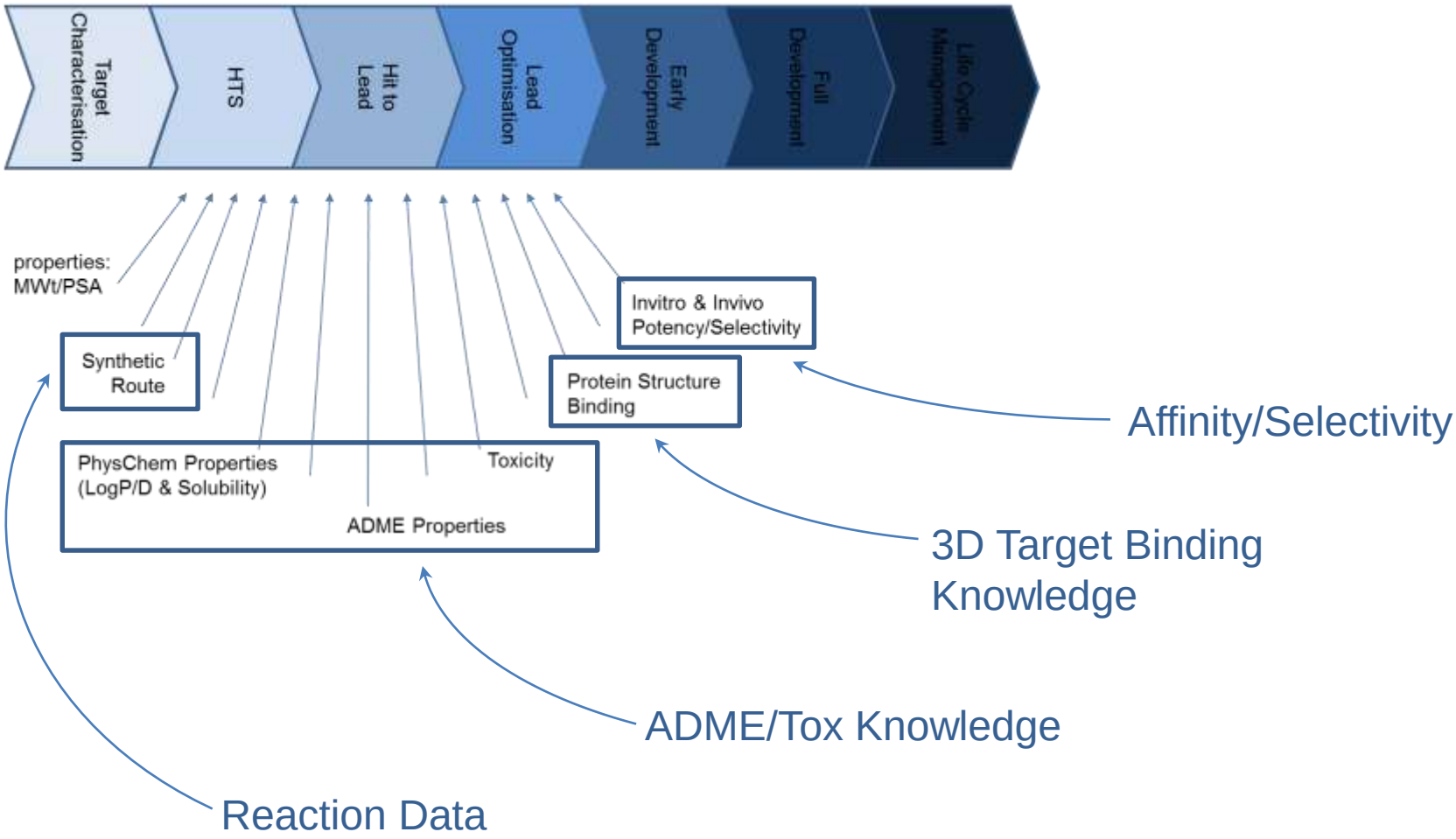
Transforming Bioactivity Data Stores into Knowledge Systems

