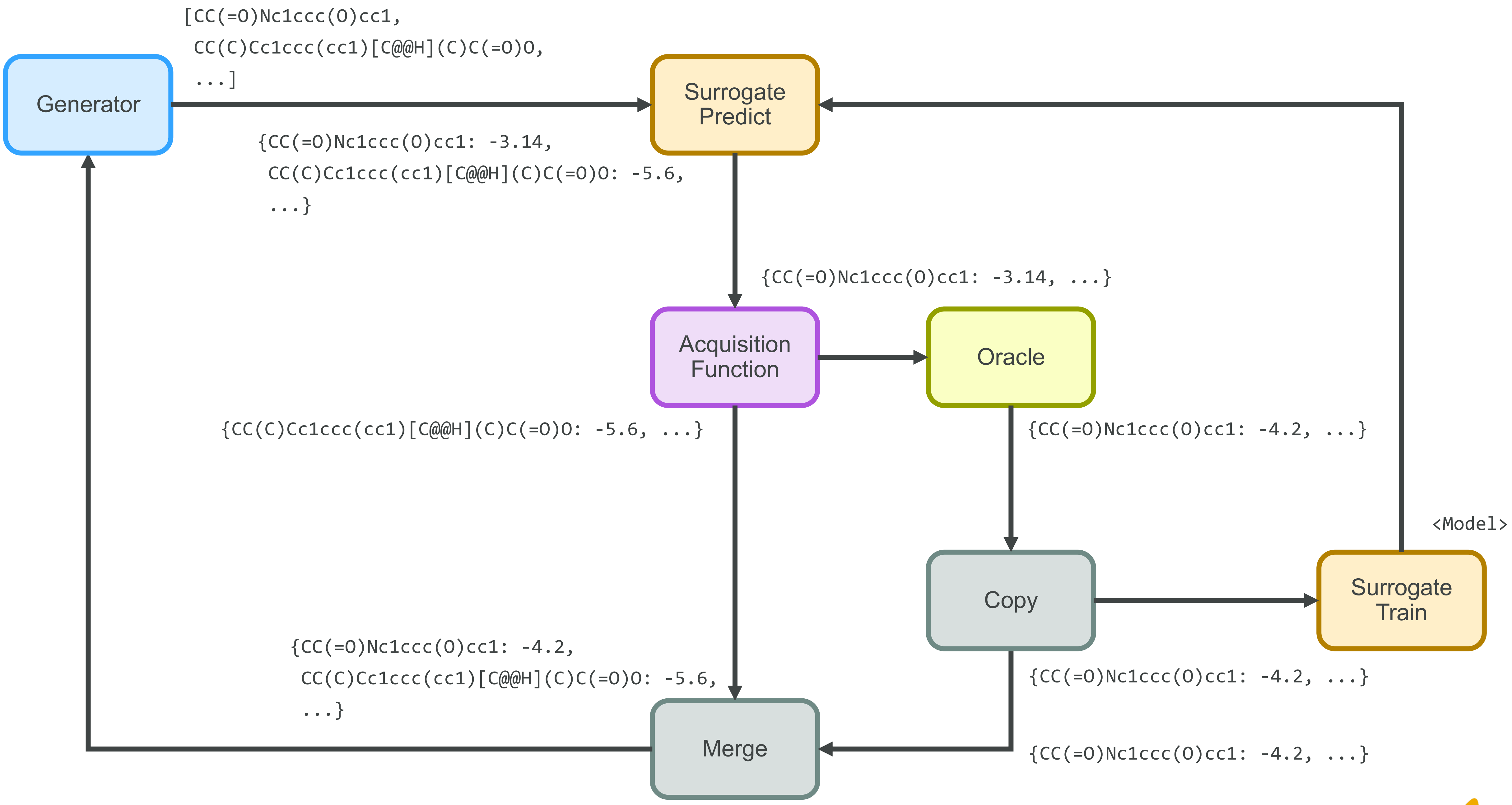
ML training

Synthesis *in silico* model

Active learning

Allows evaluating (potentially expensive) properties with a proxy model





Generator

[CC(=O)Nc1ccc(O)cc1,
CC(C)Cc1ccc(cc1)[C@@H](C)C(=O)O,
...]

