Ultra-Large Libraries and Chemical Spaces of Virtual Screening Samples with Proposed Synthetic Routes

Marc C. Nicklaus

Actyon Discovery, Inc.

XXXI Symposium on Bioinformatics and Computer-Aided Drug Discovery (BCADD-2025), 2025-10-21

Disclaimer

The views and opinions presented here represent those of the speaker. They are not the view of, nor speaking for, my former employer NCI, NIH; nor of my new company, Actyon Discovery, Inc.
The results reported here are from published papers or preprints co-authored by the speaker.

Motivation: Very large chemical spaces becoming more numerous (see, e.g., https://www.biosolveit.de/chemical-spaces/); but enumerated libraries still exist. Does it have to be one or the other?

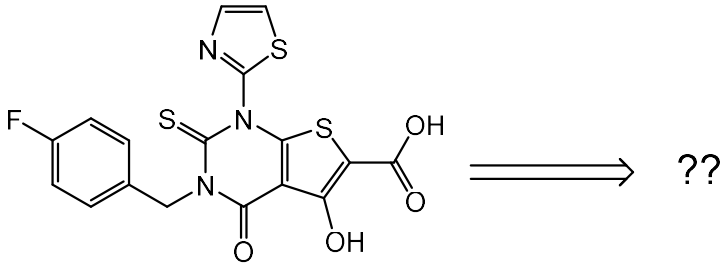
Outline:

1. SAVI (2020 Library)
   (Synthetically Accessible Virtual Inventory)
2. SAVI Space
3. SLICE
4. Comparison of enumerated libraries with chemical spaces

Why SAVI?

Maybe We're Asking the Wrong Question in CADD

Designed molecule may or may not be synthetically tractable



⟹ ??

quoted: $900 for 50 mg -- but no predicted synthetic route;
couldn't be synthesized (at that cost)

Why do we grapple with this kind of problem?

Maybe we should instead ask:
What can I make easily, reliably, and cheaply?

Exploiting Decades of Research into Synthesis Prediction

SAVI Methodology:

- Derived from the LHASA project[1]; which started in late 1960s/early 1970s
- Expert-system type approach
- Language pair CHMTRN & PATRAN[2]
- LHASA: retrosynthetic; SAVI: forward-synthetic

Advantages:

- Does not need large training database to add new transform: Can easily add new rules
- Goes way beyond the typically used SMIRKS transforms
- Has scoring and KILL capability

Sonogashira coupling:
[#6;$(C=C−[#6]), $(c:c):1] [Br,I].[CH1;$(C#CC):2] ≫ [#6:1] [C:2]

Does this really tell you all about the “chemical logic” of all possible Sonogashira reactions?

[1] LHASA—Logic and Heuristics Applied to Synthetic Analysis, Pensak and Corey, *ACS Symp. Ser.*, 1977
[2] Judson *et al.*, *JCIM,* 2020 60 (7), 3336-3341.