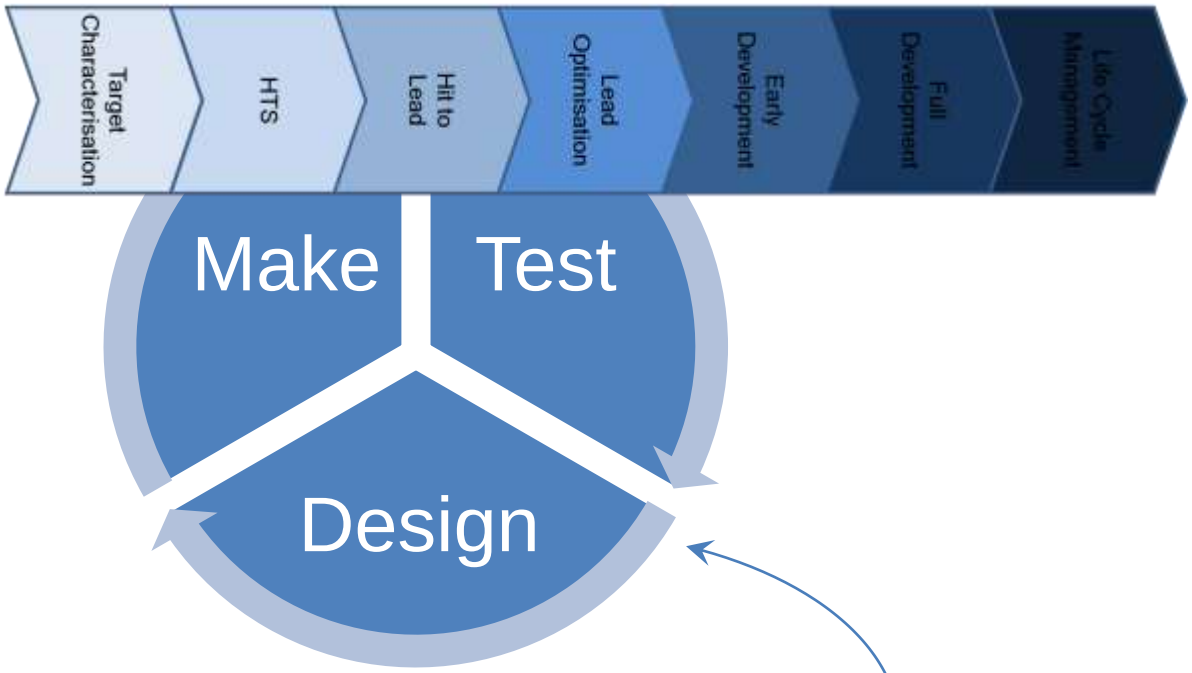
James A. Lumley, Research IT, Eli Lilly

Global Pharma R&D Informatics Congress – Dec 2017

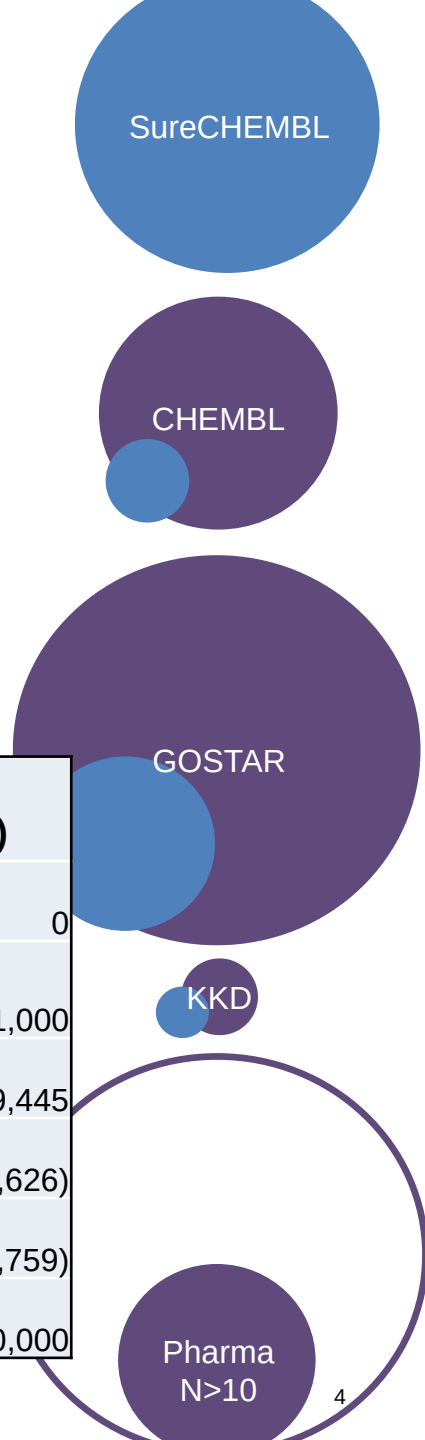
Motivation:
Customer + Consumable Data

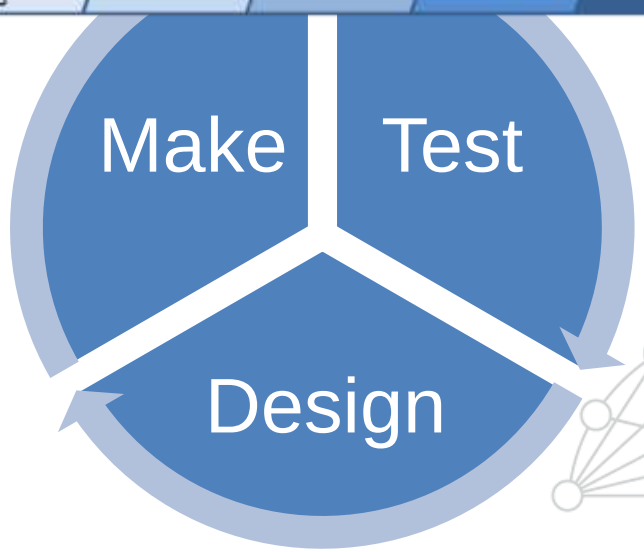
Knowledge Bases Impacting Small Molecule Discovery



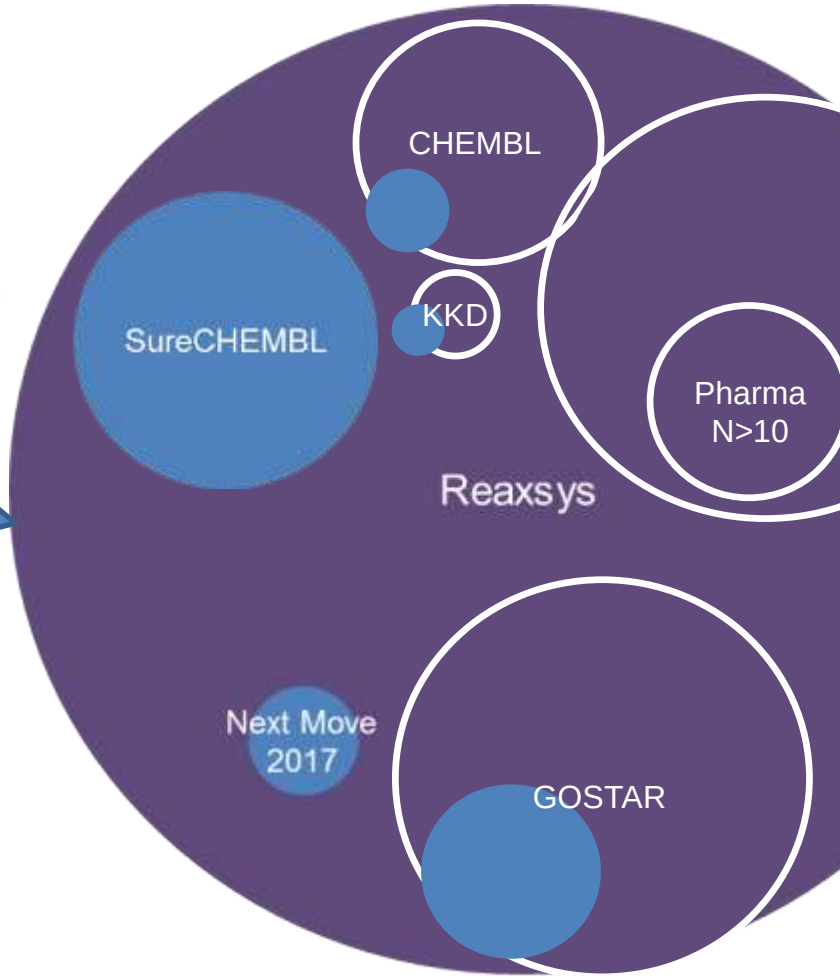


Database	Compounds (substances)	Bioactivities	Targets (Assays)
SureCHEMBL Nucleic Acids Research, 44, D1, 2016	17,000,000	0	0
ChEMBL 2017 https://www.ebi.ac.uk/chembl/	1,735,442	14,675,320	11,000
GOSTAR 2017 https://www.gostardb.com/databases.jsp	6,600,000	22,000,000	9,445
Kinase Knowledge Database F1000Res. 2016; 5: Chem Inf Sci-1366.	682,289	1,775,368	500 (33,626)
Pharmaceutical Bioactivity Data MedChemComm 2017, 8, 2067-2078	> 2,000,000?	5,350,628	(7,759)
Reaxsys MedChem (closed source) https://www.elsevier.com/solutions/reaxys	(6,500,000)	33,500,000	20,000