Surrogate Predict

{CC(=O)Nc1ccc(O)cc1: -3.14,
CC(C)Cc1ccc(cc1)[C@@H](C)C(=O)O: -5.6,
...}

Acquisition Function

{CC(=O)Nc1ccc(O)cc1: -3.14, ...}

Oracle

{CC(C)Cc1ccc(cc1)[C@@H](C)C(=O)O: -5.6, ...}

{CC(=O)Nc1ccc(O)cc1: -4.2, ...}

<Model>

Copy

Surrogate Train

{CC(=O)Nc1ccc(O)cc1: -4.2,
CC(C)Cc1ccc(cc1)[C@@H](C)C(=O)O: -5.6,
...}

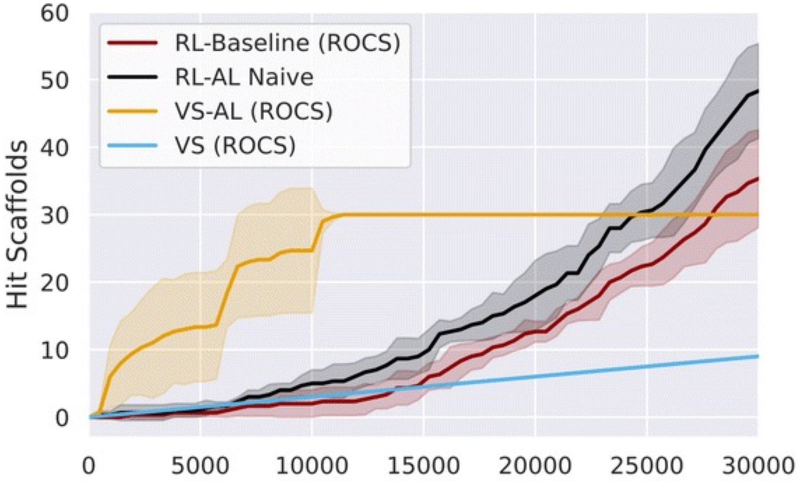
{CC(=O)Nc1ccc(O)cc1: -4.2, ...}

Merge

{CC(=O)Nc1ccc(O)cc1: -4.2, ...}

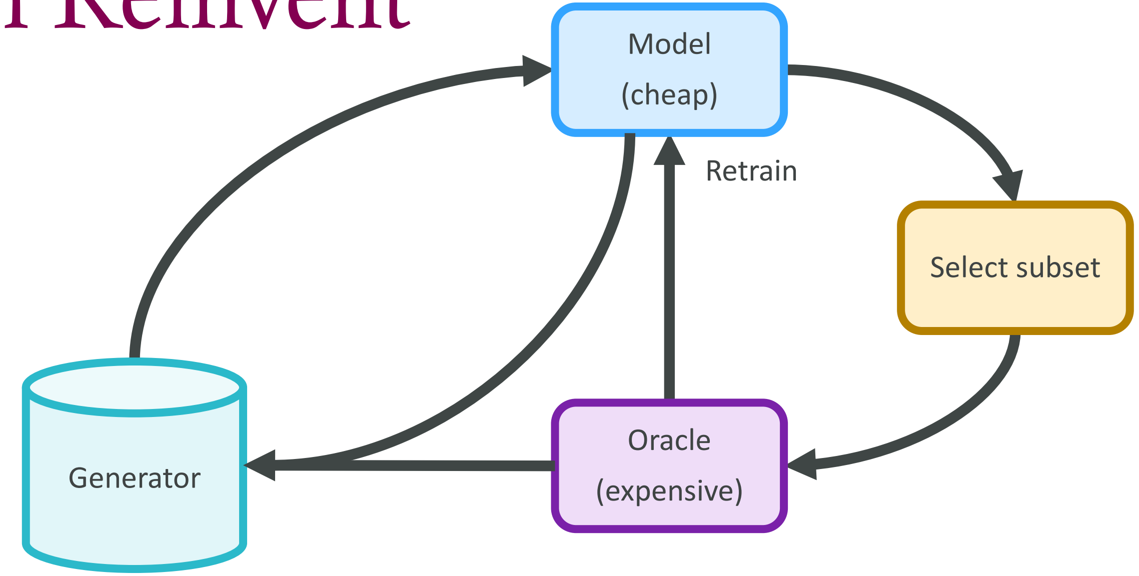
Free lunch: Active learning with Reinvent

