October 30, 2024

Streamlining the Design of DNA-Encoded Libraries at GSK Using KNIME

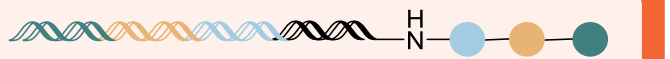
Brittany Smith, GSK

Who is Encoded Technologies (ET), GSK?

GSK

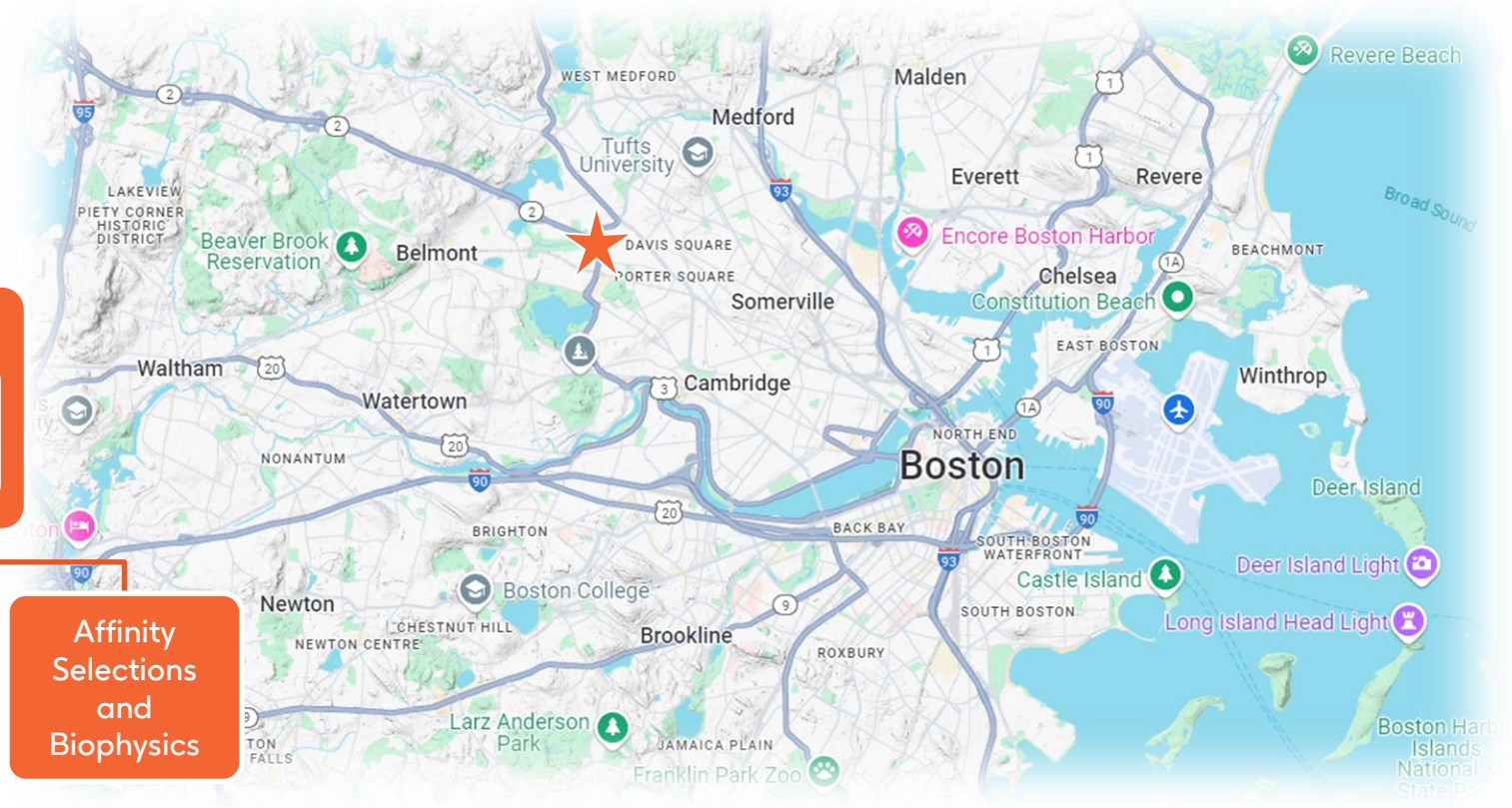
MMD/R&D

Encoded Technologies



DNA-encoded libraries (DELs)

- Platform Chemistry
- Data Science and Engineering
- Lead Discovery
- Affinity Selections and Biophysics



Molecular Modalities Discovery (MMD), Research and Development (R&D)

GSK

Acknowledgments

Katelyn Billings
• former GSK employee

Melissa Grenier-Davies

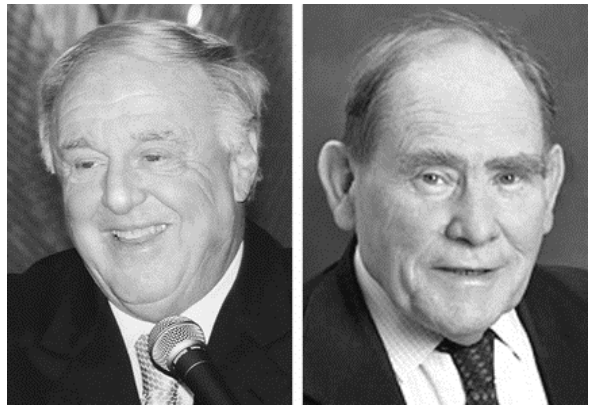
Carol Mulrooney

Sarah Carden
• former GSK employee

DNA-Encoded Library (DEL) Technology

Screen billions of compounds in a single tube

- First proposed by Sydney Brenner and Richard Lerner in 1992
- Libraries of small molecules covalently encoded by unique DNA sequences
- Linked phenotype with genotype allows for pooling, affinity selection to be conducted in a single vessel as mixtures
- Power of molecular biology techniques – PCR amplification, high-throughput sequencing



Many companies now have DEL partnerships or are internally investing in DEL technology

