Lead Optimization Design Cycle in Orion

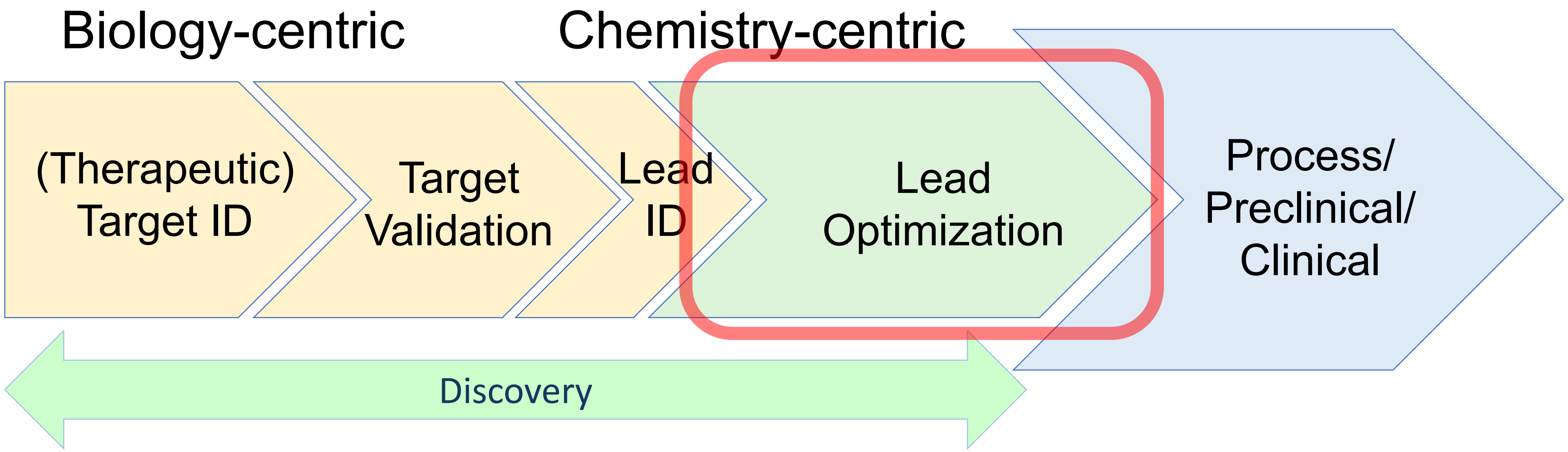
Matt Geballe
Head of Cheminformatics & Data Science
CUP 2023

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Agenda

- Lead Optimization
- Generating Synthetically Accessible Candidates
- Machine Learning and Lead Opt

Drug Discovery Process



J. Am. Chem. Soc. 2016, 138, 43, 14257–14263

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

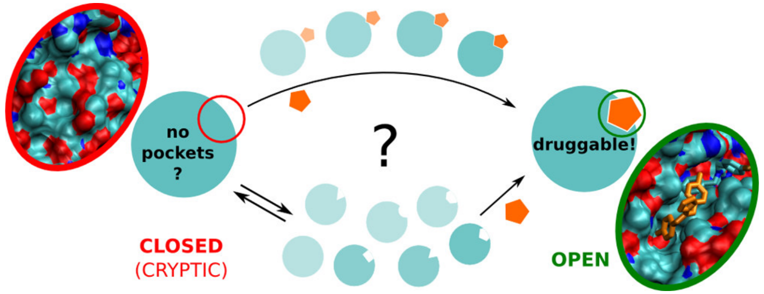
Drug Discovery Process

Biology-centric | Chemistry-centric

(Therapeutic) Target ID → Target Validation → Lead ID → Lead Optimization → Process/ Preclinical/ Clinical

Discovery

Cryptic pocket detection



2D Ligand-Based

Screening by 2D structure and/or fingerprint (GraphSim TK)

3D Ligand-Based

Search billions of molecules in minutes. Extremely fast screening by shape and chemical-features similarity (ROCS®, FastROCS™, FastROCS™ Plus)

3D Structure-Based

Dock billions of molecules with Gigadock™, for the most rigorous results, and with Gigadock Warp, when speed and affordability matters (Gigadock, Gigadock Warp)

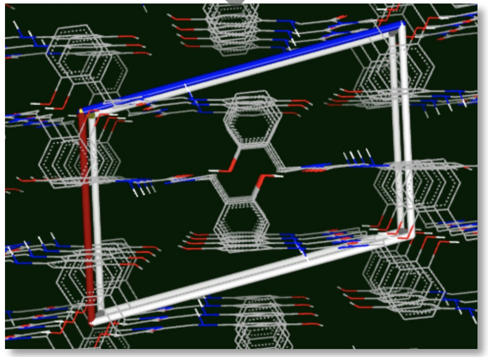
Combined 3D Ligand- and Structure-Based

Ready-to-use workflows that combine the best of 3D ligand- and structure-based methods (FastROCS Plus, Gigadock Warp)

Property Calculation and Filter

Calculate molecular properties and filter out undesirables (FILTER)