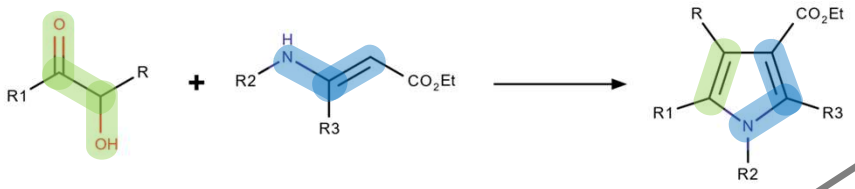
CHMTRN/PATRAN Transform Excerpt

```
                TRANSFORM 2267
                NAME Sonogashira Coupling
...
                R-C#C-R1  ==>  R-X  +  H-C#C-R1

                R = Aryl or Vinyl; X = Br,I
                ...RATING 60
                TYPICAL*YIELD              VERY*GOOD
                RELIABILITY               GOOD
                REPUTATION                GOOD
                HOMOSELECTIVITY           FAIR
                HETEROSELECTIVITY         GOOD
                ORIENTATIONAL*SELECTIVITY     NOT*APPLICABLE
                CONDITION*FLEXIBILITY       FAIR
                THERMODYNAMICS            GOOD
1D*PATTERN
                C[RINGS=NO]#C-C[ARYL=YES]
1D*PATTERN
                C[RINGS=NO]#C-C=C
...
...1D*PATTERN
...             C[RINGS=NO]#C-C[HS=2]-C[ARYL=YES]
...1D*PATTERN
...             C[RINGS=NO]#C-C[HS=2]-C=,#C
NEW*1D*PATTERN
                C[RINGS=NO]#C-C[ARYL=YES] => C^1[RINGS=NO]#C^2[HS=1]
                + Cl,Br,I-C^3[ARYL=YES]
NEW*1D*PATTERN
                C[RINGS=NO]#C-C=C => C^1[RINGS=NO]#C^2[HS=1] + Cl,Br,I-
                C^3=C
2D*PATTERN

                        H   Br,I,Cl
                        |   |
                C#C-C%,=C => C#C + C%,=C

END*PATTERNS
```

```
                KILL IF ANYWHERE THERE IS AN ACID*HALIDE OR: ANHYDRIDE OR:
                EPOXIDE OR: ISOCYANATE
                FOR EACH IODINE ATOM ANYWHERE DO I_CHK
                  IF THERE IS ONLY ONE ATOM ALPHA TO SPECIFIED*ATOM 1
                THEN KILL
...
                TURN OFF BIT 1
                FOR EACH BROMINE ATOM ANYWHERE DO BR_CHK
                  IF THERE IS ONLY ONE ATOM ALPHA TO SPECIFIED*ATOM 1
                AND:IF  &
                    AN ATOM BETA TO SPECIFIED*ATOM 1 IS MULTIPLY BONDED
                THEN  &
                    TURN ON BIT 1
...
... Qualifiers for R group.  If the halide is aromatic
... withdrawing groups on the ring activate and vice versa.
                IF ATOM*3 IS AN AROMATIC ATOM
                BEGIN BLOCK1
                  DESIGNATE AS THE CURRENT*RING  &
                    THE RING CONTAINING THE ATOMS ALPHA TO ATOM*3
                OFFPATH
                  SAVE AS 1 THE ATOMS ON*CURRENT*RING
                  ADD 10 IF THERE IS A WITHDRAWING BOND  &
                    ON SAVED*ATOM 1 OFF*CURRENT*RING
                  RAISE*RATING SLIGHTLY IF THERE IS A WITHDRAWING BOND  &
                    ON SAVED*ATOM 1 OFF*CURRENT*RING
                  SUBTRACT 10 IF THERE IS A DONATING BOND  &
                    ON SAVED*ATOM 1 OFF*CURRENT*RING
                  LOWER*RATING SLIGHTLY IF THERE IS A DONATING BOND  &
                    ON SAVED*ATOM 1 OFF*CURRENT*RING
                  IF THERE ARE SIX ATOMS ON*CURRENT*RING
                  BEGIN BLOCK2
                    ADD 10 IF THERE IS A NITROGEN ATOM ON*CURRENT*RING
                    RAISE*RATING SLIGHTLY IF THERE IS A NITROGEN ATOM
                ON*CURRENT*RING
                  BLKEND BLOCK2
                BLKEND BLOCK1
```

SAVI-2020 Components

Transforms



LHASA transforms written in CHMTRN/PATRAN

Building Blocks

Enamine

Yurii Moroz (Ukraine)
155k building blocks

Philip Judson (UK)
53 transforms

SAVI

Product Generation
CADD Group

Hitesh Patel (US)

Chemoinformatics engine

CACTVS

Wolf D. Ihlenfeldt (Germany)

Xemistry chemoinformatics

SAVI-2020: Product Counts

"Plus" class: has not encountered any SUBTRACT.

Class	SAVI products	Unique	Percentage
Plus	**1,094,782,440**	**976,051,945**	**62.61%**
Neg0	609,262	579,532	0.03%
Neg10	54,775,204	48,036,148	3.13%
Neg20	82,180,372	80,366,188	4.7%
Neg30	516,116,725	457,508,945	29.52%
Total	1,748,464,003	1,526,316,392	

Non-uniqueness due to multiple possible synthetic routes for some molecules – welcome feature of SAVI.