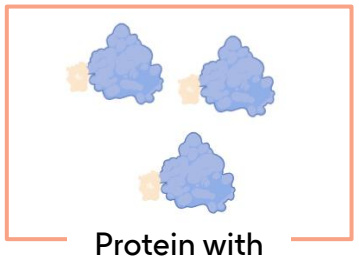
GSK PNAS 1992, 89, 5381-5383 | Angew. Chem. 2017, 56, 1164-1165

DNA-Encoded Library Technology (ELT)

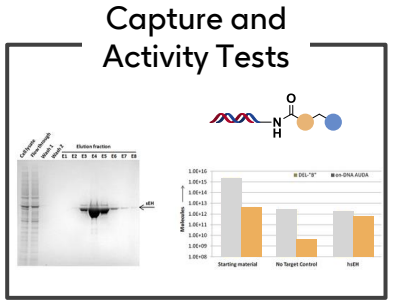
Affinity-based screening of DNA-encoded libraries

Advantages

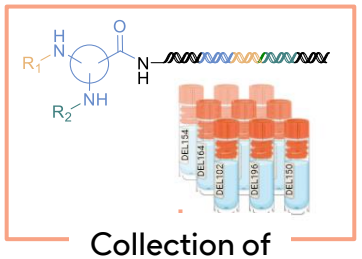
- ✓ Screen billions of compounds in a single tube
- ✓ Binding mode agnostic (potential for allosteric binder ID)
- ✓ Minimum protein requirement (<1 mg)
- ✓ Fast turn-around time/low cost (automation not required)