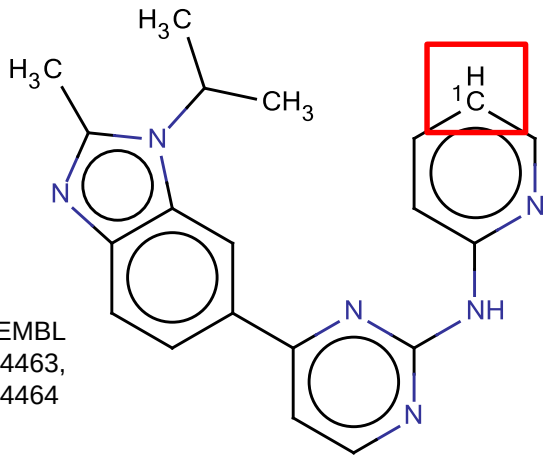
Augmented Intelligence



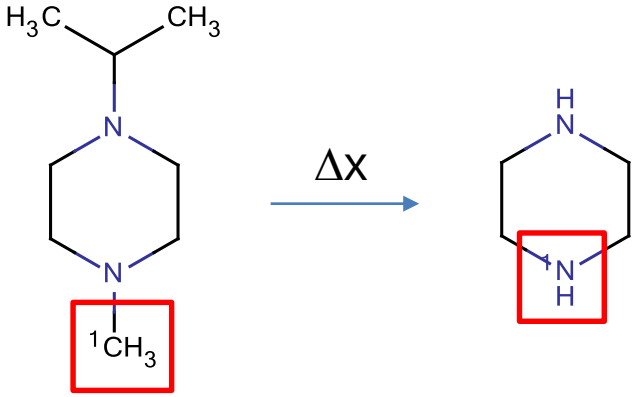
Making it Work: Transforming Datasets via MMP/S Generation

Matched Molecular Pair Generation

Fast, unsupervised approach after Hussain and Rea (JCIM 2010 22,50,339-48). Bonded atoms split/fragmented are tagged with isotopic label. Pairs found via string match on canonicalised, isotopically labelled smiles. Supports atom attachment point.



CHEMBL 3604463, 3604464



CHEMBL 3604463

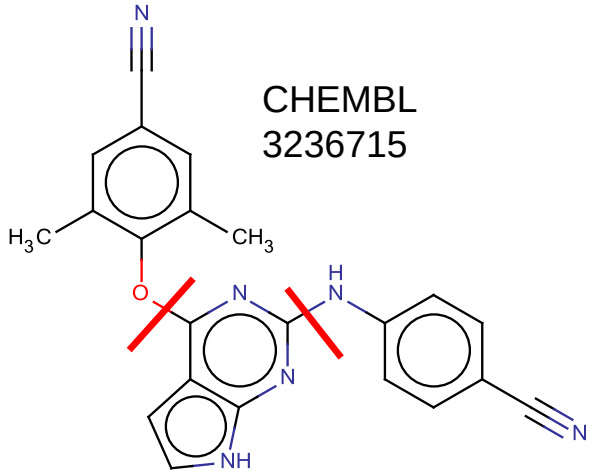
CHEMBL 3604464

Context:
Fc1c(c2[n](C(C)C)c([n]c2)C)[n]c(Nc2ccc(**[1CH]**=O)cc2)[n]c1

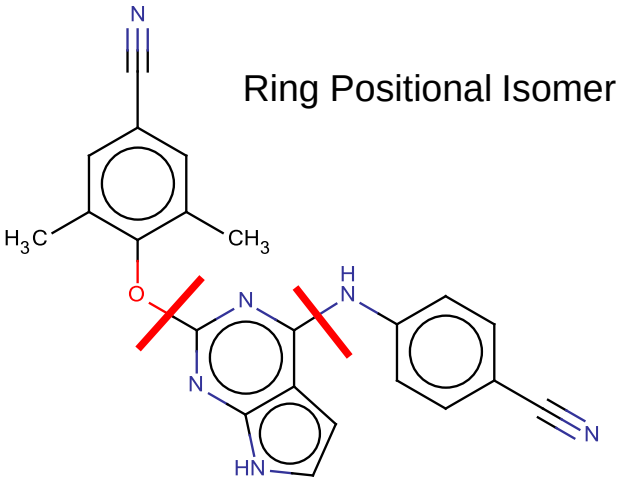
Fragment (Pair):
CHEMBL3604463: **[1CH]**N1CCN(C(C)C)CC1 → CHEMBL3604464: CNC1C**[1NH]**CC1

More Complex Transformations

Allows subtle matching of positional isomers and removes false positive matches, for example:



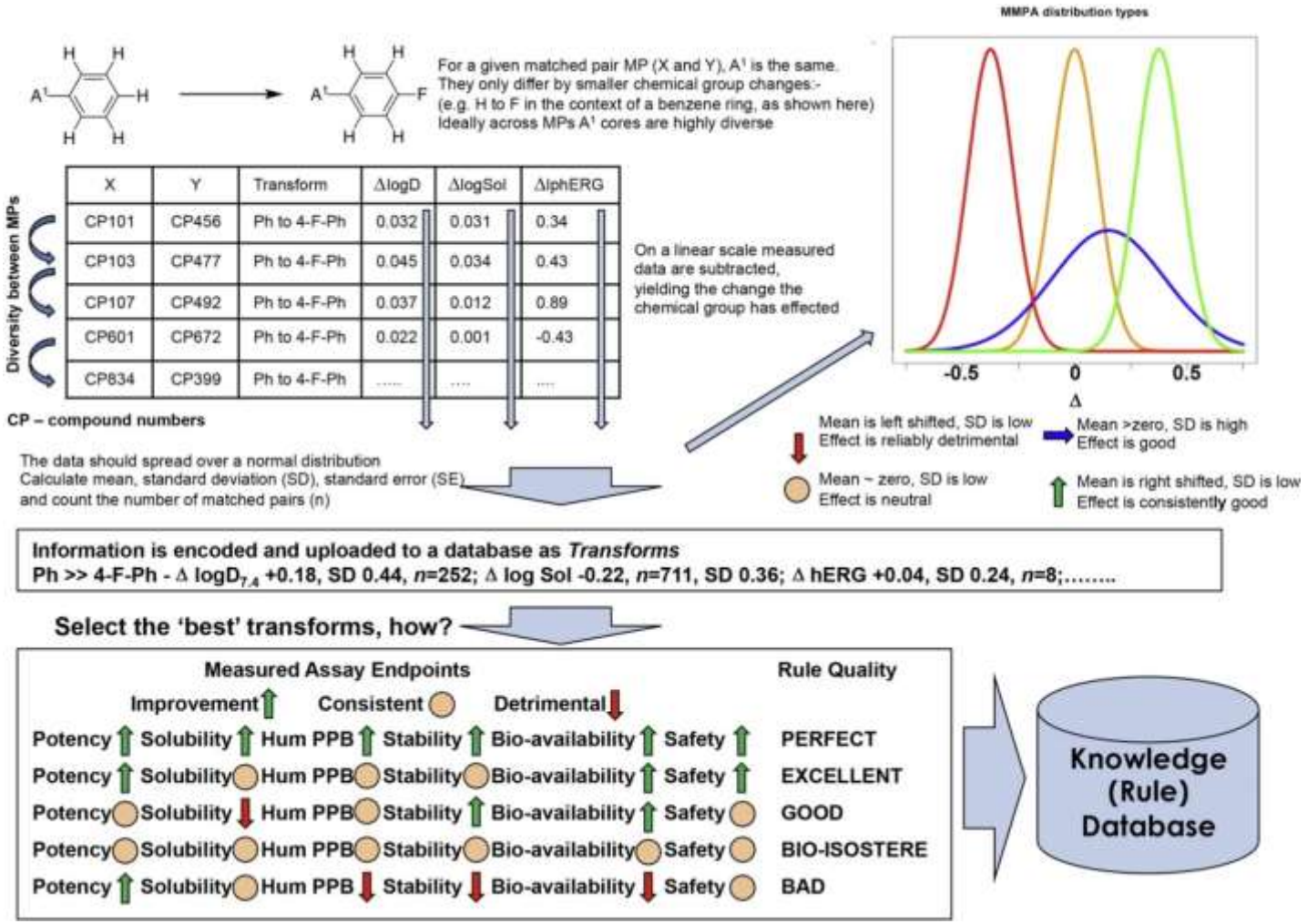
CHEMBL
3236715



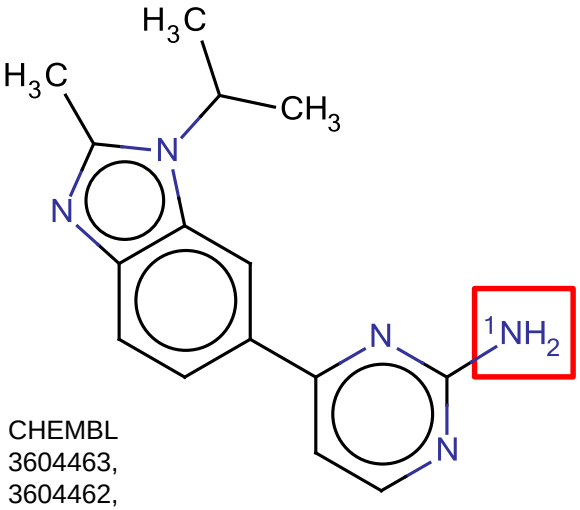
Ring Positional Isomer

Context: **[1OH]**c1c(C)cc(cc1)C#N.**[2NH2]**c1ccc(C#N)cc1
Fragment Pair: [nH]1c2c(cc1)**[1cH]**[n]**[2cH]**[n]2 → [nH]1c2c(cc1)**[2cH]**[n]**[1cH]**[n]2

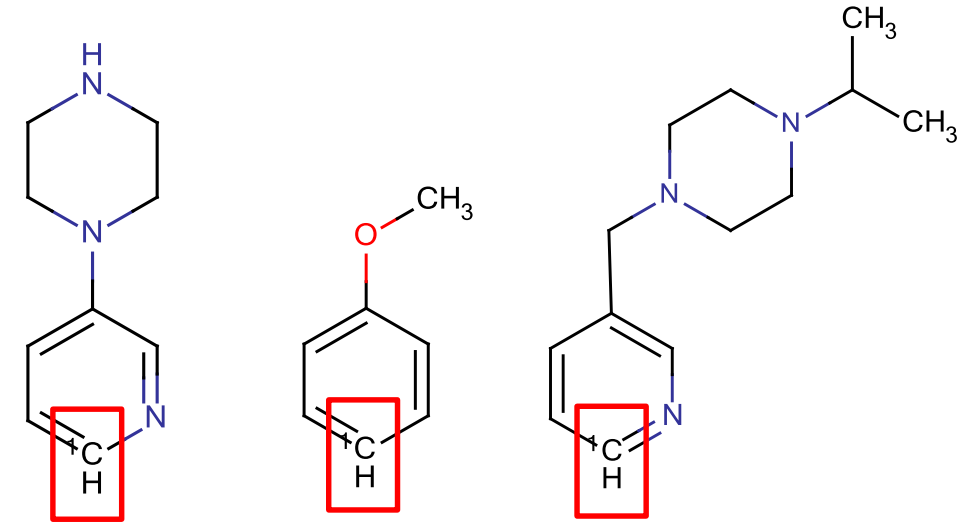
- Summarisation of an MMP across many molecule pairs derives a Transform. Transforms encode and formalise Tacit Knowledge in a highly visual form to aid Chemist in deciding “what to make next”