Generating this billion+ SAVI database is computationally costly: >5 million CPU hours on NIH Biowulf cluster. It's also large (downloadable from https://cactus.nci.nih.gov/download/savi_download/):

- SDFs in total: 4.4TB
- SMILES tables in total: 1.1TB

Patel, H., Ihlenfeldt, WD., Judson, P.N. *et al.* SAVI, *in silico* generation of billions of easily synthesizable compounds through expert-system type rules. *Sci Data* **7**, 384 (2020). https://doi.org/10.1038/s41597-020-00727-4

SAVI-Space –

Combinatorial Encoding of a Billion-Size Synthesizable Virtual Inventory

- **SAVI-Space:** Design a **compressed, fragment-based and reaction-driven data structure** (Fragment Space) for improved accessibility.
  - **SAVI-Space**[1] encodes transformations combinatorially, designed to enhance speed, scalability, and accessibility
  - Created by Malte Korn (Matthias Rarey group, Univ. Hamburg)
  - Three different versions of SAVI-Space were generated

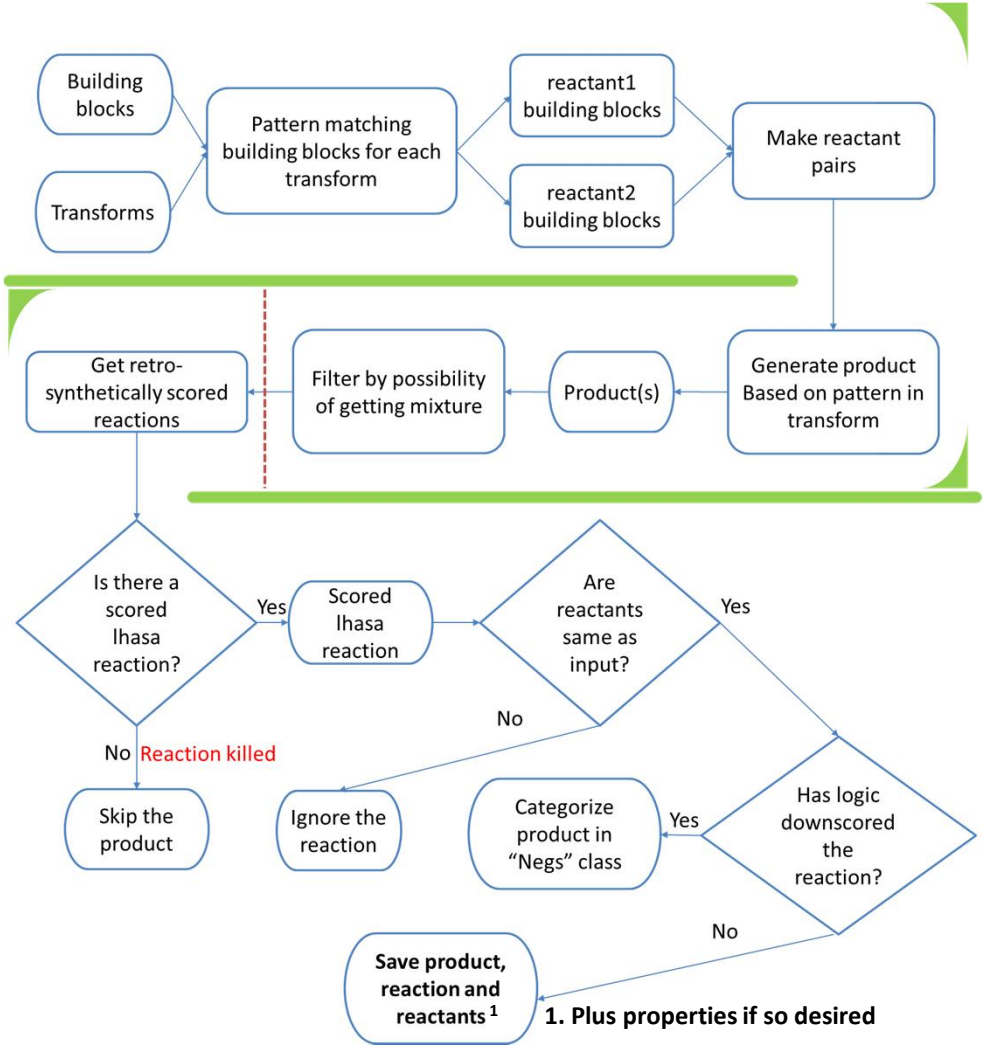
1. Korn, M., Judson, P., Klein, R. *et al.* SAVI Space—combinatorial encoding of the billion-size synthetically accessible virtual inventory. *Sci Data* **12**, 1064 (2025). https://doi.org/10.1038/s41597-025-05384-z