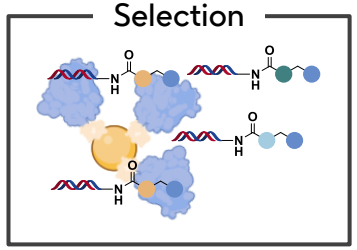
Protein with Affinity Tags



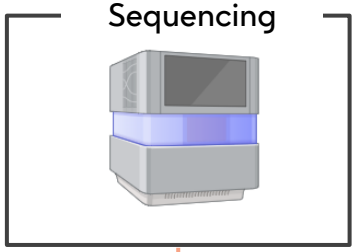
Capture and Activity Tests



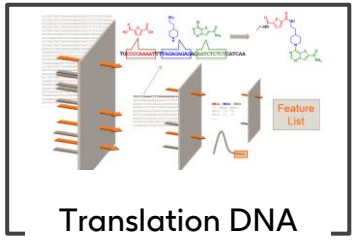
Collection of DELs



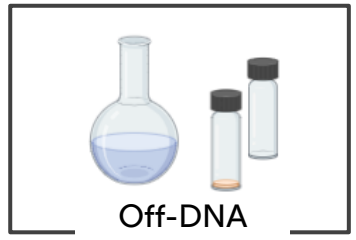
Affinity Selection



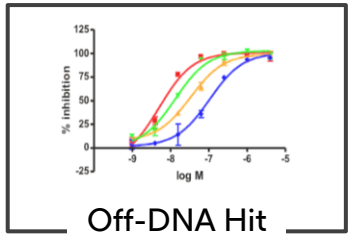
Amplification + Sequencing



Translation DNA Tags to Structures



Off-DNA Synthesis



Off-DNA Hit Evaluation

Qualified Hits for Lead Generation Efforts

Tool Compounds for Target Validation

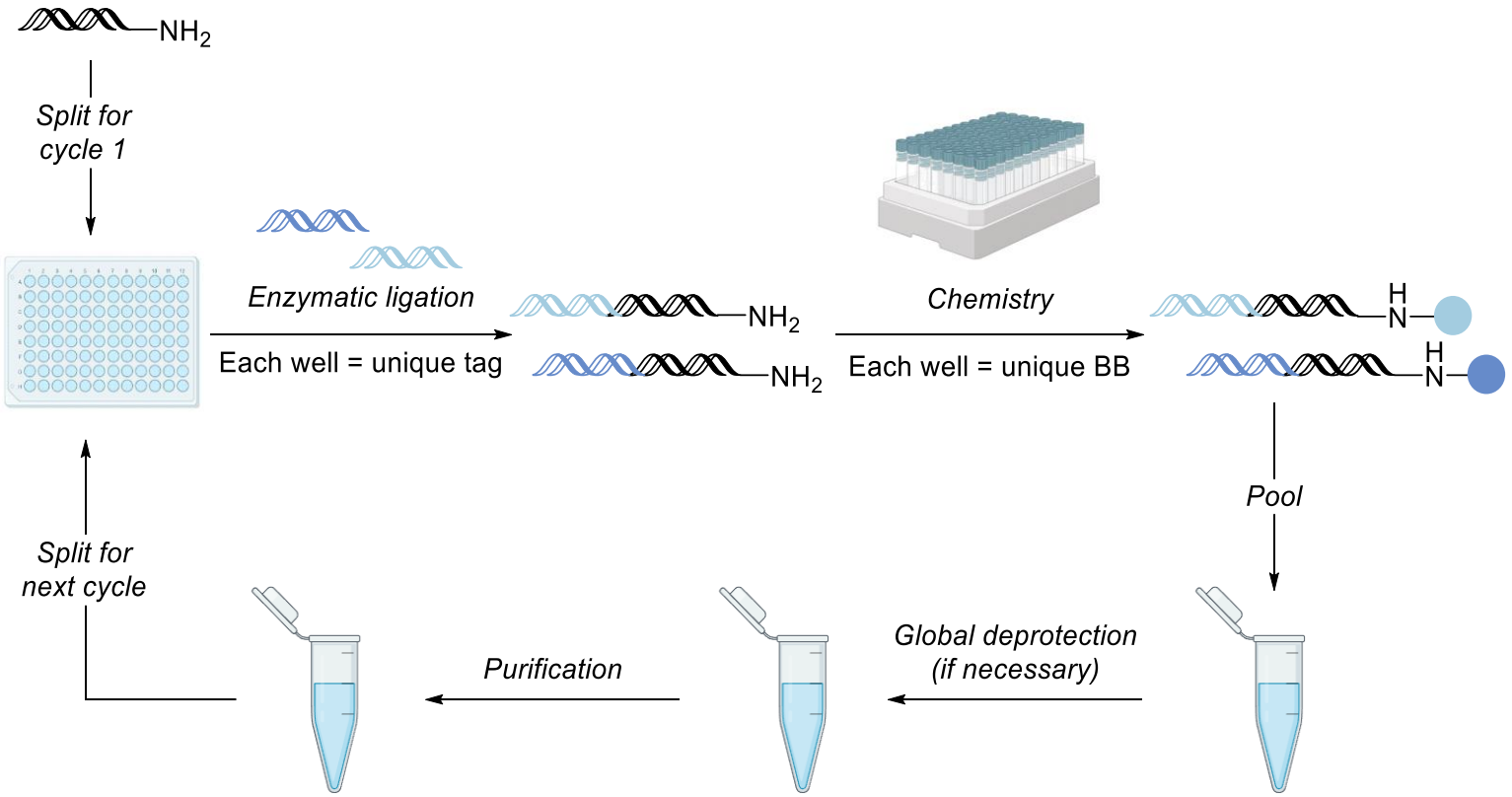
Off-DNA Small Molecules

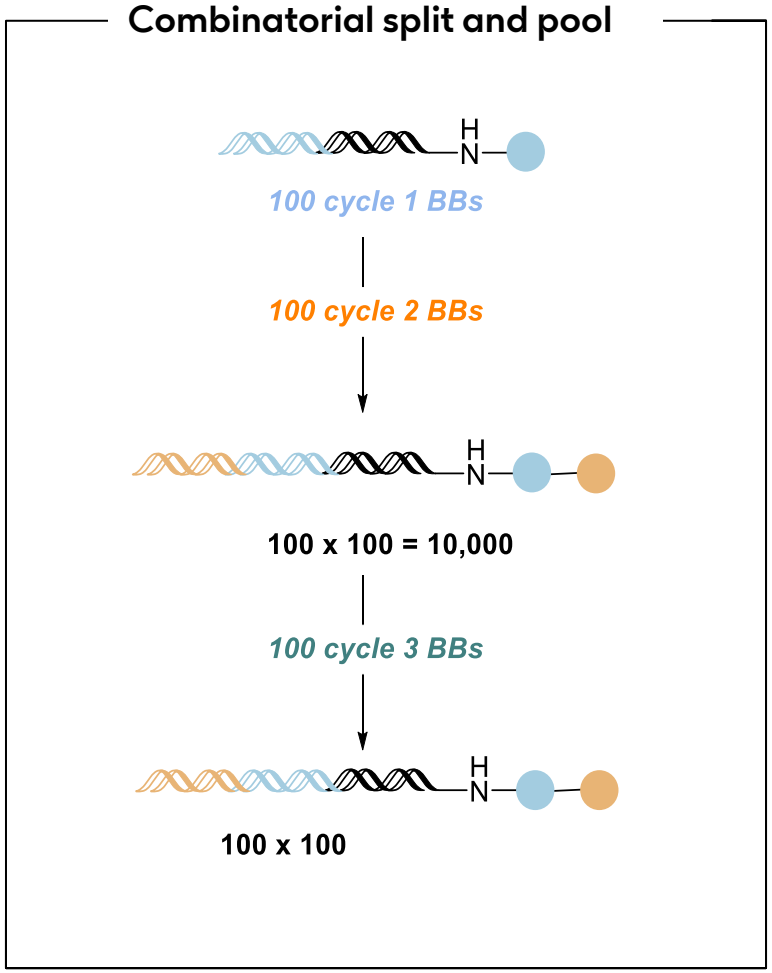
Images created in BioRender.com

GSK

DNA Encoded Library Technology (ELT)

Library Synthesis





Images created in BioRender.com

GSK

External Software Enabling DEL Design, Synthesis and Translation

In addition to internally built software

Early DEL design

Building block (BB) selection, virtual library enumeration

jupyter

python™

RDKit
Open-Source Cheminformatics and Machine Learning

Open for Innovation KNIME

Schrödinger

Visualization and BB Scoring

TIBCO® Spotfire®

Implementation

Database, library registration

ORACLE®

Virtual compound enumeration

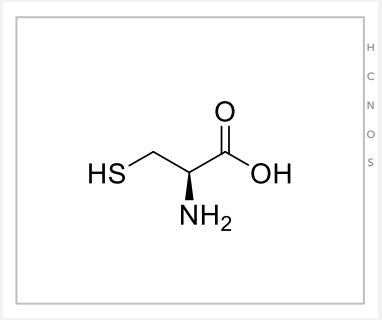
python™

GSK

SMILES and SMARTS

Encoding molecules for data analysis

Sigma Aldrich



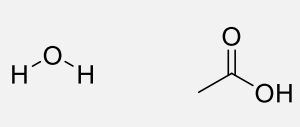
Thousands-billions of compounds

How do we encode large numbers of compounds?

SMILES

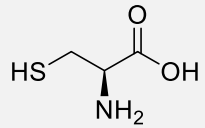
Simplified **M**olecular **I**nput **L**ine **E**ntry **S**ystem