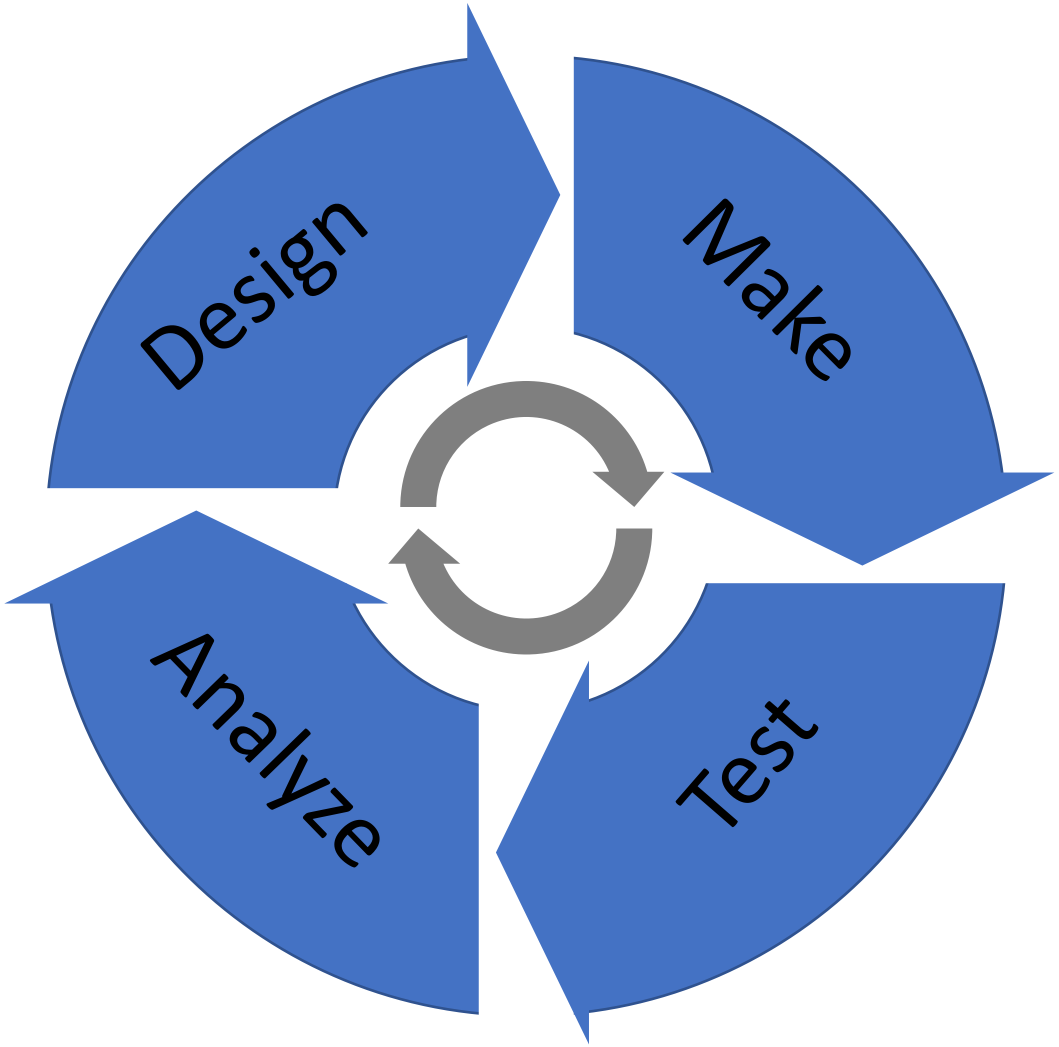
orion®



J. Am. Chem. Soc. 2016, 138, 43, 14257–14263

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

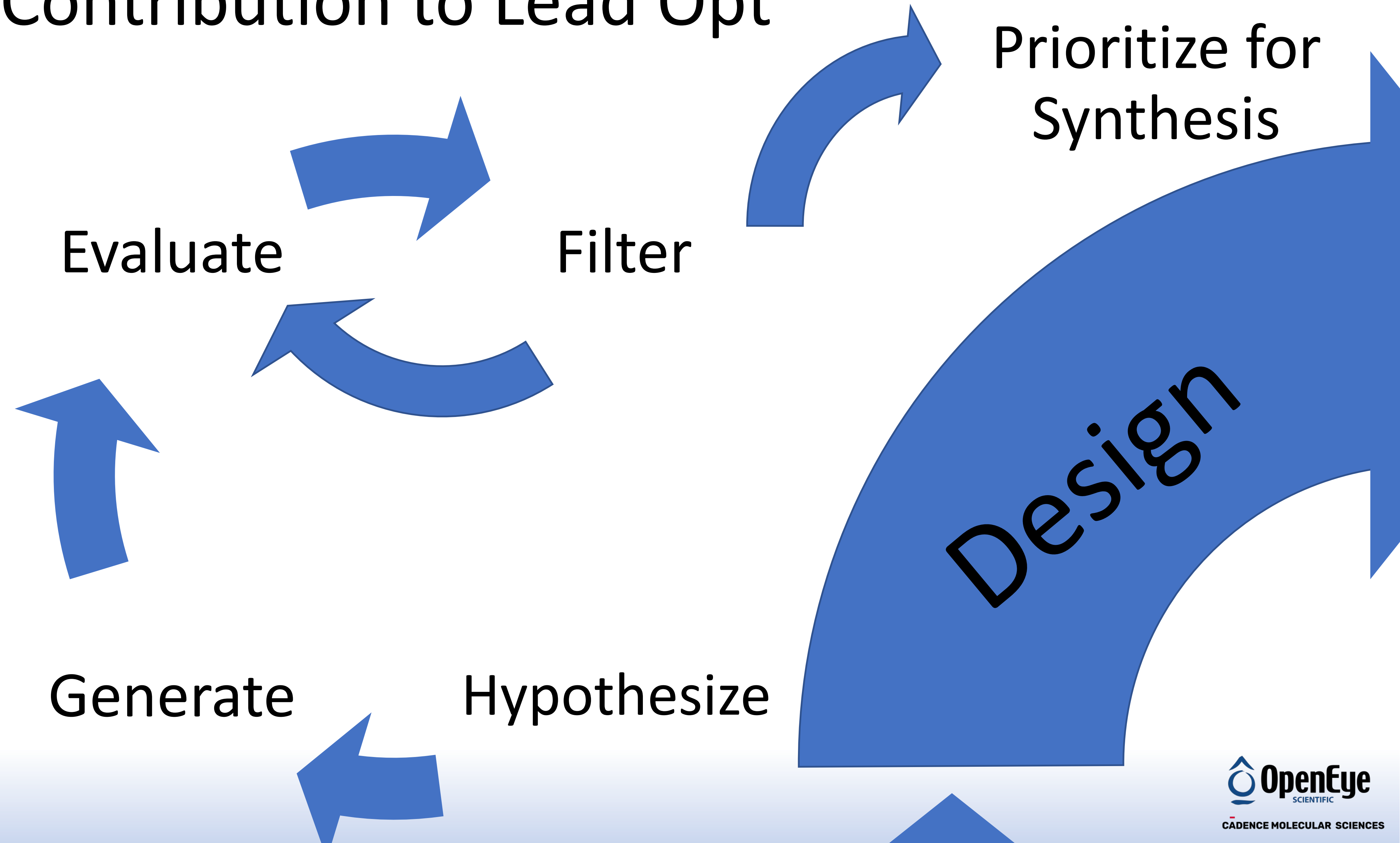
DMTA Cycle in Lead Optimization



- Multiple goals
  - Affinity, ADME-Tox, (Permeability), (Solubility) Selectivity, Synthesizability
  - Priorities shift over Time
- Cycle Time
  - Sequential Process
  - Timelines

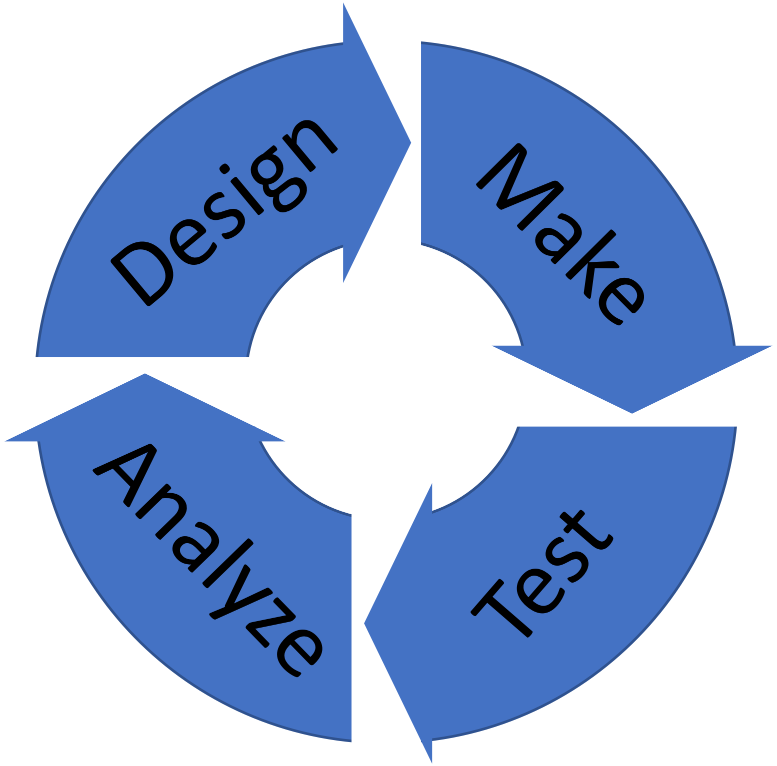
Schneider, P., Walters, W.P., Plowright, A.T. *et al.* Rethinking drug design in the artificial intelligence era. *Nat Rev Drug Discov* **19**, 353–364 (2020). https://doi.org/10.1038/s41573-019-0050-3

CADD Contribution to Lead Opt



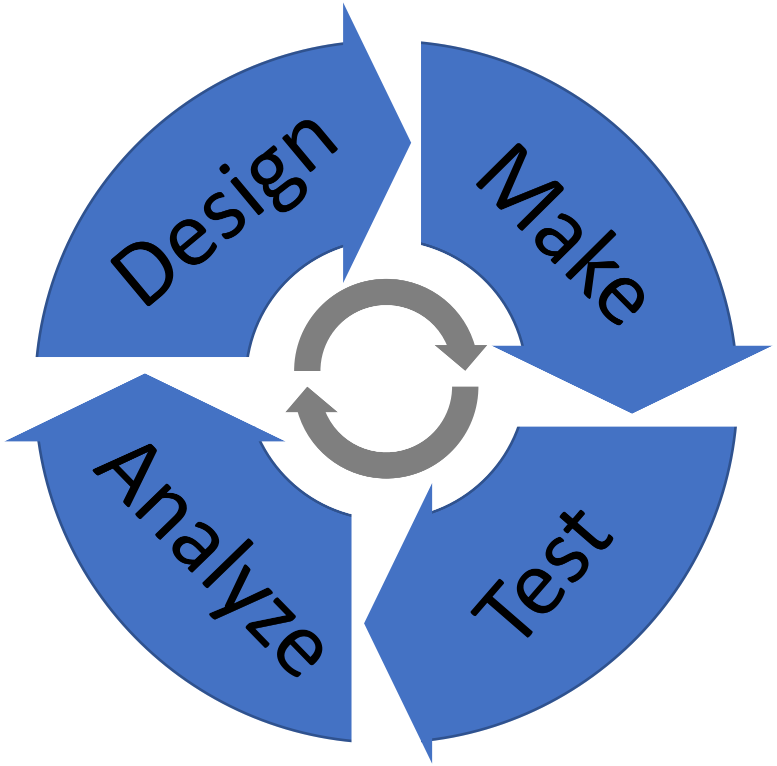
Lead Opt as (too?) Simple Math

$$\text{Cycle Success} \propto P_{Design}P_{Make}P_{Test}P_{Analyze}$$



Lead Opt as (too?) Simple Math

$$\text{Cycle Success} \propto P_{Design} P_{Make} \cancel{P_{Test}} P_{Analyze}$$



OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Lead Opt as (too?) Simple Math

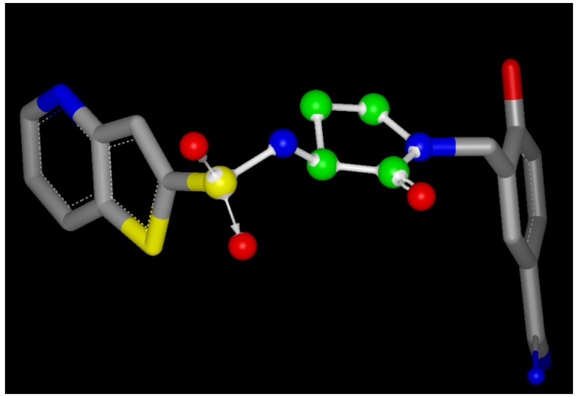
$$\text{Cycle Success} \propto P_{Design} P_{Make} \cancel{P_{Test}} P_{Analyze}$$

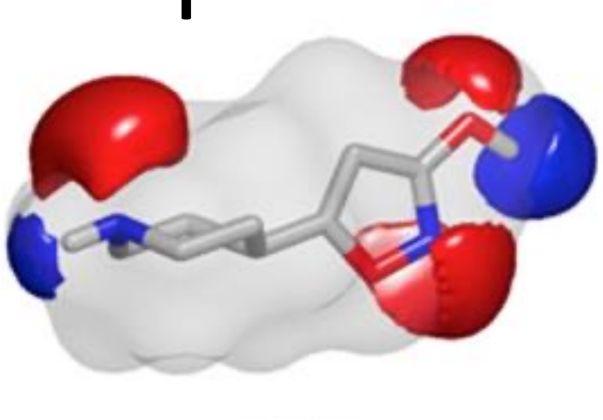
$$\text{LeadOpt Success} \propto P_{Improve} P_{Make}$$

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

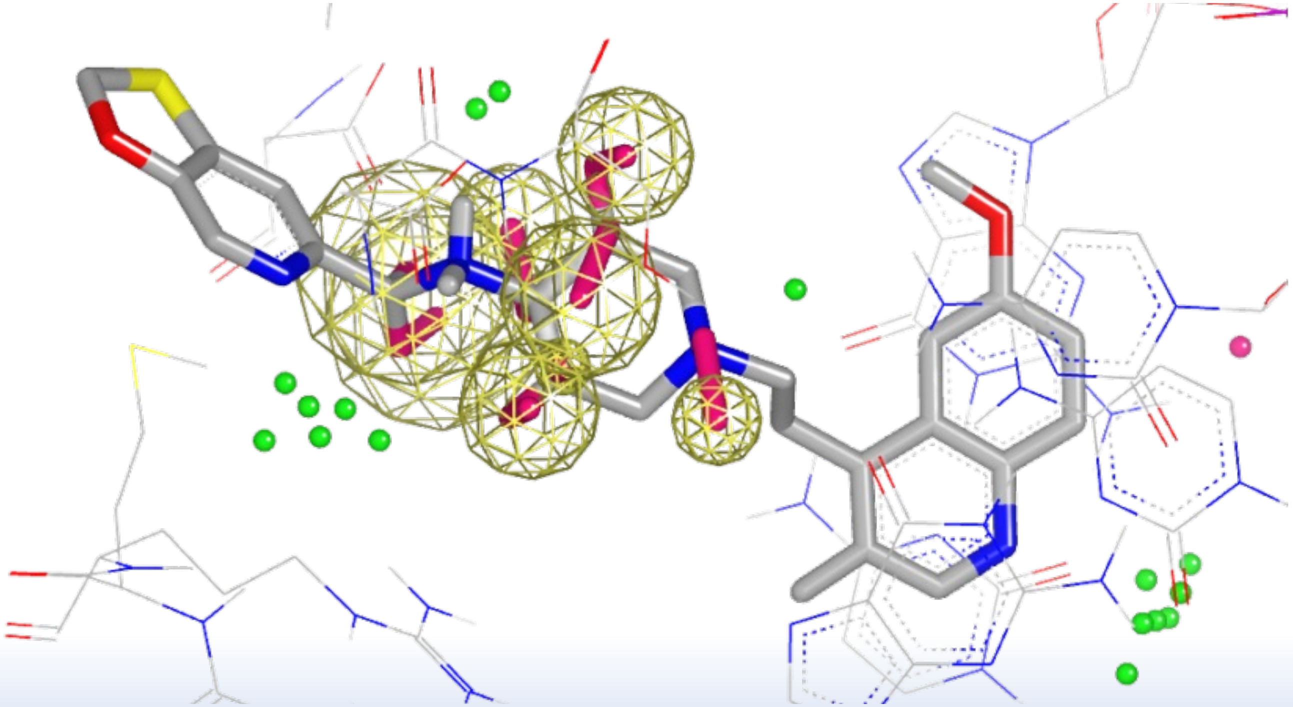
Traditional OpenEye Apps/TKs for Lead Opt

- BROOD
- EON
- FreeForm





- SZMAP/GamePlan
- POSIT



OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

What Orion Brings to the Lead Opt Table

- Platform
  - Central Data Management
  - Scaling Compute Capacity
  - UI for Analysis & Collaboration

- Foundational OE Tools

- New Tools
  - Generating Accessible Candidates
  - ML Models
  - Induced Fit Docking
  - Affinity Prediction