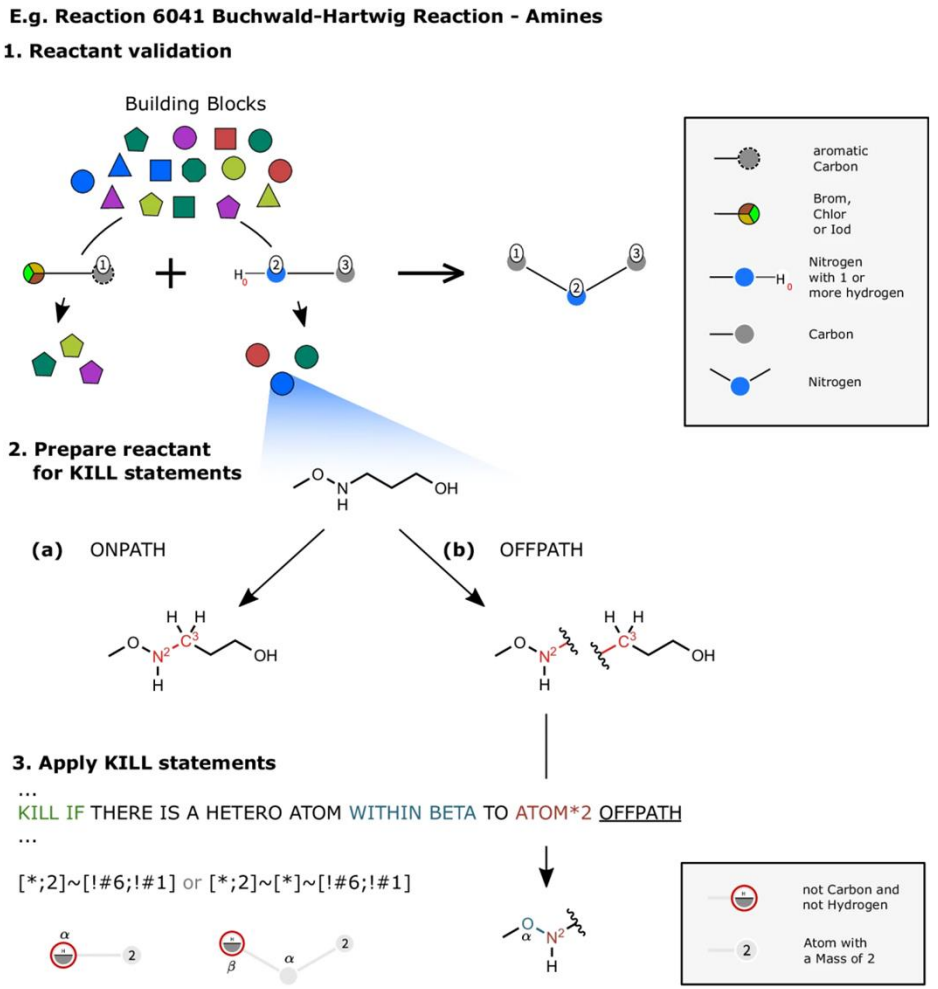
SAVI-Lib & SAVI-Space Versions

Version	Building Blocks Used	Number of Products	KILL Rate	Comput-ational Time/h	Disk Space Needed
SAVI(-Lib) 2020	143k[1]	1.75B	51%	5,000,000	SDF: 4.4T SMILES: 1.1T
SAVI-Space-2020 (Lib-2020 rules)	139k	2.34B	29%	3	2.1GB
SAVI-Space-2020	138k	2.40B	36%	3	0.8GB
SAVI-Space-2024	256k	7.55B	29%	10	1.4GB