Used to define atoms, bonds, and valences of molecules



O

CC(O)=O

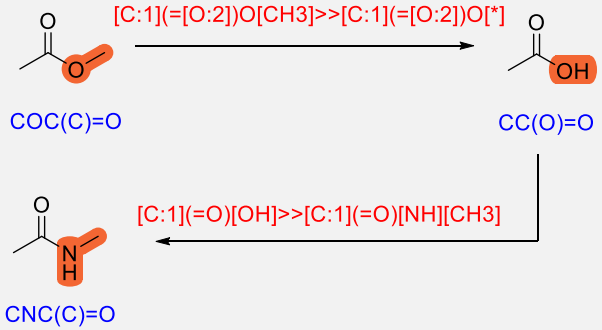


C([C@@H](C(=O)O)N)S

SMARTS

SMILES **Ar**bitrary **T**arget **S**pecification

Used to define chemical transformations



https://www.daylight.com/smiles/index.html

Building Block Selection

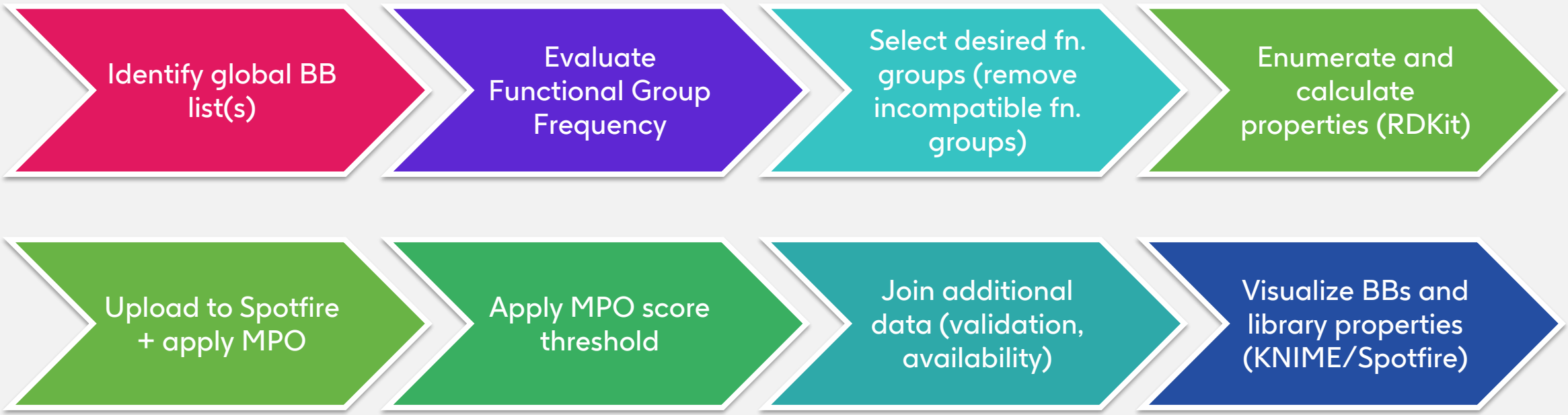
Background

- DNA-encoded libraries designed for developability
  - Goal of new library design is to improve the quality of the hits identified in ET selections in order to generate more lead-like hits for medicinal chemistry programs
  - Physicochemical properties of building blocks and number of cycles of chemistry used to build the library both contribute to the overall property distribution

- Building block selection for GSK DELs
  - Previously used primarily validation yield to filter potential building blocks
  - Subsequently implemented strict filters that included validation yield and various properties
  - Most recently implemented Multi-Parameter Optimization (MPO) score threshold to select building blocks

Building Block Selection Workflow

Selecting building blocks to include in DNA-encoded libraries



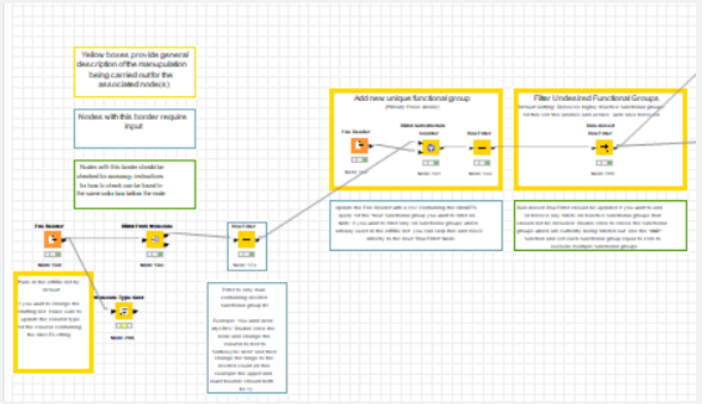
Building Block Selection Workflow

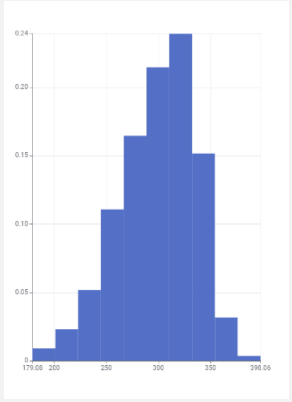
Using cheminformatics and visualization software

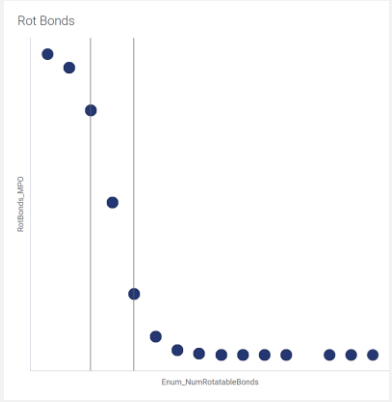
Use KNIME to assess properties and substructures flags for lists of commercial and internal BBs

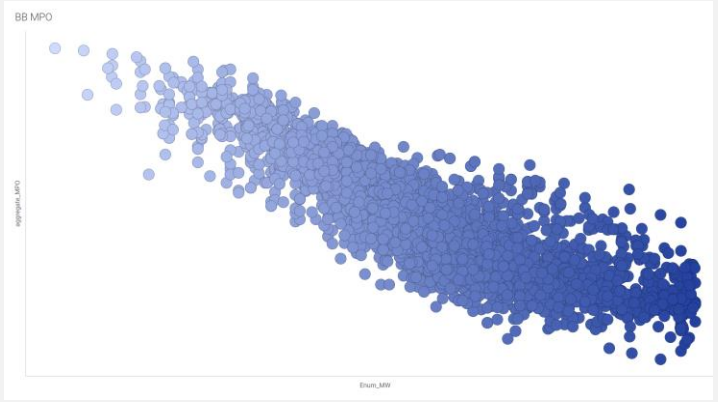
Import BB list(s) into Spotfire and calculate MPO scores

Use visualization tool to evaluate and select BBs, using MPO scores as a guide









Open for Innovation KNIME

Open for Innovation KNIME

Marcus Farmer (Spotfire)