Agenda

- Lead Optimization

- Generating Synthetically Accessible Candidates

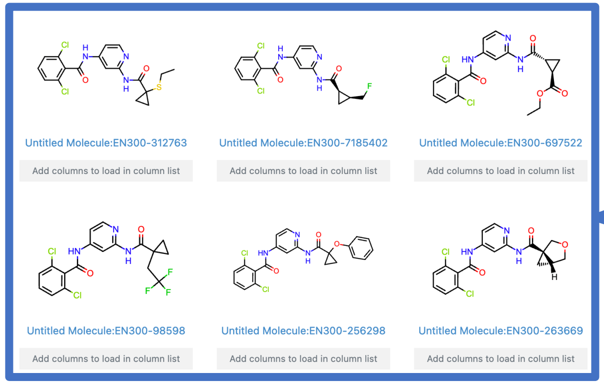
- Machine Learning and Lead Opt

Consider Synthetic Effort at the Beginning

Prioritize fo
Synthesis

Evaluate

Filter



Design

Focused Library Generation

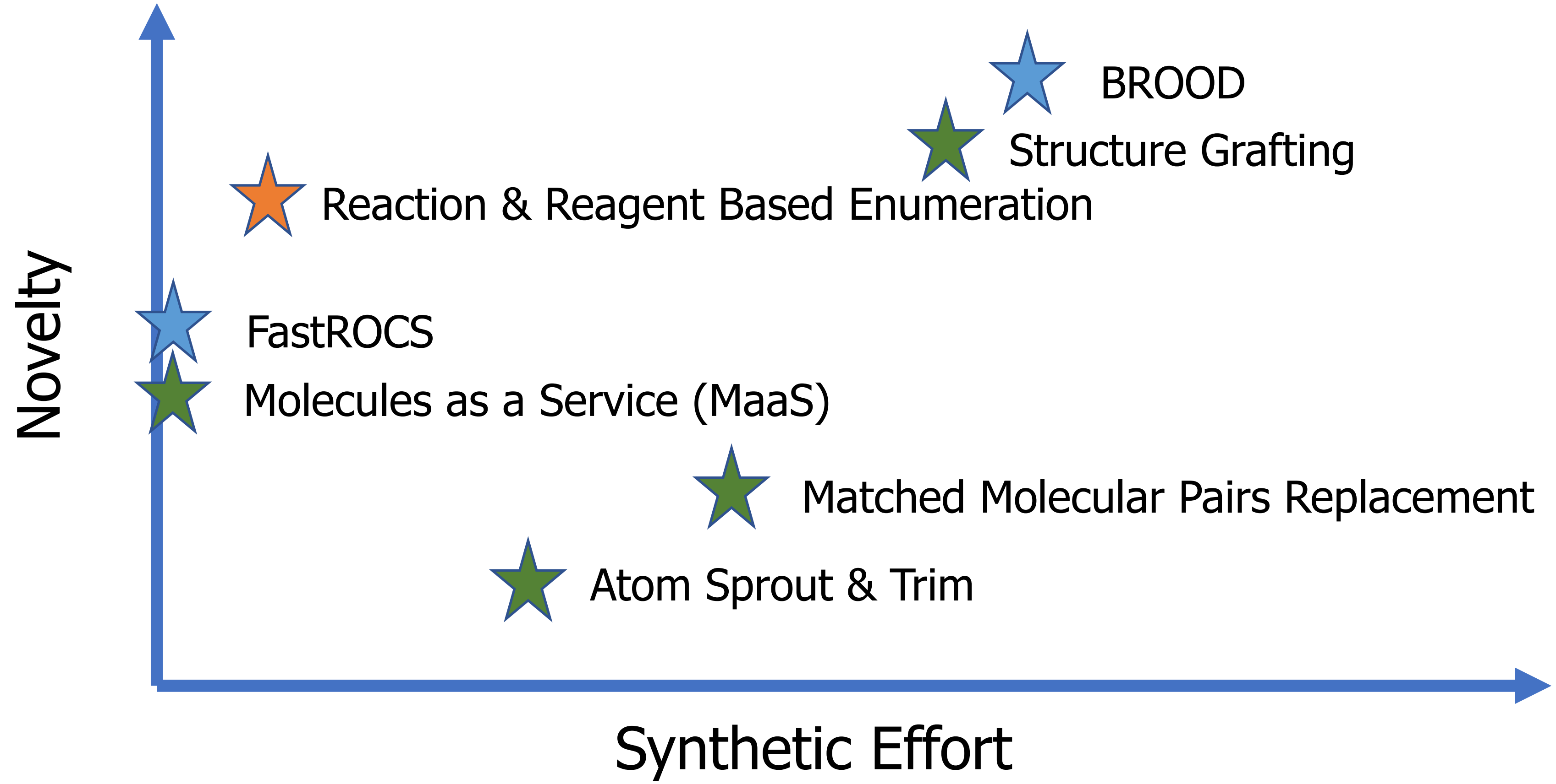
P_{Make}

Generate

Hypothesize

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Library Generation – ”Ideas” on Demand



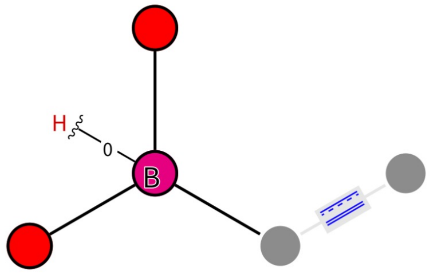
- Searching Purchasable Compounds
- 3D Fragment Replacement
- Virtual Structure Editing
- Virtual Chemistry Enumeration

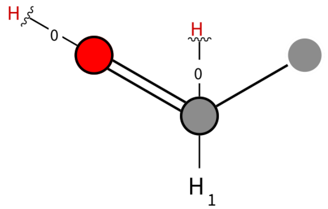
All methods have push-button execution and capability to build customizable databases/libraries

Building the Foundation – Moieties

Shared definition of **MOIETIES** for all higher-level data structures to use:

<MOIETY>	
Suzuki Boronics	[BD3]([#6]=,:[#6])(-O)-O
Aldehydes	[OD1]=[CD2H]-[#6]
Alkyl bromide	CBr
…	





Moieties

Visualization created with: https://smarts.plus/

Building the Foundation – Reagent Types

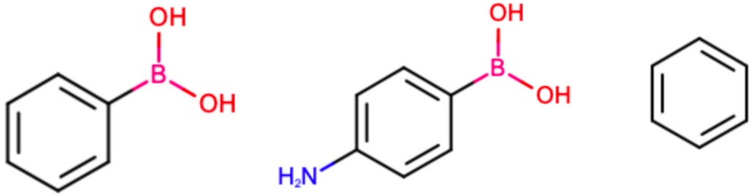
Qualifies compounds as **REAGENT TYPE**
based on **MOIETIES** present and/or absent:

Reagent Types

Moieties

<REAGENT_TYPES>

Reaction: **Suzuki_cross_coupling; Reagent: Suzuki_boronics**
suzuki_boronics = 1
acetal = 0
acid_chlorides = 0
aldehydes = 0
amidines = 0
amines_aliphatic_primary = 0
amines_aromatic_primary = 0
arylbromide = 0
carboxylic_acids = 0
chloroformates = 0
…etc.



Yes No No

Building the Foundation – Reaction Definitions

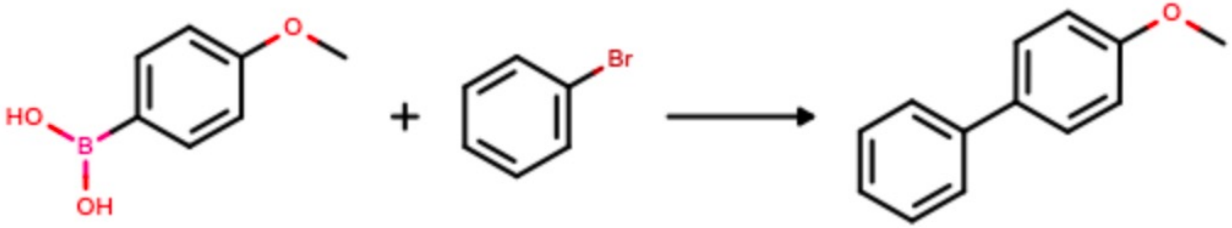
Define **REACTION** by list of classified reagents, SMIRKS to specify transformations of reagents to product structures

<REACTIONS>
Suzuki_cross_coupling
[c!$(c:c=O):1]-[Cl,Br:2].[BD3:3]([#6:4]=,:[#6])(-[O:5]) [O:6]>>[#6:4]-[c:1]
<REAGENTS>
Aryl_bromides
Suzuki_boronics
</REAGENTS>

Reaction Definitions

Reagent Types

Moieties



Definitions for half-reactions (single reagent processing) and retro-transforms are also present

Reaction/Reagent-Based Enumeration

Chemical Moiety Definitions

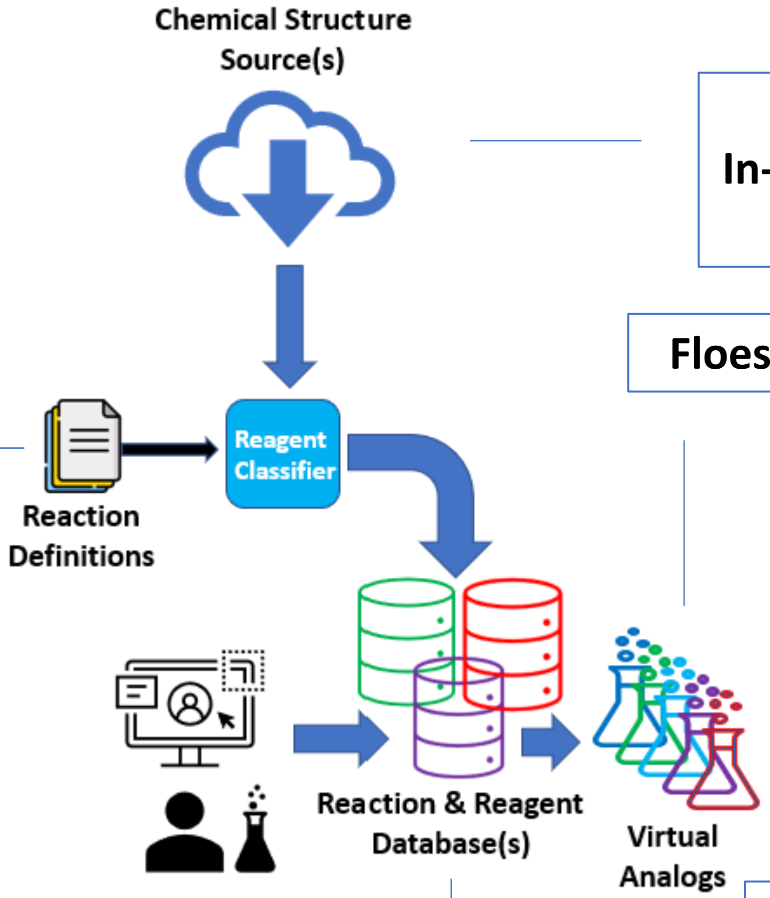
Reaction: Suzuki_cross_coupling

Reaction: Buchwald-Hartwig

Reaction: Amide Synthesis

Reaction Description
- Reaction Name
- Transformation SMIRKS
- Depiction SMIRKS
- Reactants
  - Amine
  - Acid
- Amine constraints
- Acid constraints
- Reagent half-reactions
- Retro reaction transformation(s)