1. Out of an original 155k building blocks provided by Enamine

Workflows

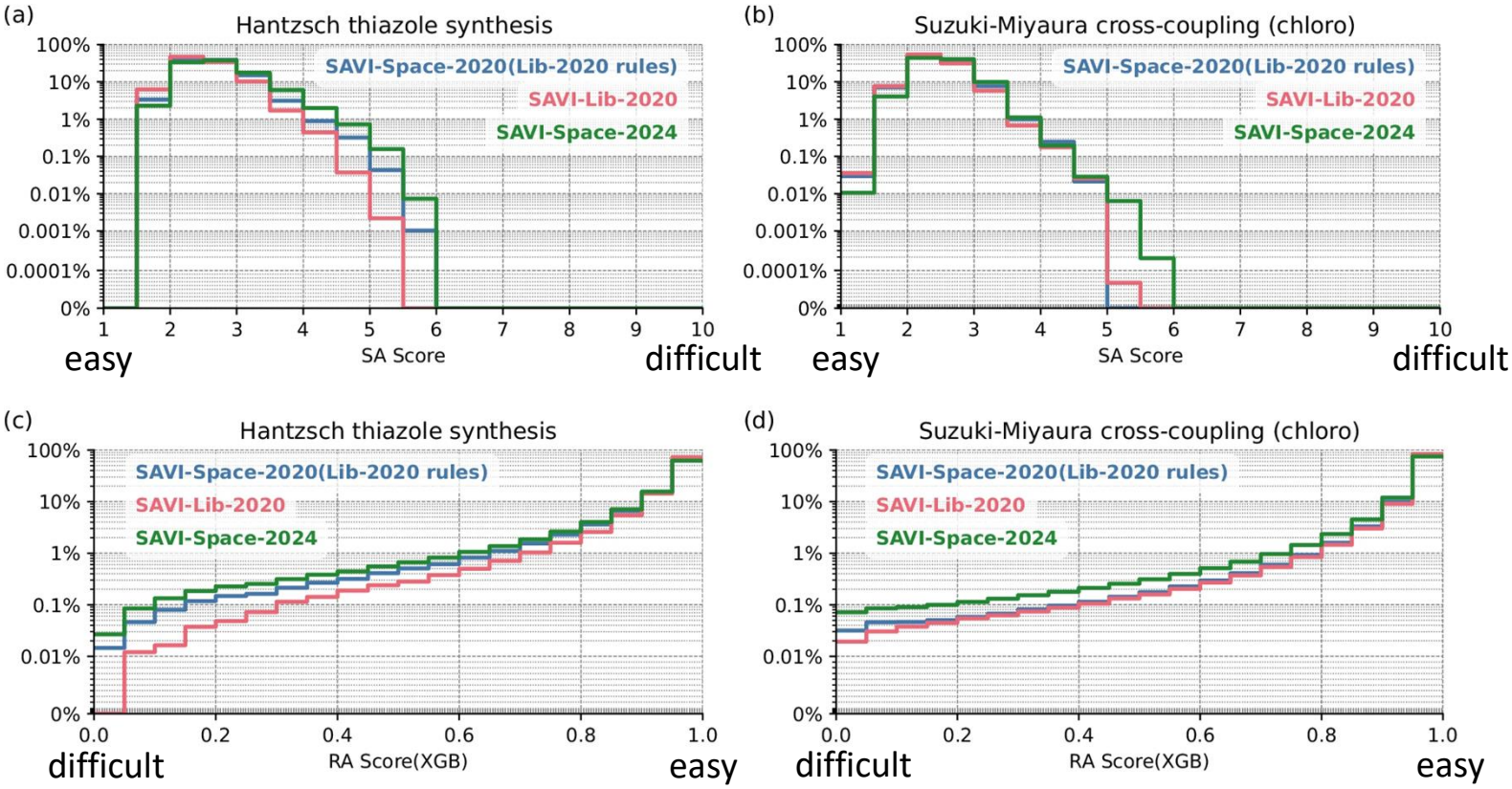

Building blocks
Transforms
Pattern matching building blocks for each transform
reactant1 building blocks
reactant2 building blocks
Make reactant pairs
Generate product Based on pattern in transform
Product(s)
Filter by possibility of getting mixture
Get retro-synthetically scored reactions
Is there a scored lhasa reaction?
Yes
Scored lhasa reaction
Are reactants same as input?
Yes
No
No Reaction killed
Skip the product
Ignore the reaction
Categorize product in "Negs" class
Yes
Has logic downscored the reaction?
No
Save product, reaction and reactants [1]
1. Plus properties if so desired