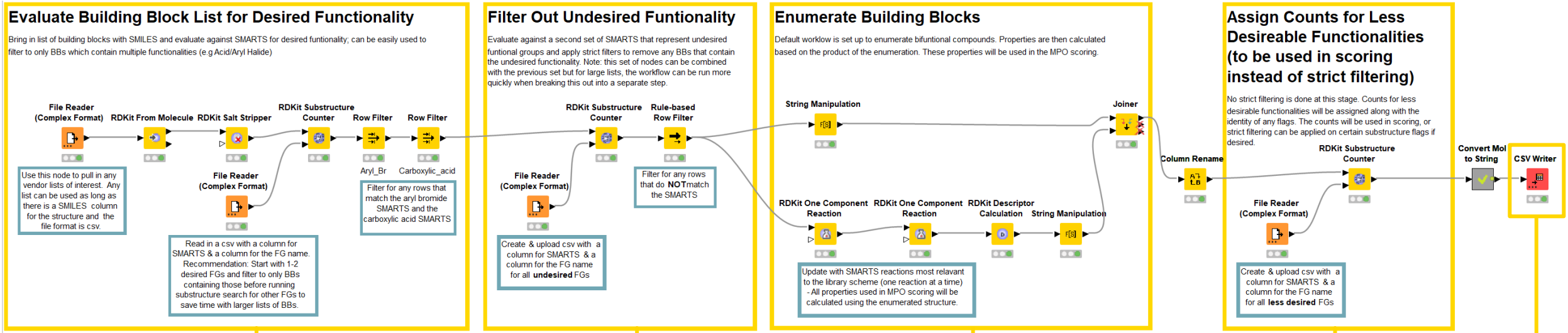
TIBCO Spotfire®

TIBCO Spotfire®

GSK

KNIME Workflow

Building block evaluation



Brings in list of compounds and evaluates for desired functional groups | strict filtering is applied to only include compounds containing the desired functional handles

Evaluates remaining compounds and applies strict filters to eliminate any undesired functional groups (e.g. reactive electrophiles)

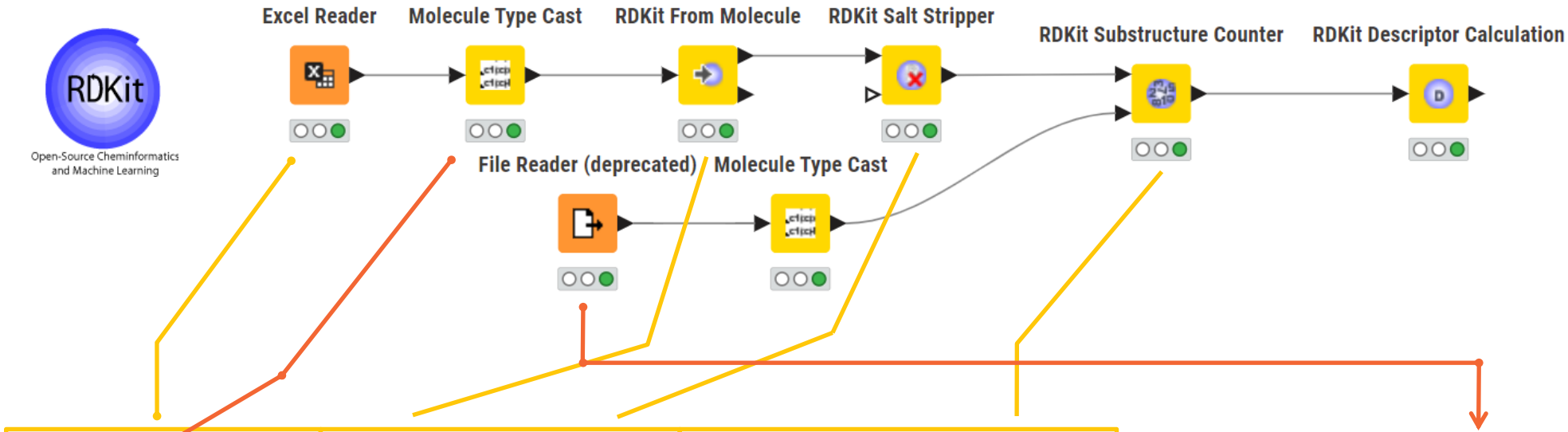
Enumerates structures so that property and substructure analysis will be representative of compounds incorporated into the library

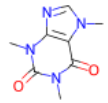
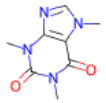
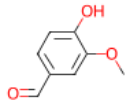
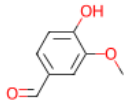
Provides substructure count for parameters to be considered in the MPO scoring

Generates csv file for use in Spotfire to score compounds

KNIME Workflow

RDKit

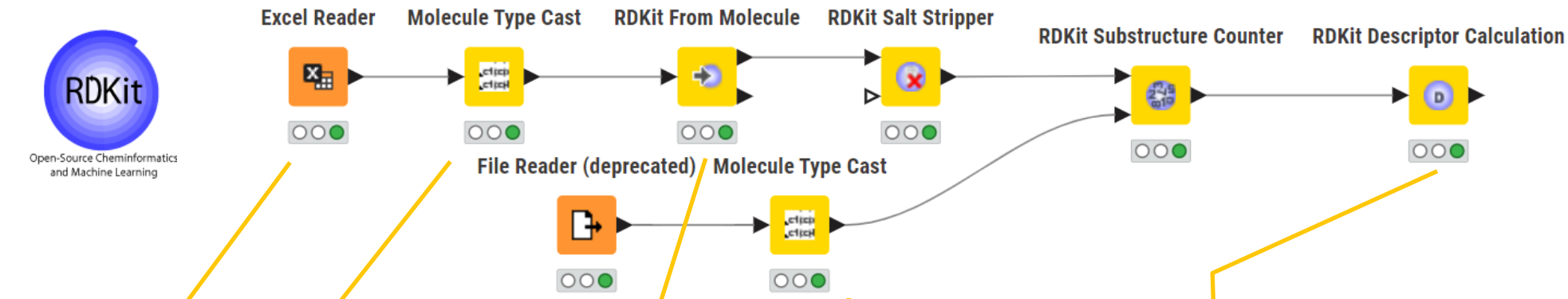


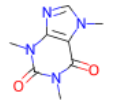
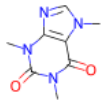
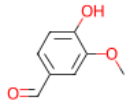
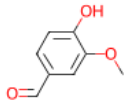
Name *String*	SMILES *Smiles*	pSMILES (RDKit Mol) *RDKit Molecule*	Salt Stripped Molecule *RDKit Molecule*	carboxylic_acid *Number (integer)*	aldehyde *Number (integer)*	aromatic_OH *Number (integer)*
Caffeine	Cn1cnc2n(C)c(=O)n(C)c(=O)c12			0	0	0
Vanillin	COc1cc(C=O)ccc1O			0	1	1