58 reactions currently defined, custom definitions possible but *non-trivial*

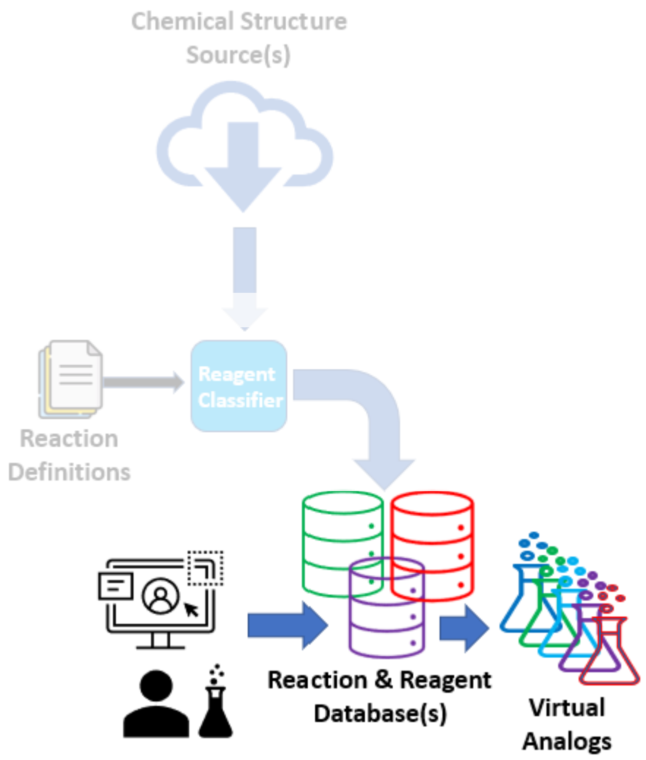


Can use SMILES or .sdf files of In-house building blocks, vendor catalogs, or ZINC tranches

Floes available for focused library generation

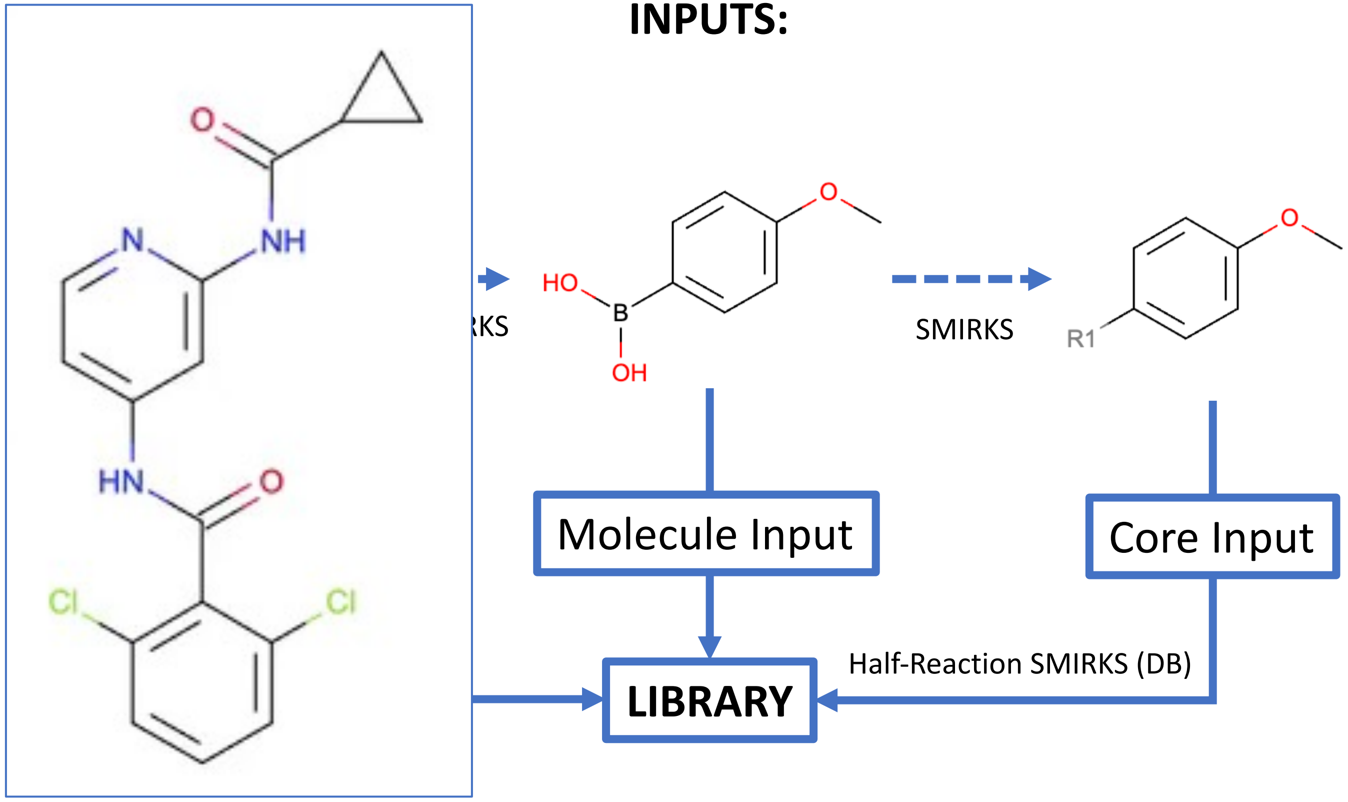
Default database available
Tools for building databases included

Reaction/Reagent-Based Enumeration

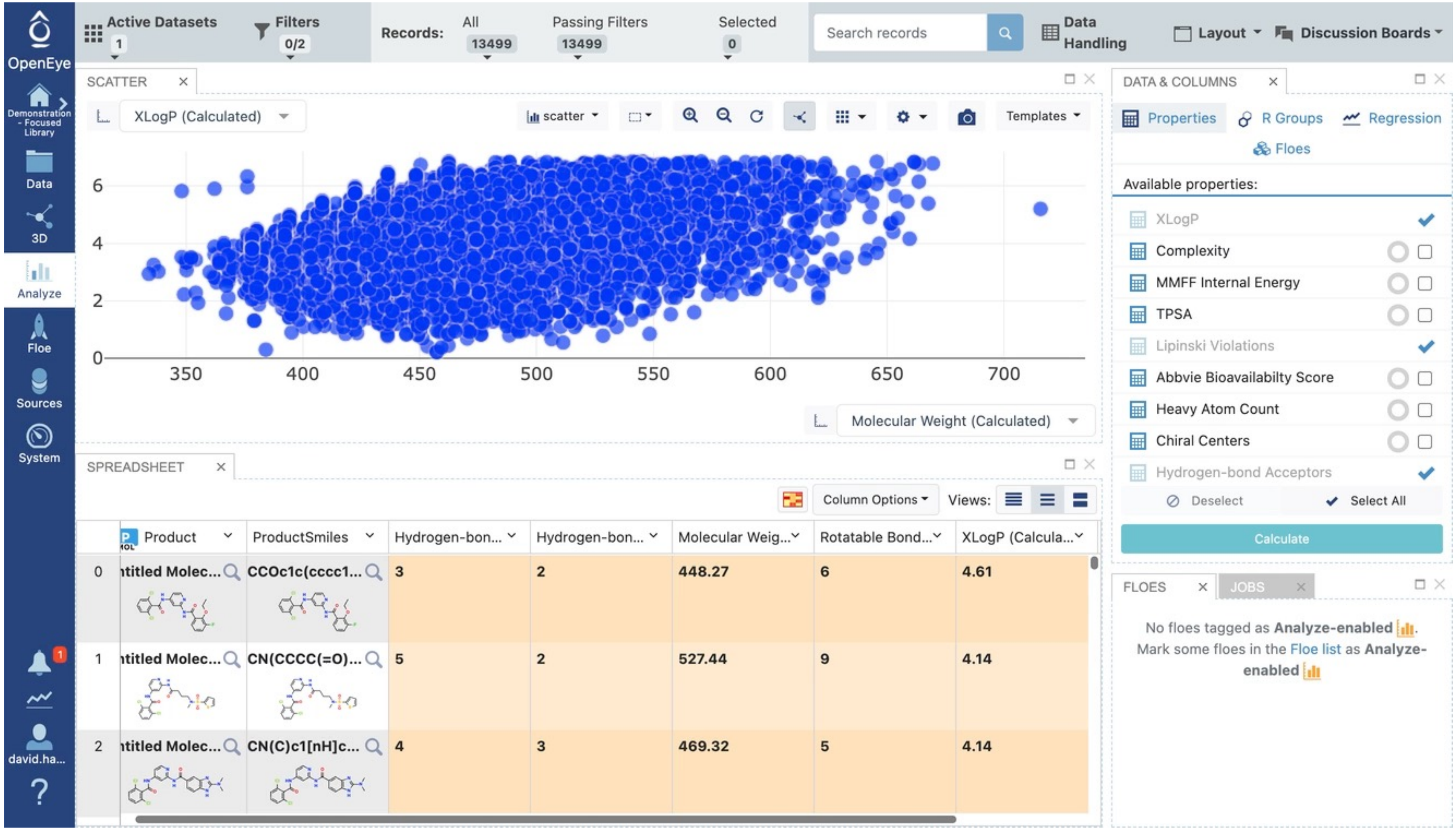


Floes for Enumeration

INPUTS:



Filtering and Visualization with Orion



Agenda

- Lead Optimization

- Generating Synthetically Accessible Candidates

- Machine Learning and Lead Opt

Machine Learning in the Lead Opt Context

- Paucity of (Good, Local) Data
  - Prefer simpler models
  - Explore global & local models
  - Beware overfitting

- Good ML Hygiene
  - Domain of Applicability
  - Confidence/Error Bars
  - Explainable Predictions

- Utility in Lead Opt:
  - Tools for assessing "modelability"
  - Filtering and Prioritization

Unlikely

"Cleanliness becomes more important when godliness is unlikely."
— P. J. O'Rourke

creative commons cc-by-sa 3.0 license

Model Building in Orion

- Model Building Floes shipped in 2022

- Floe to prepare data

- Floes to train Models

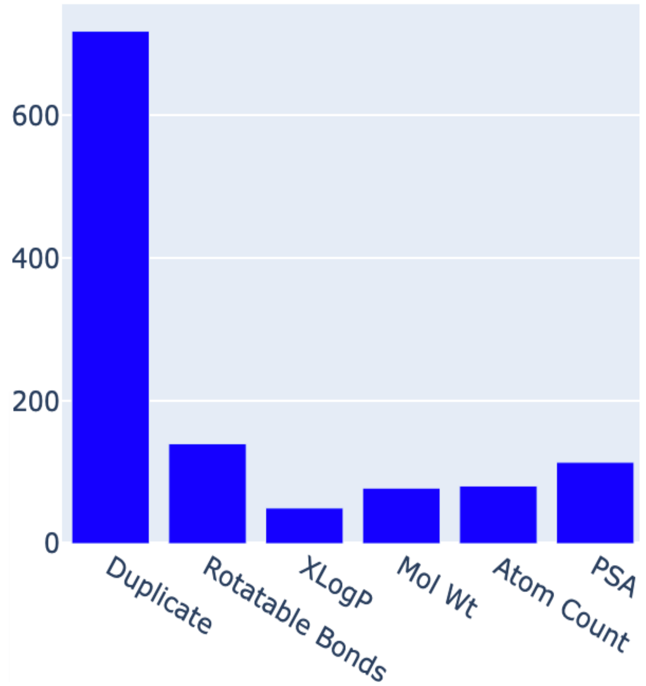
- Floes to apply Models for prediction

- Pre-trained Solubility model

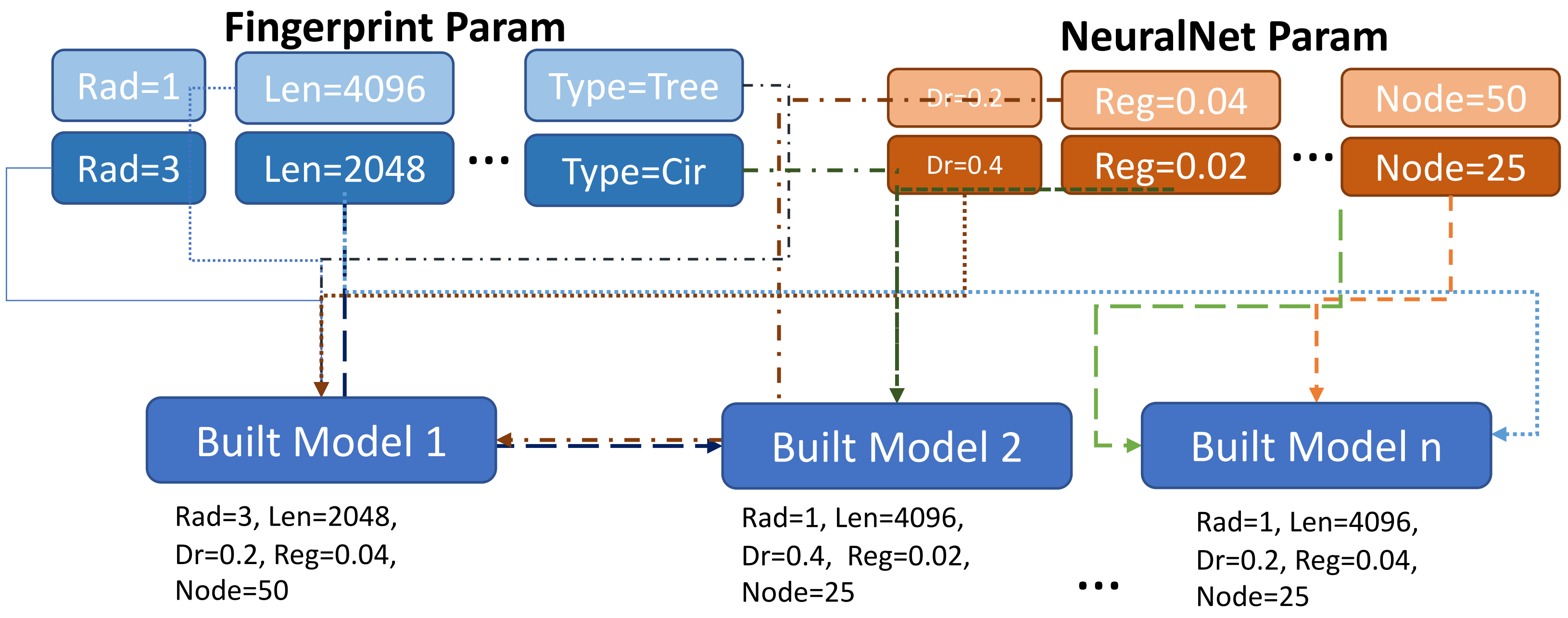
ChEMBL_030: http://www.ebi.ac.uk/chembl

Training Data Pre-Processing

- Molecule standardization
- Simple property distributions
- Duplicate handling
- Response (or class) distribution



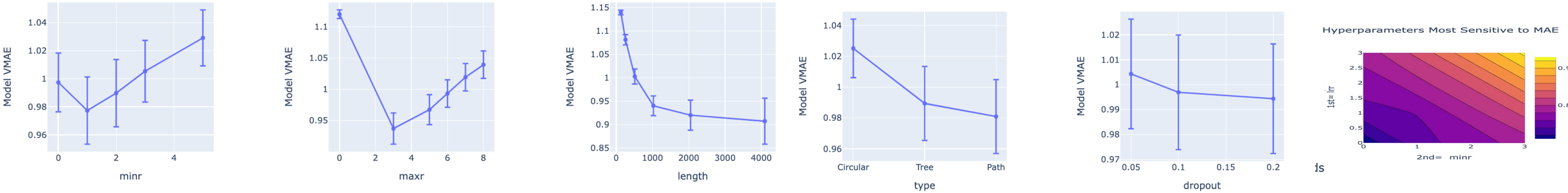
Train Multiple Models



Grid-Based Parameter Search

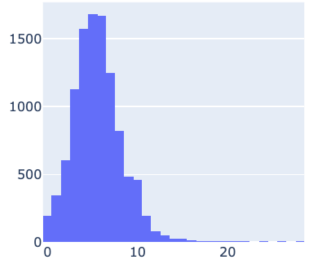
Model Building Floe Report

Summary of Model Hyperparameters on Validation Set



List of Fully-connected Neural Network Models Generated

Model #	Record #	MinRad	MaxRad	Bit-Length	FP-Type	Dropout	Learning Rate	Hidden Layers	Reg Layers	VMae	VLoss	Link
1	122	0	3	4096	Path	0.4	0.0005	250 150 100	0.04 0.02 0.02	6.99e-01	1.49e+00	Model link
2	126	0	3	4096	Path	0.1	0.0005	250 150 50	0.04 0.02 0.02	7.03e-01	1.56e+00	Model link
3	141	0	3	4096	Path	0.2	0.0005	250 150 100	0.04 0.02 0.02	7.07e-01	1.53e+00	Model link
4	33	1	3	4096	Tree	0.4	0.0005	250 150 100	0.04 0.02 0.02	7.08e-01	1.49e+00	Model link
5	56	1	3	4096	Path	0.1	0.0005	250 150 50	0.04 0.02 0.02	7.14e-01	1.58e+00	Model link
6	128	0	3	4096	Tree	0.4	0.0005	250 150 100	0.04 0.02 0.02	7.16e-01	1.48e+00	Model link
7	86	0	3	4096	Tree	0.1	0.0005	250 150 50	0.04 0.02 0.02	7.19e-01	1.54e+00	Model link
8	58	1	3	4096	Path	0.2	0.0005	250 150 100	0.04 0.02 0.02	7.19e-01	1.56e+00	Model link
9	15	1	3	4096	Tree	0.1	0.001	250 150 100	0.04 0.02 0.02	7.24e-01	1.43e+00	Model link
10	16	1	3	4096	Tree	0.1	0.0005	250 150 100	0.04 0.02 0.02	7.27e-01	1.61e+00	Model link
11	108	0	3	4096	Tree	0.2	0.0005	250 150 100	0.04 0.02 0.02	7.32e-01	1.60e+00	Model link
12	67	0	3	4096	Tree	0.1	0.0005	250 150 100	0.04 0.02 0.02	7.32e-01	1.58e+00	Model link
13	105	0	3	4096	Path	0.4	0.0005	250 150 50	0.04 0.02 0.02	7.35e-01	1.63e+00	Model link