E.g. Reaction 6041 Buchwald-Hartwig Reaction - Amines
1. Reactant validation
Building Blocks
aromatic Carbon
Brom, Chlor or Iod
Nitrogen with 1 or more hydrogen
Carbon
Nitrogen
2. Prepare reactant for KILL statements
(a) ONPATH
(b) OFFPATH
3. Apply KILL statements
...
KILL IF THERE IS A HETERO ATOM WITHIN BETA TO ATOM*2 OFFPATH
...
[*;2]~[!#6;!#1] or [*;2]~[*]~[!#6;!#1]
not Carbon and not Hydrogen
Atom with a Mass of 2

Challenges encountered and (mostly met) in the path from SAVI-Lib to SAVI-Space

- Wrote transpiler translating LHASA transform patterns into reaction SMARTS patterns
- Atoms and bonds can have properties in LHASA patterns – difficult to replicate in SMARTS
- KILL statements:
  - Translated to filter rules applied to reactants
  - Created manually (300 KILLs)
  - KILLs may combine atoms from reactant A and reactant B
  - Some CHMTRN keywords are difficult/impossible to translate to SMARTS
    - LESS*HINDERED; IN THE SAME RING AS . . .
    - For translated SMARTS expressions, additional information is stored (JSON)
- Protecting group handling in SAVI-Lib difficult to replicate in SAVI-Space
- Exact comparison of Space with Lib difficult – unless you enumerate Space
  - SAVI-Space had 34% more products than SAVI-Lib
  - Recapitulation of 1000 Lib compounds in Space: 56% - 100%, avg.: 95%

Synthetic accessibility



SA: synthetic accessibility score. RA: retrosynthetic accessibility score.

Searching for Drugbank Molecules in REAL Space and SAVI Space

Queries: Drugbank approved 2025/07; 300 < MW < 500 (1074 cpds)
Search methodology: SpaceLight - topological fingerprint similarity (fCSFP2,5)

- REAL Space 2024:
  - 48 billion products
- Results:
  - 54 exact matches
  - 236 with high similarity (>0.8)
- Resources:
  - 48min (2.7sec/cpd)

- SAVI Space 2024:
  - 7.6 billion products
- Results:
  - 69 exact matches
  - 415 with high similarity (>0.8)
- Resources:
  - 22min (1.3sec/cpd)

Hardware: Intel(R) Core(TM) i5-8500 CPU @ 3.00GHz; 6 cores

SLICE (SMARTS and Logic In ChEmistry)

Idea was: (a) Can we make SAVI faster; (b) SAVI transforms easier to write?

SLICE consists of:

1. SLICE Designer: GUI to define SMARTS patterns, configure atom and bond properties, and establish chemical constraints and logic.
   Programmed in JavaScript; generates XML files with SMARTS and logic.

2. SLICE Engine: uses XML files to generate virtual libraries from specified building blocks.
   Programmed in Java. Generates enumerated SD and/or CSV files.

Available at: https://github.com/tarasovan/SLICE-public
Nouleho Ilemo S., Delannée V. *et al.*, https://doi.org/10.26434/chemrxiv-2025-m6klm