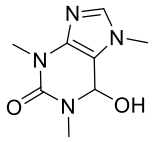
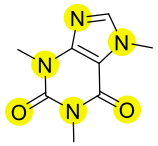
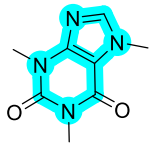
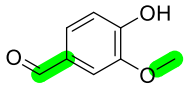
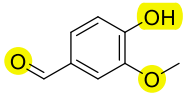
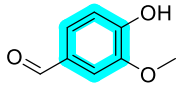
name *String*	query_smarts *String*
carboxylic_acid	[$([OX1H0-,OX2H1][CX3&H1](=O)),$([OX1H0-,OX2H1][CX3](=O)[#6])]
aldehyde	[CH1](=O)[#6]
aromatic_OH	[c;R]([OH1])

KNIME Workflow

RDKit



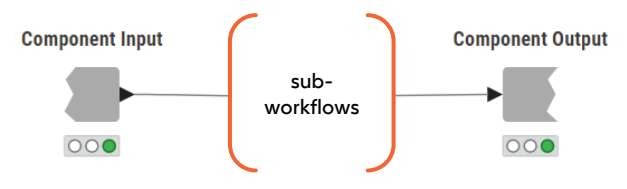
Name *String*	SMILES *Smiles*	pSMILES (RDKit Mol) *RDKit Molecule*	Salt Stripped Molecule *RDKit Molecule*
Caffeine	Cn1cnc2n(C)c(=O)n(C)c(=O)c12		
Vanillin	COc1cc(C=O)ccc1O		

ExactMW *Number (double)*	NumRotatableBonds *Number (integer)*	NumHeteroAtoms *Number (integer)*	NumAromaticRings *Number (integer)*
194.08	0	6	2
			
152.047	2	3	1
			

Upskilling with KNIME Courses

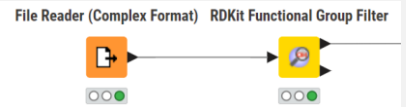
L4-CH: Introduction to Working with Chemical Data

Components
Organizing the workflow



- Cleaner workflow
- Grouping of nodes per sub-workflows BB MPO Scoring, Flagging, Filtering
- Duplication of similar processes

RDKit: Functional Group Filters and Similarity Searches

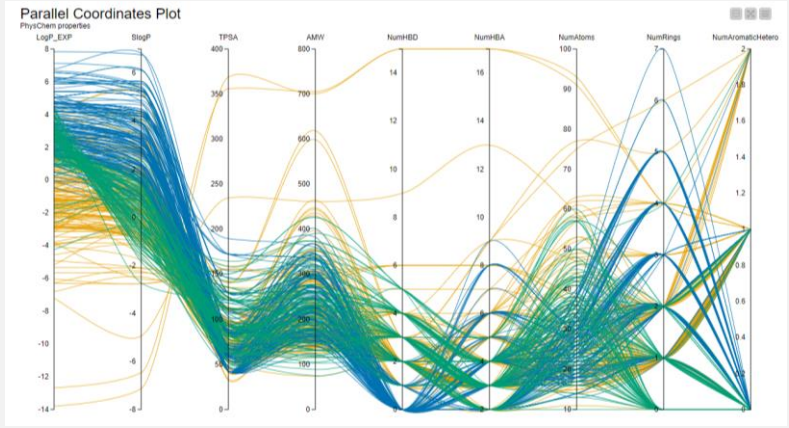


List of available functional group filters

Active	Functional Group	Qualifier	Count
☐	Acid Chloride	=	0
☐	Aromatic Acid Chloride	=	0
☐	Aliphatic Acid Chloride	=	0
☑	Carboxylic Acid	=	1
☐	Aromatic Carboxylic Acid	=	0

- Easy selection of fn. groups at specific counts BB Fn. Groups (Desirable and incompatible)
- Customizable node can be used with uploaded SMILES list