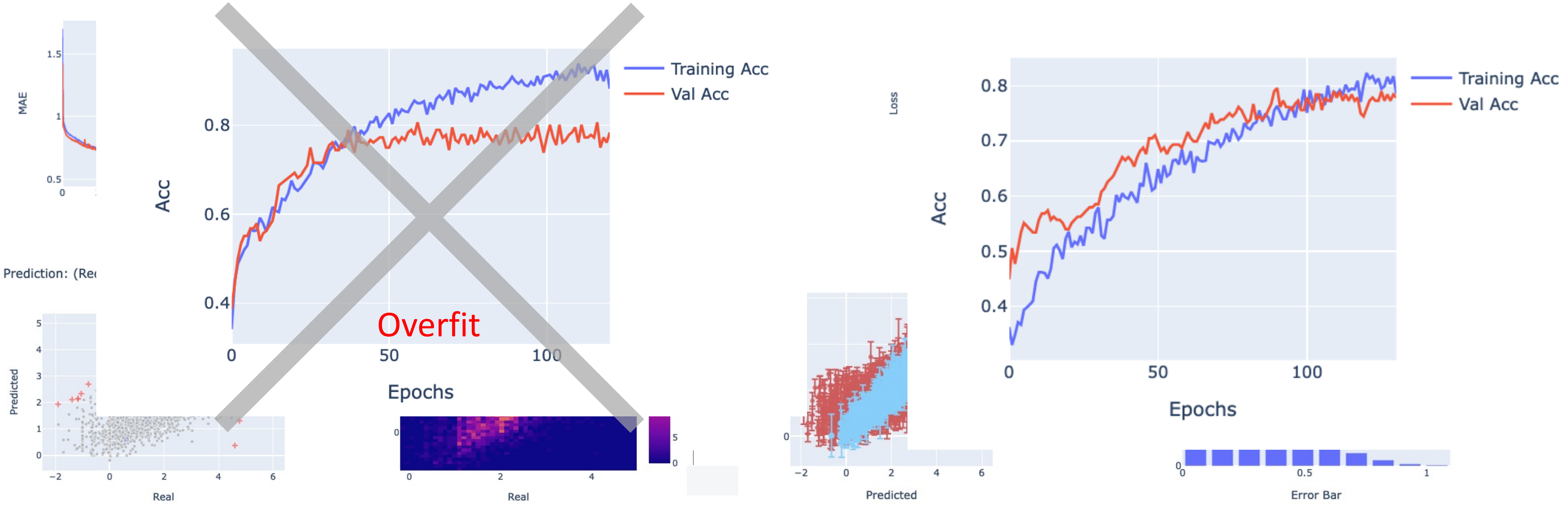
CADENCE MOLECULAR SCIENCES

Training: Floe Report and Model Details

NN Hyperparam and Fingerprint Parameters

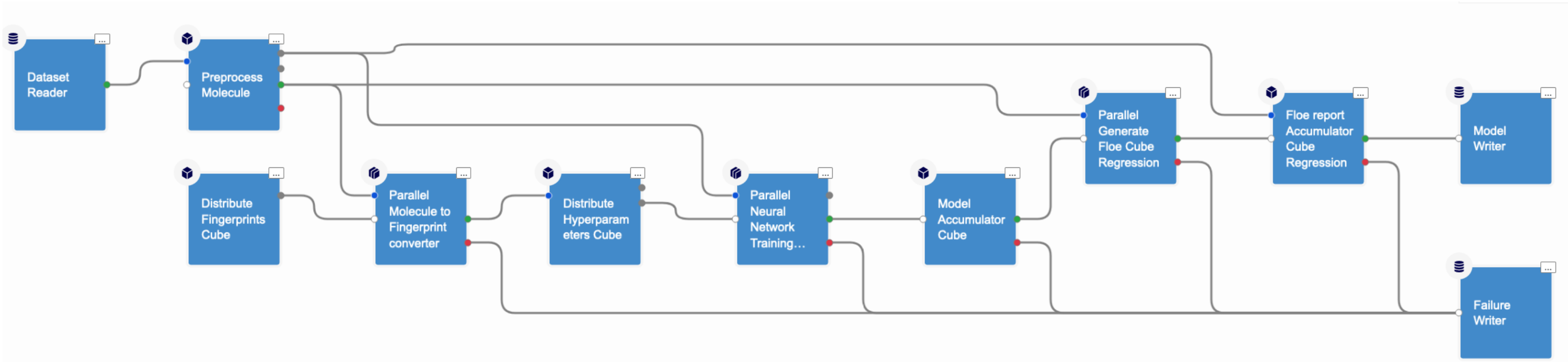
Dropout	Max Epochs	Hidden Layers 0	Hidden Layers 1	Hidden Layers 2	Reg Layers 0	Reg Layers 1	Reg Layers 2	Learning Rate	FP Type	FP BitLength	Min Rad	Max Rad
0.1	1000	250	150	100	0.06	0.04	0.02	0.0001	T	4096	0	3

Neural Network Epoch Training Plots



OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Model Building Summary



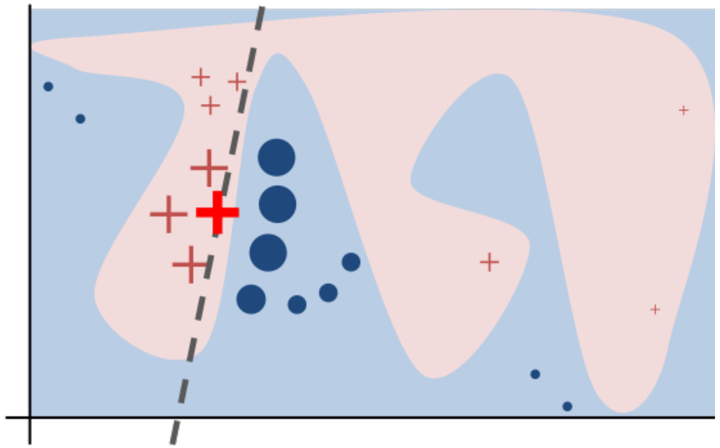
- Input
  - Molecules (or custom feature vectors)
  - Response property
- Training Setup
  - Regression or Classification
  - Automatic Train/Test/Validation splits and cross-validation
- Build Multiple Models (in Parallel)
  - Featurization Parameters (FPs for now)
  - Neural Network hyperparameters
- Comprehensive Floe Report
  - Summary with Model Ranking
  - Individual Model Performance

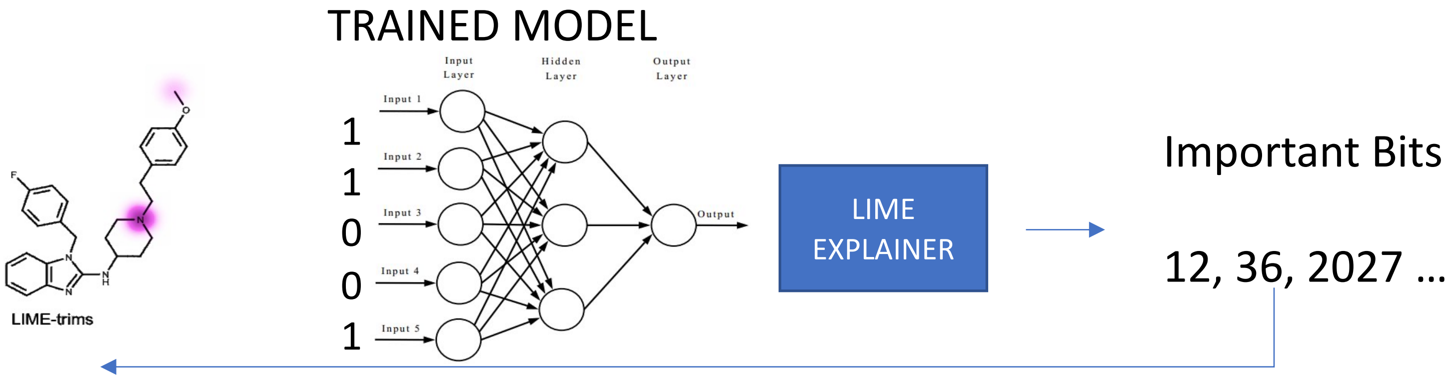
Making Predictions

- Predictions in Parallel
  - Take advantage of Orion, predictions can be applied at scale
- Confidence and Domain of Applicability
  - Models retain knowledge of training set property distributions
  - Multiple ways to assess uncertainty (property box, dropout)
  - Train and supply separate Bayesian uncertainty model
  - High/Med/Low assessment with rationale
- Explainable Predictions