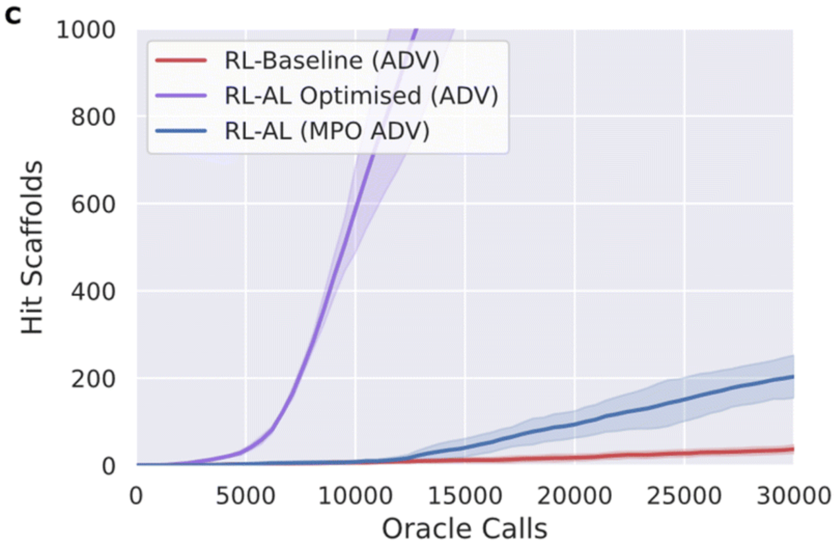
COX2, REINVENT + ROCs (PoC)