Series Generation



CHEMBL 3604463, 3604462, 3604464



CHEMBL 3604463 **8.7** | CHEMBL 3604462 **8.4** | CHEMBL 3604464 **6**

Increasing Activity (pIC50)*

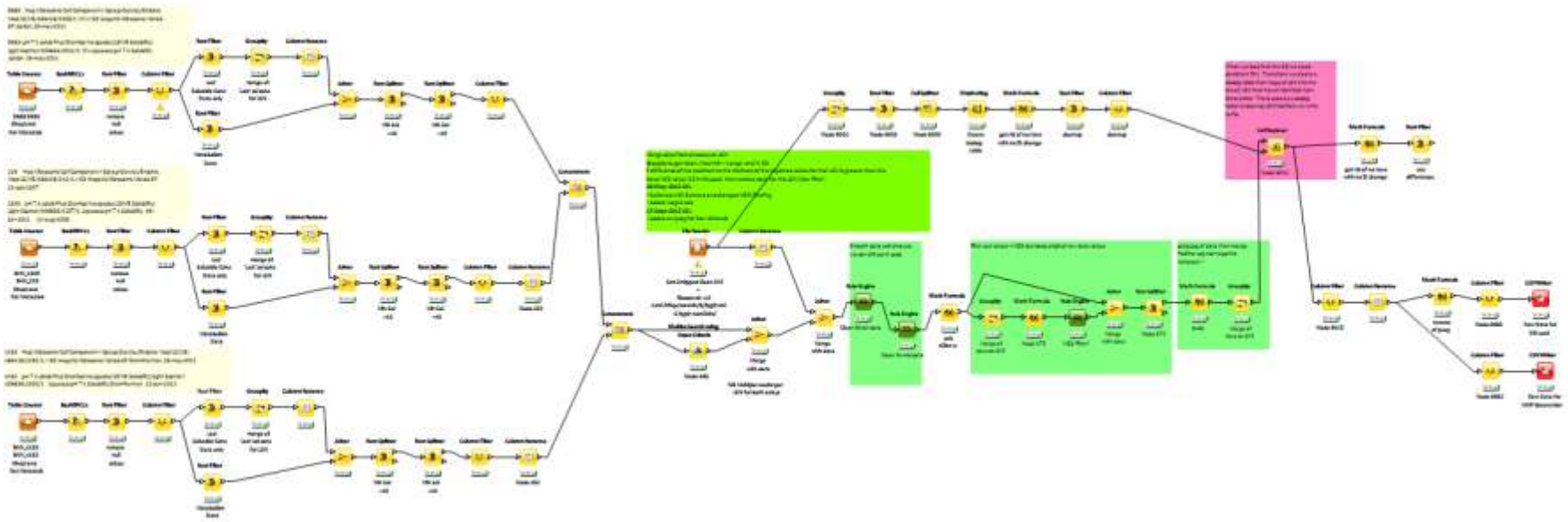
For Bioactivity data where changes are dominated more by 3D structural effects, longer series show better predictive success than pairs

J. Med. Chem., 2014, 57 (6), pp 2704–2713

* Order by pIC50 CDK1/CyclinB [Bioorg Med Chem Lett (2015) 25:3420-3435]

1. Cleaning and Transforming the data is time consuming: Workflow Systems Help

Visual programming style of KNIME an excellent way to collaborate with each Assay ‘owner’ to capture method for cleaning / filtering data on an assay by assay basis. Example:

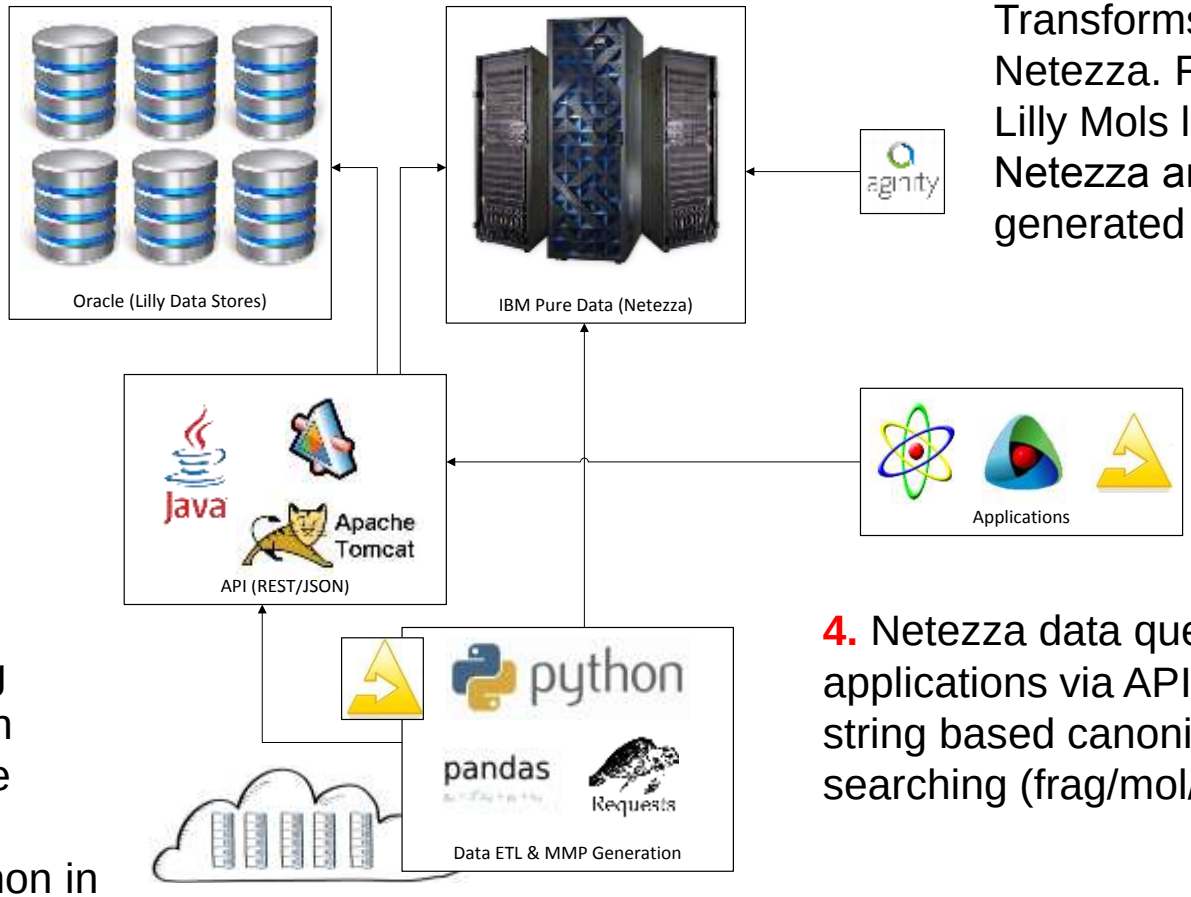


1. Retrieve data, remove null values
2. Remove compounds with poor solubility
3. Merge across assays then across unique structures (remove salts)
4. Smooth data extremes or apply other data cleaning tasks
5. Remove outliers via MSD based filter
6. Aggregate to Mean pLog/pIC50 for each molecule

2. Secure, Scalable Provisioning of Raw and Transformed Data is hard: SOA Helps

140 RESTful Webservice API endpoints deployed in 2-3 years primarily for data retrieval and workflow development, 25 necessary for MMP/S work