Local Interpretable Model-Agnostic Explanations (LIME)

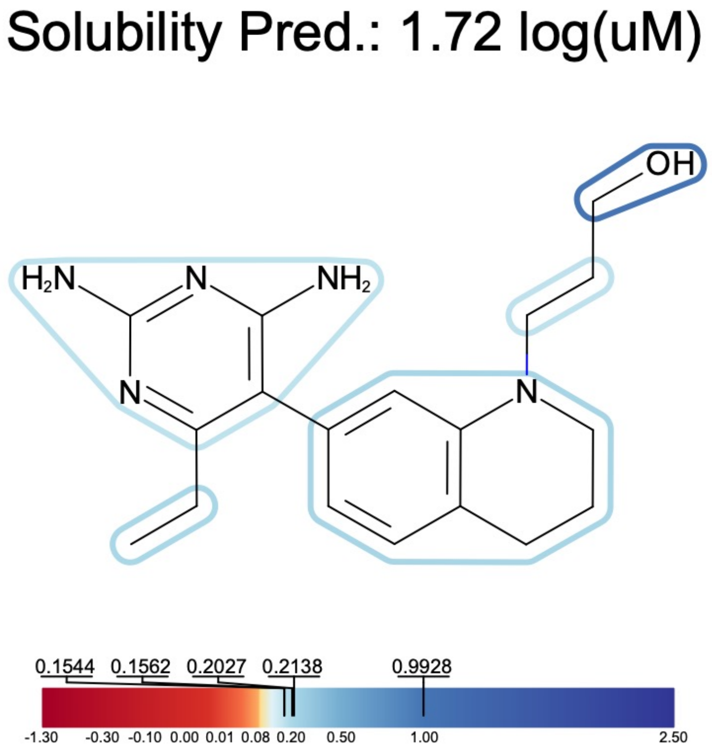
- **Interpretable:** qualitative understanding between the input variables and the response
- **Local Fidelity:** learns an interpretable model locally around the prediction.



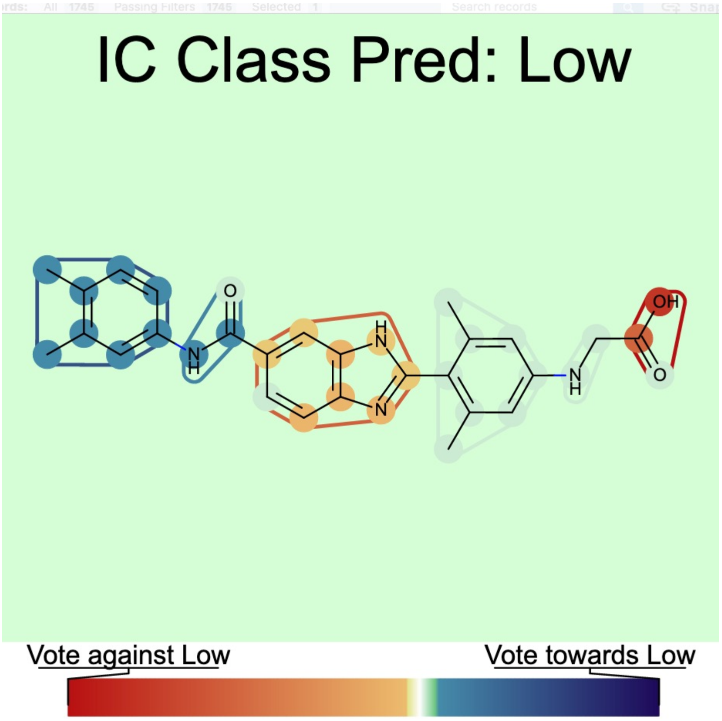


Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier." https://doi.org/10.48550/arXiv.1602.04938

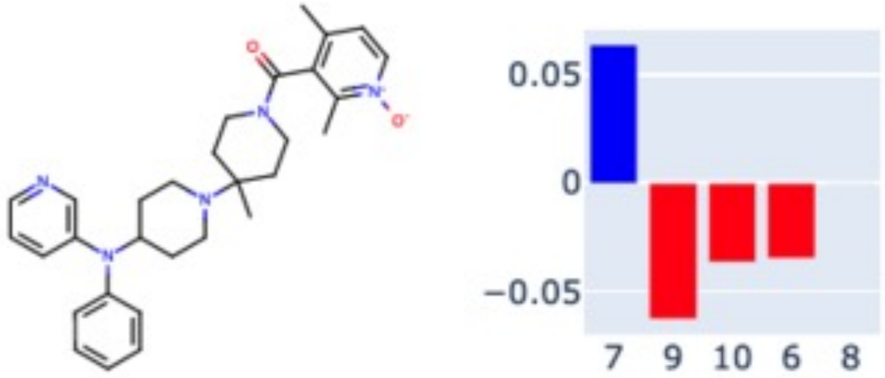
Prediction Explainer Examples



Regression



Classification



Custom Feature Vector

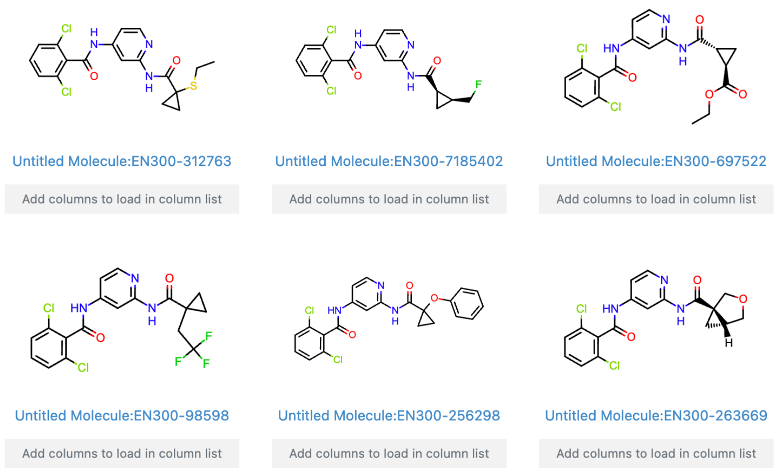
Future Work

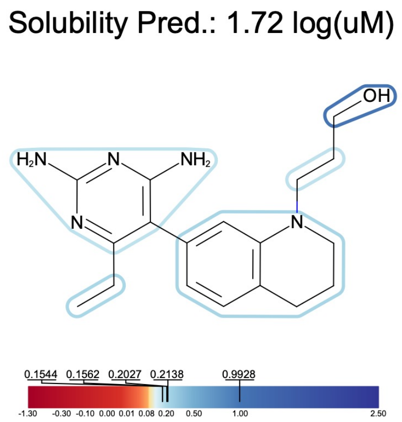
- Graph Convolutions
- 3D Representations
- Kriging
- Diffusion
- Distributed Training of Larger Models
- Hyperparameter Optimization
- Advanced Train/Test Splitting
- Model Comparison

And much more…

Conclusions

- Orion brings all of OpenEye Lead Opt capabilities into one place
  - Ways to increase $P_{Improve}$ and P_{Make}
- Robust cheminformatic foundation for building synthetically accessible lead opt libraries
- General ML tools to assist in model building with guardrails





Acknowledgements

- Burt Leland
- Sayan Mandal
- David Hamilton
- Cheminformatics & Data Science Group

- Scientific Development Teams
- Cloud Teams
- Geoff Skillman, Ant Nicholls, Bob Tolbert

Open Positions in the Cheminformatics & Data Science Group

OpenEye SCIENTIFIC
CADENCE MOLECULAR SCIENCES

Thank You


OpenEye
SCIENTIFIC
CADENCE MOLECULAR SCIENCES