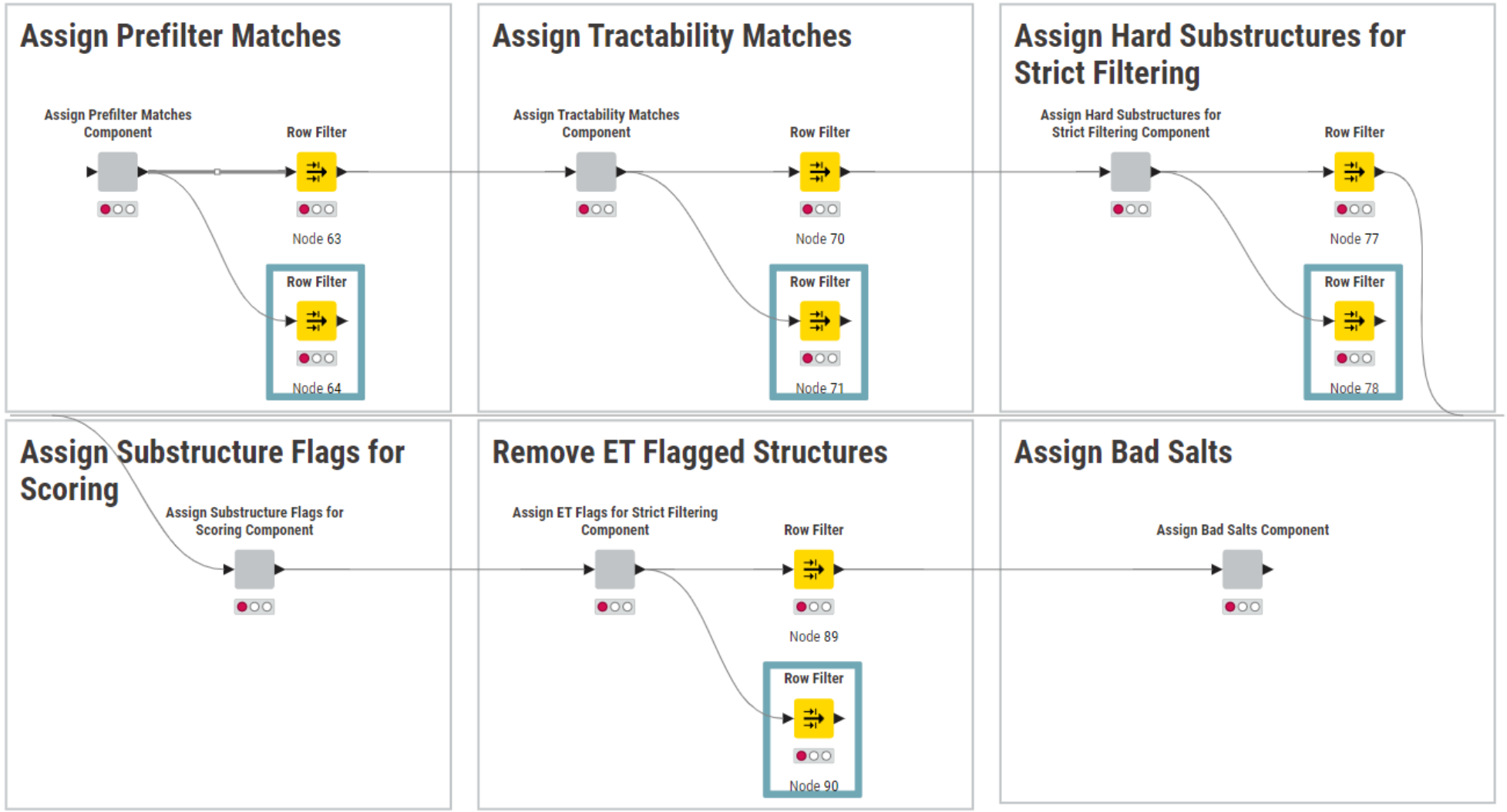
Java Nodes
Visualizing Data



- Powerful visualization of large data sets
- Allows user to make informed decisions
- Multiple properties can be visualized per compound at one time (Parallel Coordinates Plot) BB Properties
- Property distribution of an entire data set (Histogram) DEL Cycle

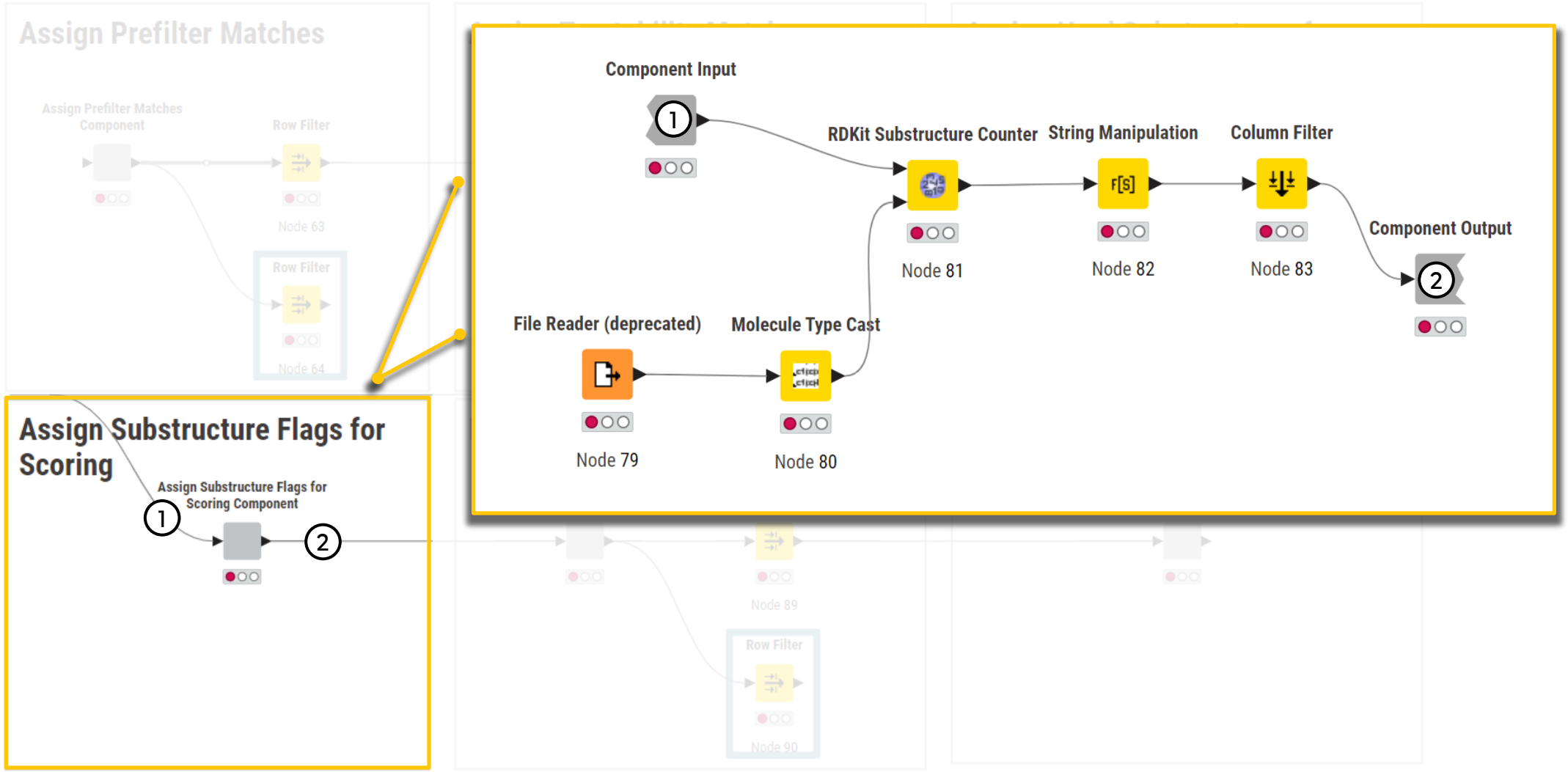
Building Block Flags

Using Components to Simplify the Workflow



Building Block Flags

Using Components to Simplify the Workflow



Prepared SMILES generation for DELs

GSK

Prepared SMILES generation for DELs

Setting up DEL definitions in our database

Carol Mulrooney

- Building blocks are enumerated to full structures by preparing their SMILES, then concatenating their prepared SMILES strings using Python code.