VS: fixed library

RL: REINVENT





ADV: AutoDock Vina

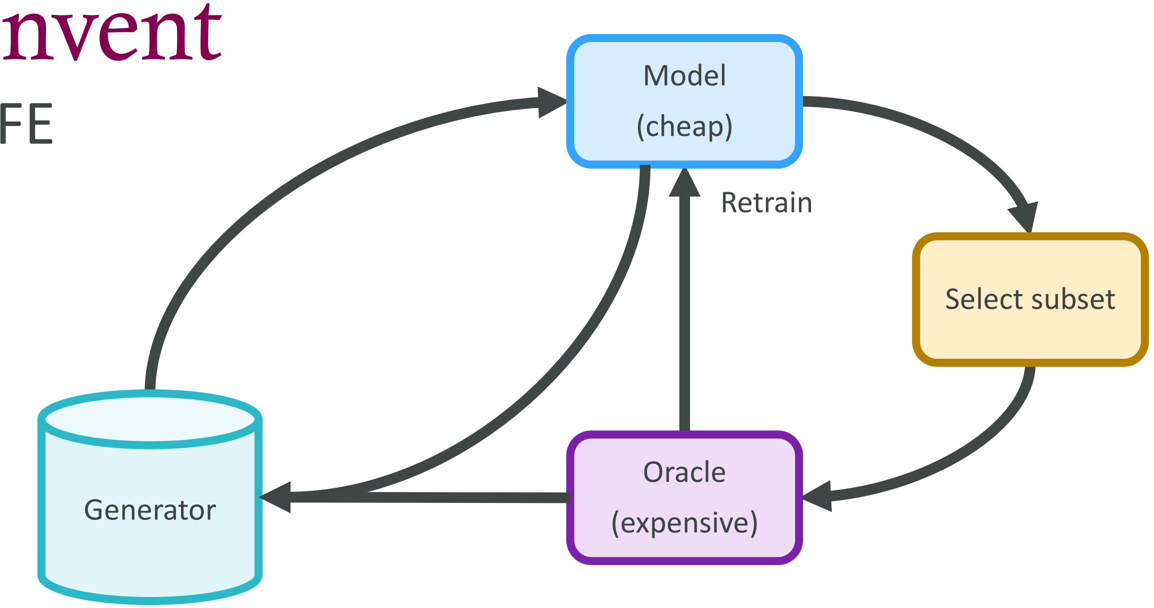
Michael Dodds

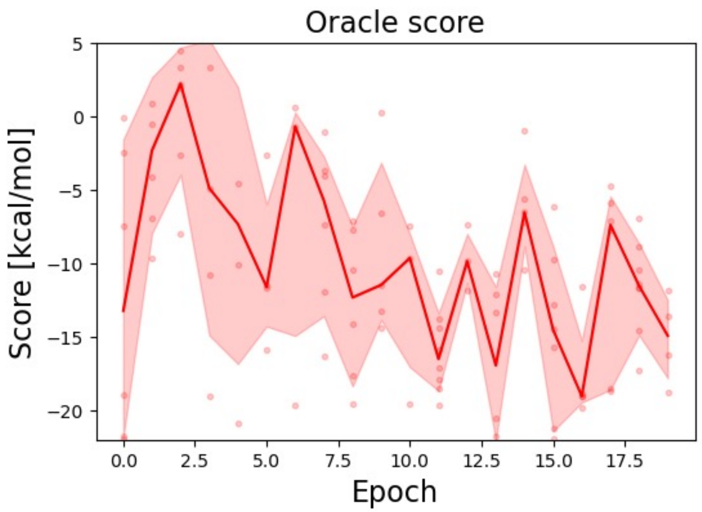
Jon-Paul Janet

Dodds, M. *et al. Chemical Science* (2024)