SOA allowed seamless iterations scaling the data backend



1. Data Extraction from oracle via API layer. Complex ETL in KNIME

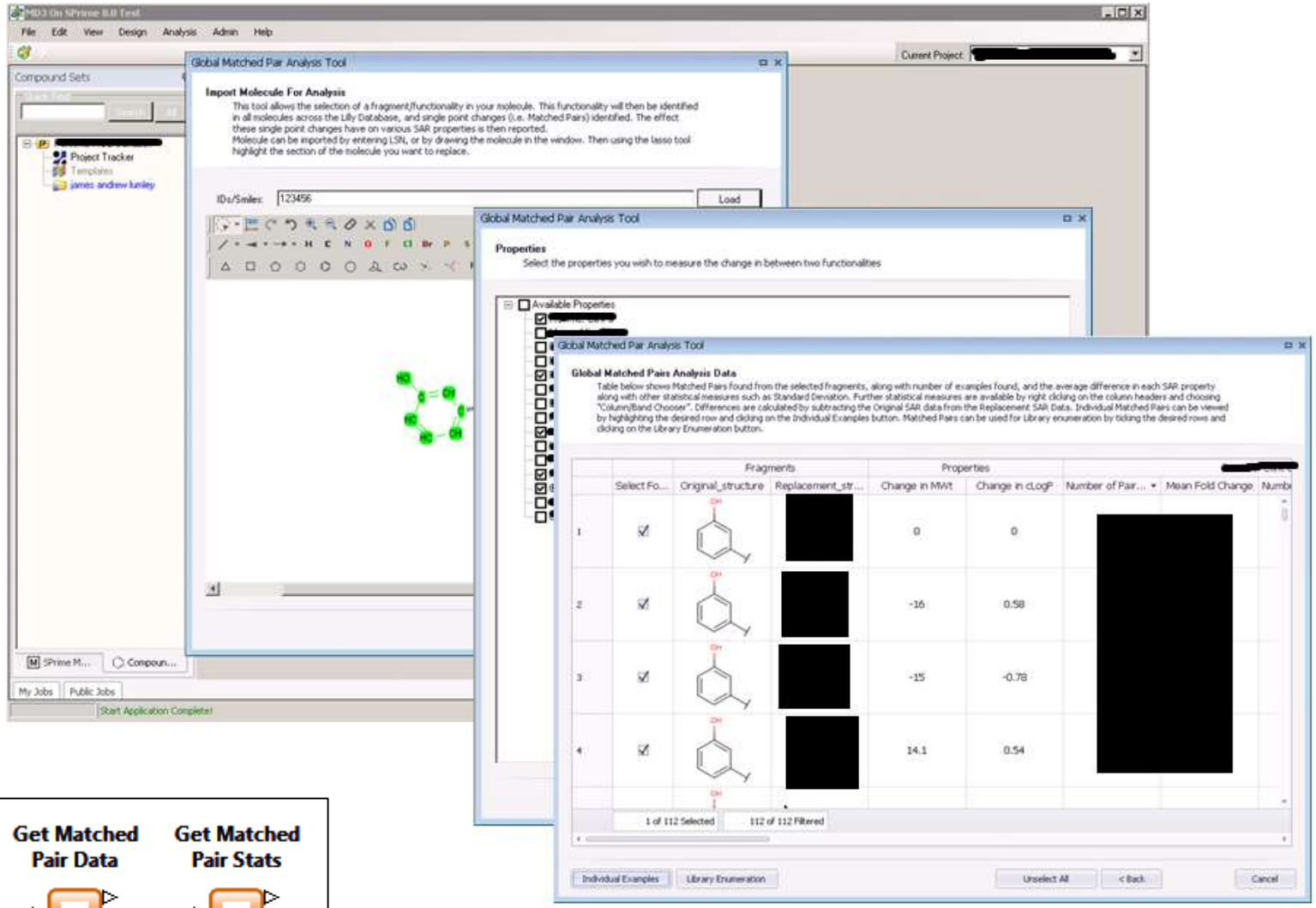
2. Data Processing workflows scaled in KNIME and include MMP/Transform generation via python in HPC/cloud

3. ADME MMP's and Transforms loaded into Netezza. Fragmented Lilly Mols loaded into Netezza and MMP's generated via SQL

4. Netezza data queryable by applications via API using string based canonical SMI searching (frag/mol/LSN)

5. System rebuilt weekly

3. Good Interfaces Matter: Multiple Interfaces drive new use cases



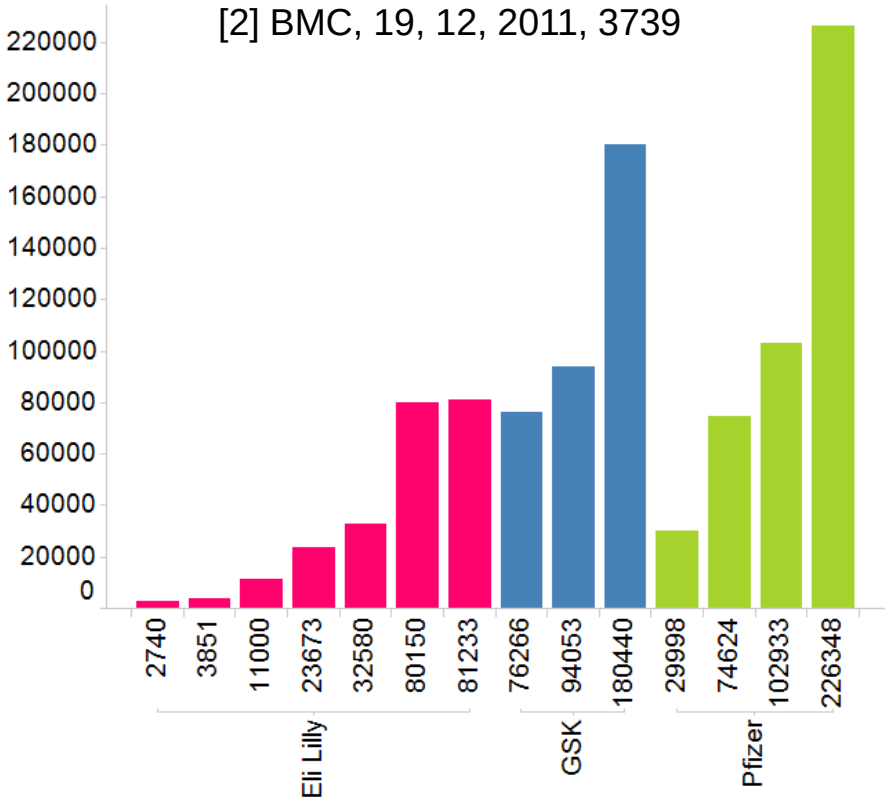
KNIME Nodes:

How good is our Knowledge?