Prepared SMILES: CC(=O)*2* + NC **Enumerated SMILES:** CC(=O)NC

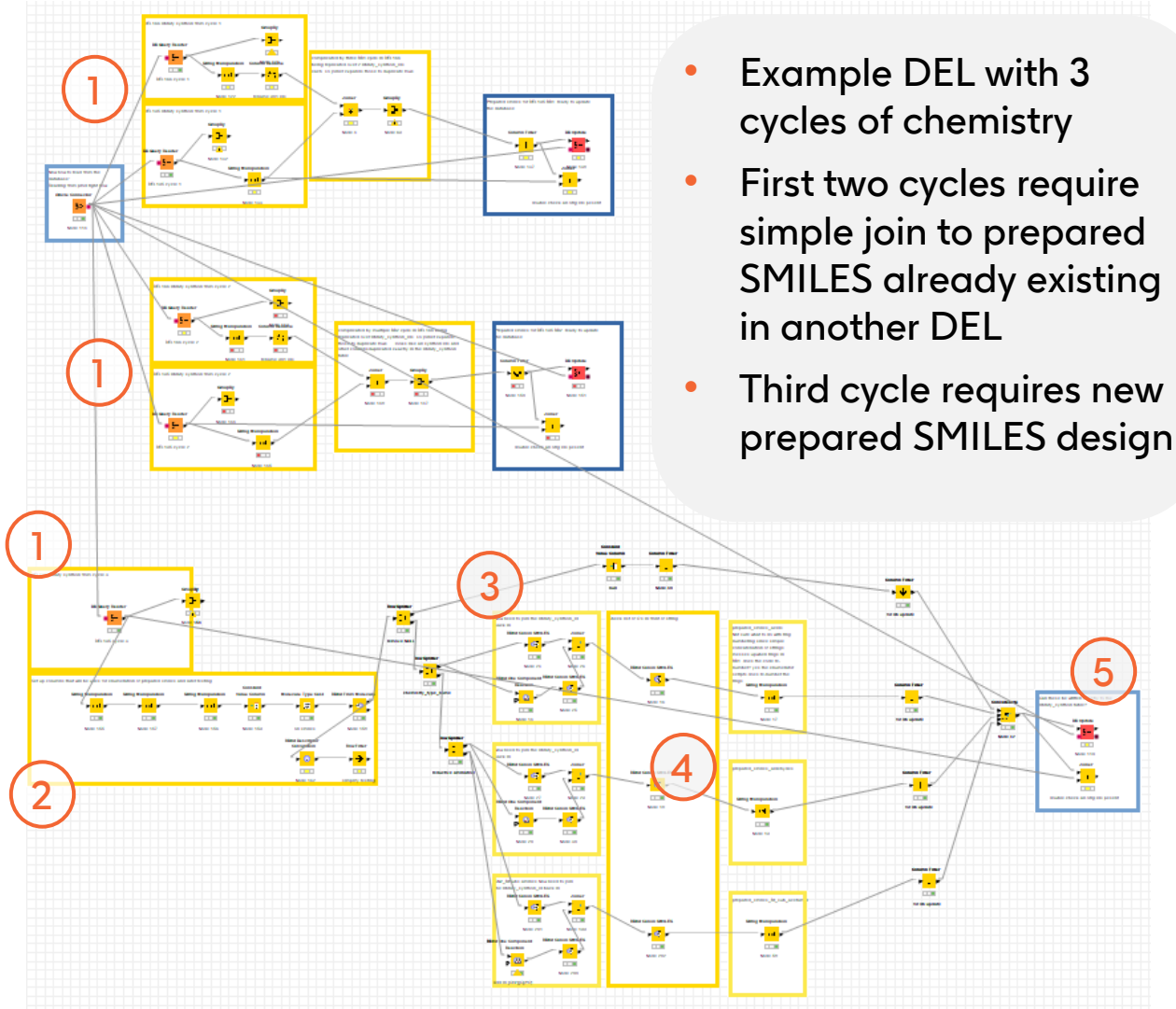
The ET platform has a tool to prepare SMILES for thousands of building blocks, over 100 DELs with 2-4 cycles chemistry

<u>But the current process results in many errors during enumeration of DEL structures:</u>

- Our DEL designs have evolved over the years from when the tool was developed
- The tool doesn't track different chemistry reaction types registered within a cycle, and building blocks with multiple reacting centers that may react differently depending on the chemistry type
- The tool doesn't provide for testing fully enumerated structures until the prepared SMILES are uploaded to database

Prepared SMILES generation for DELs

KNIME workflow to document preparation and avoid errors



- Example DEL with 3 cycles of chemistry
- First two cycles require simple join to prepared SMILES already existing in another DEL
- Third cycle requires new prepared SMILES design

1. Database query for library registration IDs and building block SMILES
2. Building block SMILES to RDKit molecule
3. RDKit 1 component reaction to prepare molecule and simulate the chemical reaction in the library synthesis
4. Generate canonical SMILES, remove unnecessary atoms using string manipulation, rename column to set prepared SMILES
5. Update database with prepared SMILES
6. Test the prepared SMILES before uploading to the database (not shown)

GSK

Summary

- DNA-Encoded Library (DEL) Design
  - Analyze, visualize, and select compounds for DELs
  - Enable chemists to work with their own data
  - Enumerate compounds and calculate physicochemical properties
  - Export data for multi-parameter optimization (MPO) scoring

- DEL Database Entry
  - Minimize error in library information database
    - Import DEL information directly from our database
    - Enumerate multiple different types of functional groups within the same cycle
    - Check structures and update the database


GSK
Ahead Together