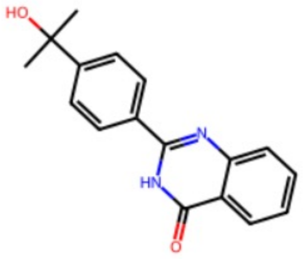
Active learning RBFE with Reinvent

Tankyrase (TNKS), LibInvent / Molformer + OpenFE

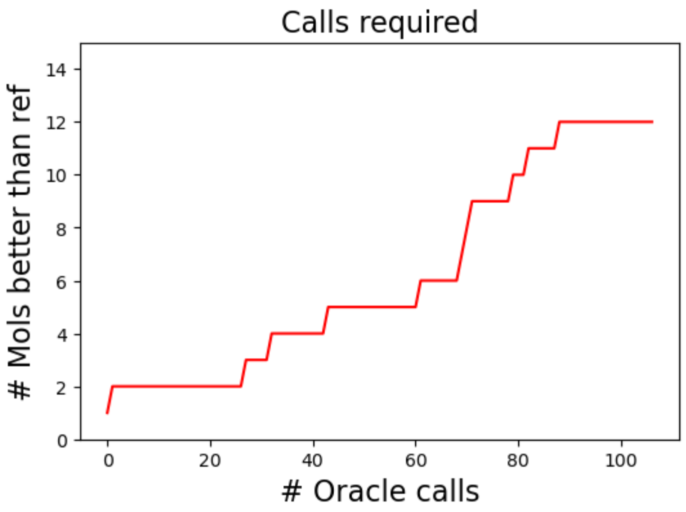



Oracle score

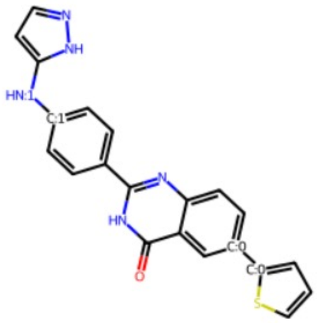
Reference



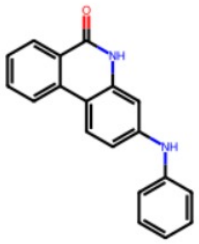
5a: -10.7531 kcal/mol


Calls required

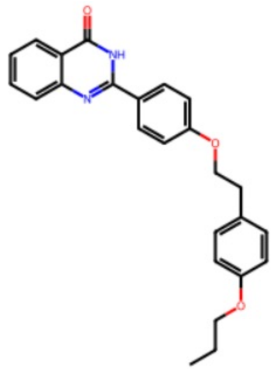
Examples generated



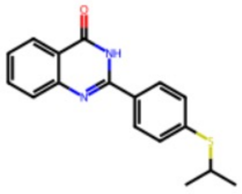
-13.6835 kcal/mol



-6.9984 kcal/mol



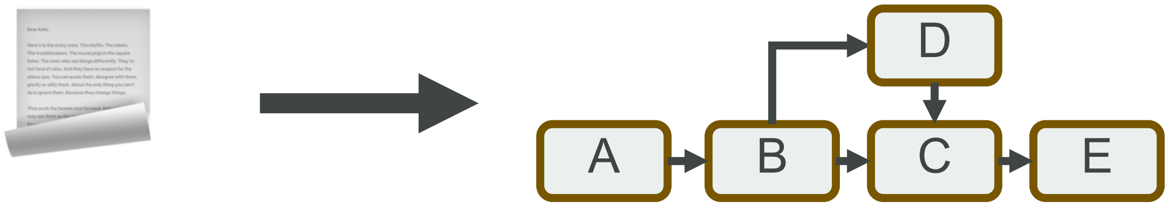
-17.4148 kcal/mol



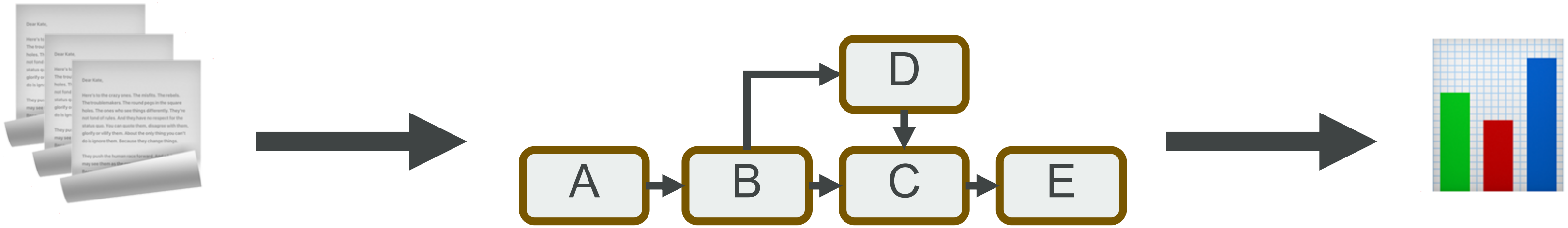
-15.9591 kcal/mol

The perks of being a workflow

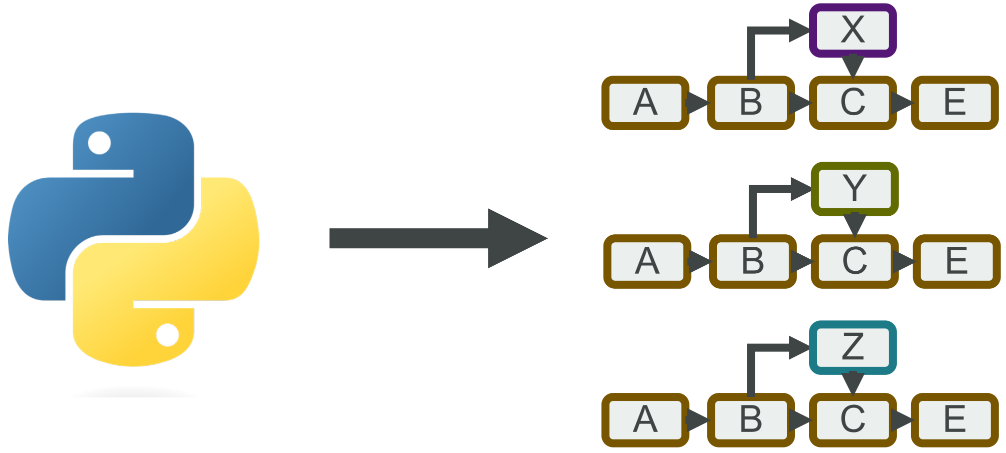
Going from serial format to executable workflows