Dataset sizes (ADME):

- >30 Lilly ADME/Tox assay datasets
- Including
  - Hepatocyte Unbound Clearance
  - Microsomal Unbound Clearance
  - CYP Inhibition
  - Permeability
  - Solubility
  - LogP/LogD
  - hERG
  - Other key ADME/Toxicity assays
- 10s of 1000's of IC/EC50 & Ki endpoints
- In progress:
  - Biochemical pIC50

7 example Lilly datasets versus literature sets:
[1] JCIM, 2010, 50 10, 1872
[2] BMC, 19, 12, 2011, 3739

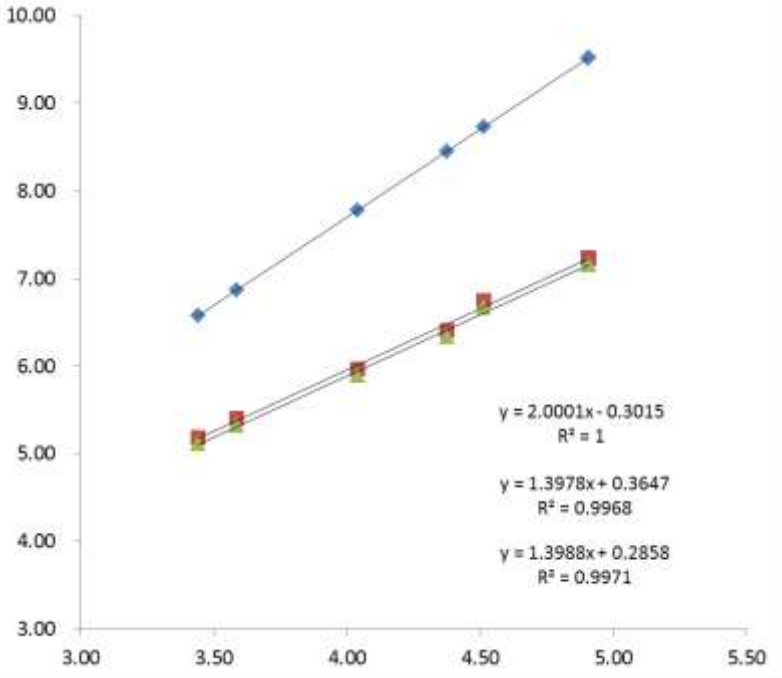


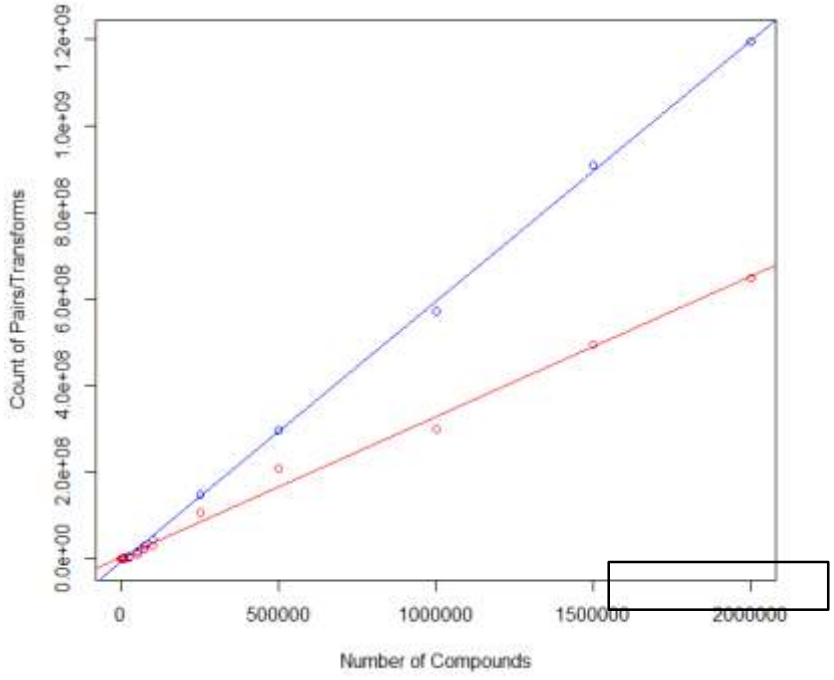
How do we assess how good our Knowledge Base is?

- **Breadth (covers all project endpoints)**:
  - hERG, Solubility, LogD, CYP (GSK, JCIM, 2010, 50 10, 1872; BMC 2009, 17, 5906)
  - Solubility, PPB, Oral Exposure, Microsomal Met. Stab. (AZ, JMC, 2006, 49, 23, 6672; MedChemComm, 2012, 3, 1518)
  - Microsomal Clearance, MDCK Perm, P-gp efflux, LogD (Pfizer, BMC, 19, 12, 2011, 3739–3749)
  - Melting point (Boehringer, MedChemComm, 2012, 3, 584)
- **Depth (statistical validity)**:
  - Reliability, Significance, Outliers… N, SD, paired *t* test (JMC, 2014, 57, 9, 3786–3802)
  - In-house metrics quantify the non-lipophilicicity contribution to a delta
- **Diversity (covers all possible chemical changes)**:
  - Unexplored

Quantifying What We Don't Know:

- All Lilly ADME datasets show linear scaling of MMP & Transform count with increasing N value (# compounds with raw data)
- For a 2M compound set we see linear scaling of derived MMPs/Transforms with a total of 1.1Bn MMP's and **0.65Bn** unique transforms





Quantifying What We Don't Know:

Compare the number of Transforms in ADME Knowledge Base to the 2M Lilly subset and the top 5K Lilly 2M subset:

646,721,257	Unique Transforms in 2M subset
20,277,018	Unique Transforms in ADME/Tox set

3.1%	Poor coverage of possible Transform space

Transforms limited to a maximum of 15 atom fragment pairs

Only ~1% are significant at N>6

Reymond et. al (MedChemCommun, 2010, 1, 30) quantified size of 15 atom chemical space via the enumeration of a database of chemically stable and synthetically feasible, C, N, O and Cl containing molecules → 28.8 billion

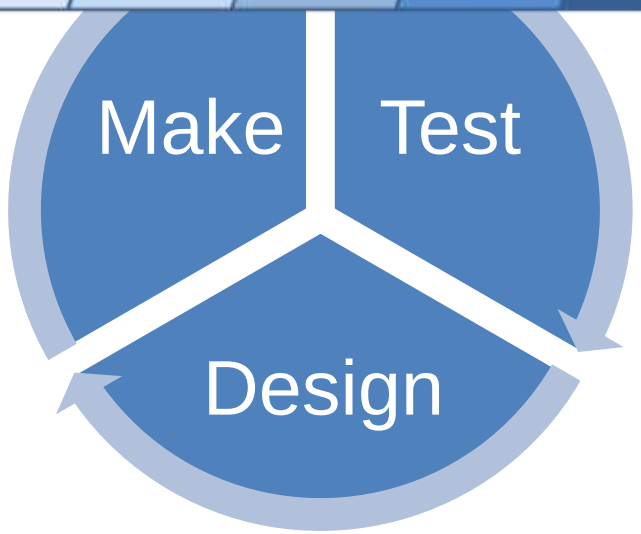
Things we Know Well:

Compare the number of Transforms in ADME Knowledge Base to the 2M Lilly subset:

646,721,257	Unique Transforms in 2M subset
20,277,018	Unique Transforms in ADME/Tox set

3.1%	Poor coverage of possible Transform space
5000	Top 5K Lilly Transforms in 2M subset
4768	Transforms in ADME/Tox set seen in top 5K

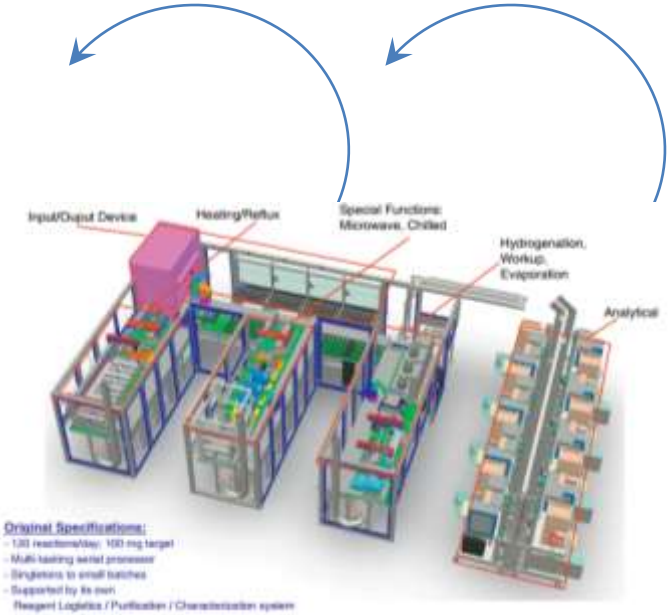
95.4%	Good coverage of Top 5K in possible Transform space suggests that common transformations are well covered but huge gaps in our knowledge of Transform space



Propose Transform ‘Sets’

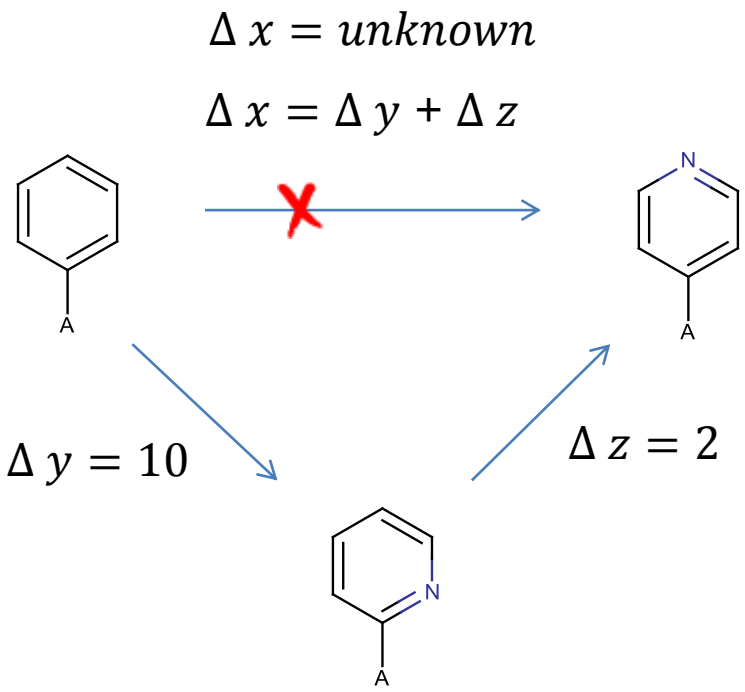
(a) Smart iteration, better quantified outcomes

(b) Unknown or less certain



Graphs as an enrichment tool:

Poor coverage of Transform space can be augmented by graph interpolation / multiplicative transforms (Keefer et. al. BMC 19, 2011, 3739–3749):



Graphs as an enrichment tool:

How well does our coverage improve as we apply graph interpolation?
Enumerate and count all multiplicative Transforms for a given ADME/Tox assay and compare to the 2M Lilly subset:

Assay	Data points	Transforms	Transforms (Interpolated)
ADME Assay 1	80,345	14,413,790	4,849,307,796
ADME Assay 2	32,720	4,616,806	838,896,430
ADME Assay 3	23,721	2,178,950	198,432,174
ADME Assay 4	11,003	781,720	49,195,320
ADME Assay 5	2,671	127,914	3,147,526

Unique Transforms in 2M subset	646,721,257

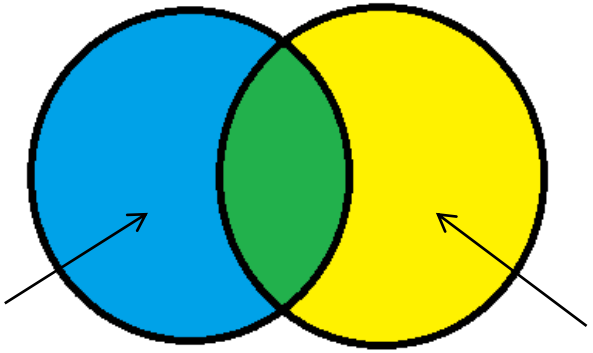
Graphs as an enrichment tool:

How well does our coverage improve as we apply graph interpolation?
Consider the coverage for a single ADME Assay using both raw and interpolated pairs versus the Lilly 2M subset. Inclusion of the Interpolated Transforms for ADME Assay 1 gives us a 10% coverage of Lilly 2M subset!

ADME Assay 1:

- 14,413,790 Transforms, 10,298,150 in Lilly 2M subset
- 4,849,307,796 Interpolated Transforms, 49,634,260 in Lilly 2M subset

ADME Assay 1:
(10,298,150 + 49,634,260)
59,932,410 Transforms



Lilly 2M subset
646,721,257 Transforms

9.3% Overlap for single assay

Summary:

- Augment chemical design with well understood chemical Transforms
- Good data / data analytics tools e.g.: KNIME accelerate team work
- Architecture Matters and done well opens up new use cases
- Defining the limits of your Knowledge helps understand how and when to apply it with confidence
- Graph based techniques and other methods offer unique approach to enrich data and fill ‘gaps’ in transform space

Acknowledgements

Research:
Jibo Wang
Prashant Desai
Jaclyn Barrett
Brianna Paisley
Dave Evans
Lewis Vidler
Luca Fenu

Statistics:
Suntara Cahya

Advanced Analytics:
Ian Watson

Research IT:
Luke Bullard
Nathan Roberts
Thomas Wilkin
Hongzhou Zhang