SLICE Designer

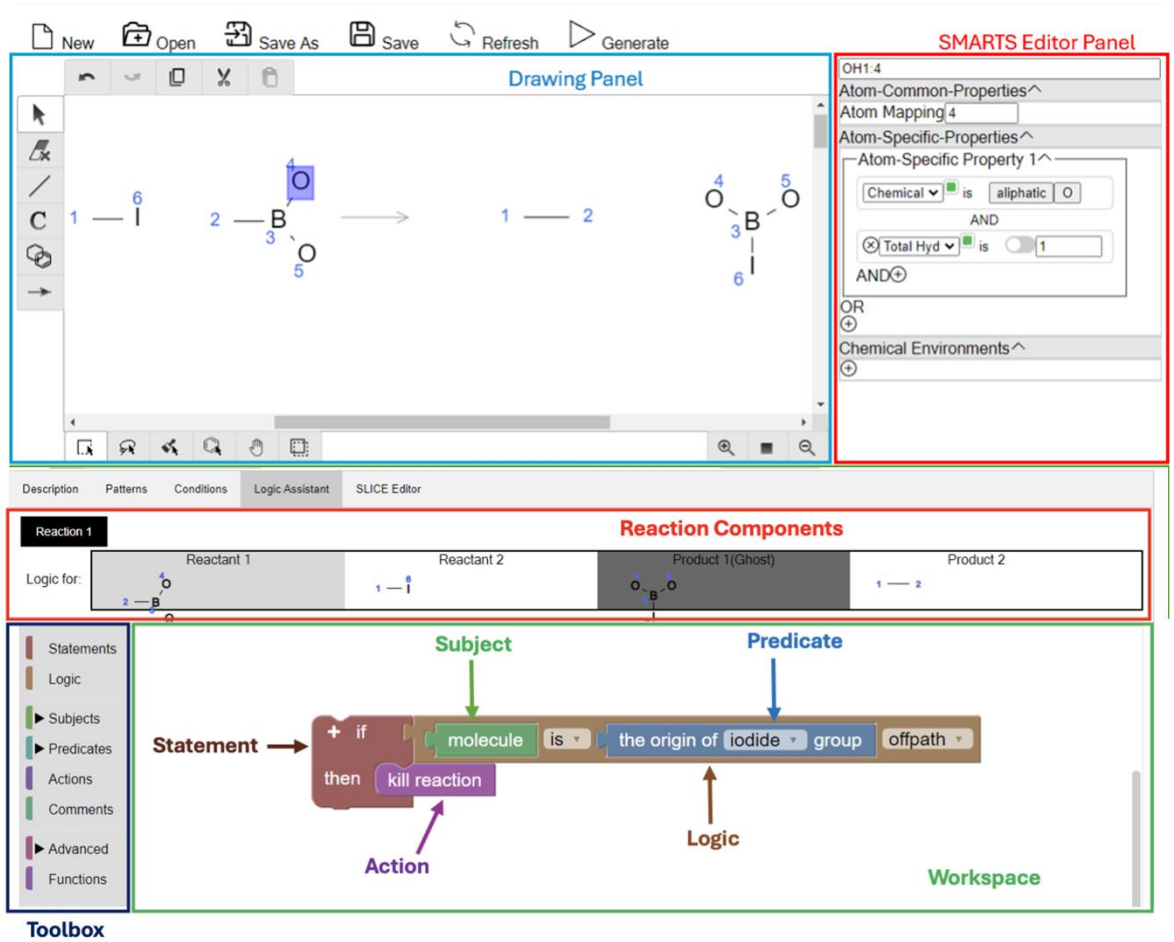
Uses:

- Custom-designed Blockly blocks (no need for traditional programming), based on a simple structure inspired by the home automation software: IFTTT (IF This Then That)

- Conditional statements with the following basic structure:

    if <*subject*> <*relation*> <*predicate*> <*where*> then <*action*>

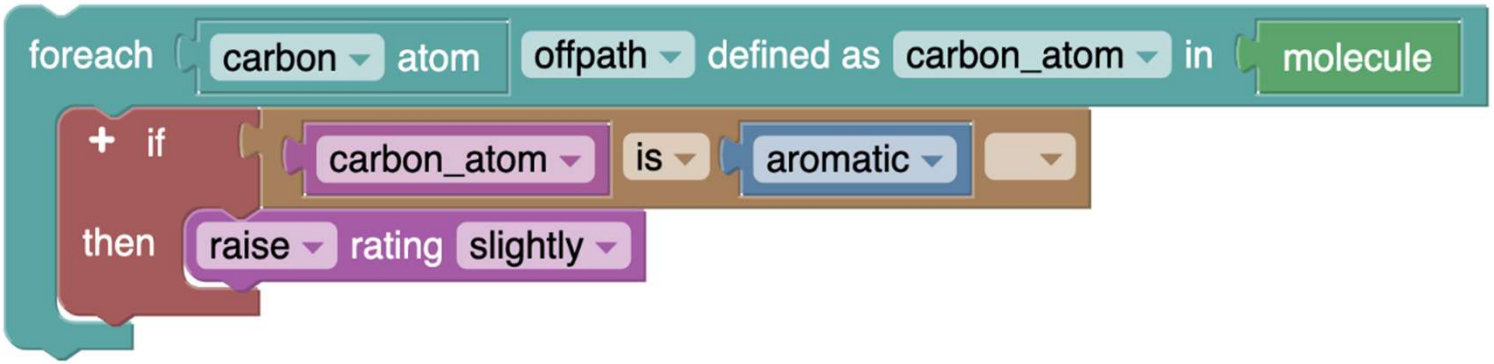
SLICE Designer: Logic Assistant Tab



SLICE Designer with Suzuki-Miyaura Cross-Coupling (Iodo) transform:

SLICE Designer: Loops

Loop syntax: foreach <*chemical_object*> <*where*> defined as <*variable_name*> in <*Object*>



SLICE Engine: Performance

Normalized time per million products for SAVI and SLICE across transforms

Transform ID	SAVI Products	SAVI Time (min)	SAVI Time/Million (min)	SLICE Products	SLICE Time (min)	SLICE Time/Million (min)	Speed-up Ratio SLICE
6005	4,839,627	5,519	1,140.5	3,472,481	92	26.5	43
7019	43,217,313	262,690	6,078.7	129,750,587	14,142	109.0	51
2201	1,190,253	16,038	13,470.4	1,769,041	128	72.3	186
7009	84,849,166	204,859	2,413.4	187,820,353	11,305	60.2	40

Note: These product sets were fully enumerated.

Enumerated vs. Fragment-Based Screening Structure Sets

	Enumerated Libraries	Fragment/Synthon-Based Spaces
Space	Large (TB)	Small (GB)
Maximal structure number	$<10^{11}$	$>10^{11}$
Time to generate	Slow	Fast(1)
Datasets	Static	Dynamic
Specific ring(-system) questions	Yes	No (not yet)
Advanced chemical logic	Yes	Difficult
Exchange of datasets	SDF, SMILES: yes	Synthon sets: not directly
Similarity/substructure search	Slow (GPU-based: faster)	Fast
Docking with external tools	Yes	No
Docking speed	Slow (GPU-based: faster)	Fast (Chemical Space Docking)
Conformational expansion	Yes (though slow, except w/GPUs)	No
External ADME/Tox predictions	Yes	No
Hardware needed	Best: HPC, GPUs	Sufficient: Desktop, laptop

(1) Addition of advanced chemical logic can cost significant human time

Is it Either Enumerated or Fragment-Based?

No, they can happily live alongside each other, if not support each other:

- SAVI Library and SAVI Space
- Enamine REAL Space (77B) now accessible as both an enumerated dataset and a synthon-based space
- FSees (Rarey group): fragment space exhaustive enumeration system
- Iterative modeling based on evolutionary approaches: Enumerate subset of fragment-based space; analyze this subset; use its properties to generate next, evolutionarily improved, subset etc.

Weber L. *et al.*, Angew. Chem. Int. Ed. Engl. (1995), 34(20), 2280-2282
Hiss J. *et al.*, Future Med. Chem. (2014) 6(3), 267–280
Liu H.-P. *et al.*, Sci. Rep. (2024) 14:24510
Reddy S. *et al.*, (2024) arXiv:2407:13779v1

Acknowledgements

- **CADD Group, CCR, NCI**
  - **Hitesh Patel** (now: OpenEye)
  - **Yuri Pevzner** (now: BMS)
  - **Stefi Nouleho**
  - **Victorien Delannée** (now: Deep Origin)
  - Olga Grushin
  - Megan L. Peach (now: FDA)
- **SBDD Group, CCR, NCI**
  - Nadya Tarasova
  - Cody Hoop
- **Enamine/Chemspace**
  - Yurii Moroz
- **Xemistry GmbH**
  - Wolf-Dietrich Ihlenfeldt
- **Lhasa Limited**
  - Philip Judson
  - Martin Ott
- **OntoChem**
  - Lutz Weber (now: Actyon Discovery)
- **ChemNavigator/MilliporeSigma**
  - Scott Hutton (now: Actyon Discovery)
- **MilliporeSigma**
  - Bret Daniel
  - Chad Hurwitz
- **Novartis**
  - Peter Ertl (now: Actyon Discovery [Advisor])
- **BioSolveIT**
  - Christian Lemmen
  - Raphael Klein
- **Univ. Hamburg**
  - Matthias Rarey
  - **Malte Korn**