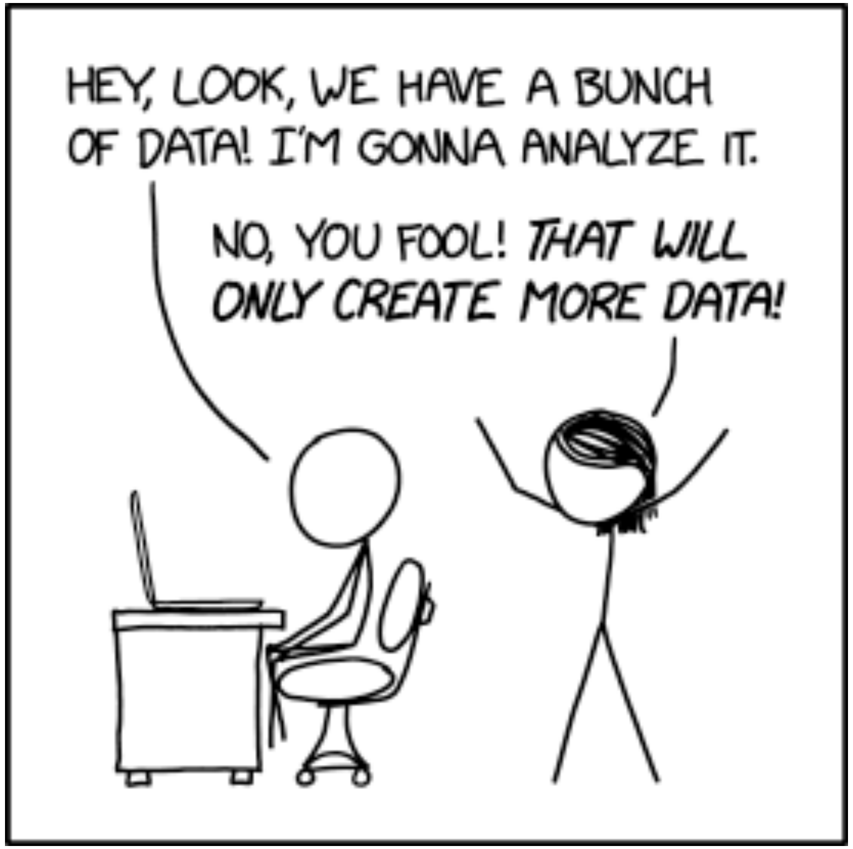
Allows automated mass execution with custom parameters & topologies



Dynamic workflow creation



Who is Maize for?


HEY, LOOK, WE HAVE A BUNCH OF DATA! I'M GONNA ANALYZE IT.
NO, YOU FOOL! THAT WILL ONLY CREATE MORE DATA!

You might be interested if you…

- Want to abstract away arcane software
- Have circular / conditional workflows
- Have awkward parallelization requirements
- Quickly iterate through different parameters

Get in touch:

thomas.lohr@astrazeneca.com

Try it out:

https://github.com/MolecularAI/maize

https://github.com/MolecularAI/maize-contrib

Acknowledgements

Michael Dodds, Lili Cao, Jon-Paul Janet, Mikhail Kabeshov, Michele Assante, Marco Klähn, Ola Engkvist

Introduction slides: Jon-Paul Janet

Confidentiality Notice