Searching the REAL data base with genetic algorithms

A talk in 2 parts.

Synthon GA

Jan H. Jensen
University of Copenhagen

GABBY

Noel O'Boyle
Sosei Heptares, UK

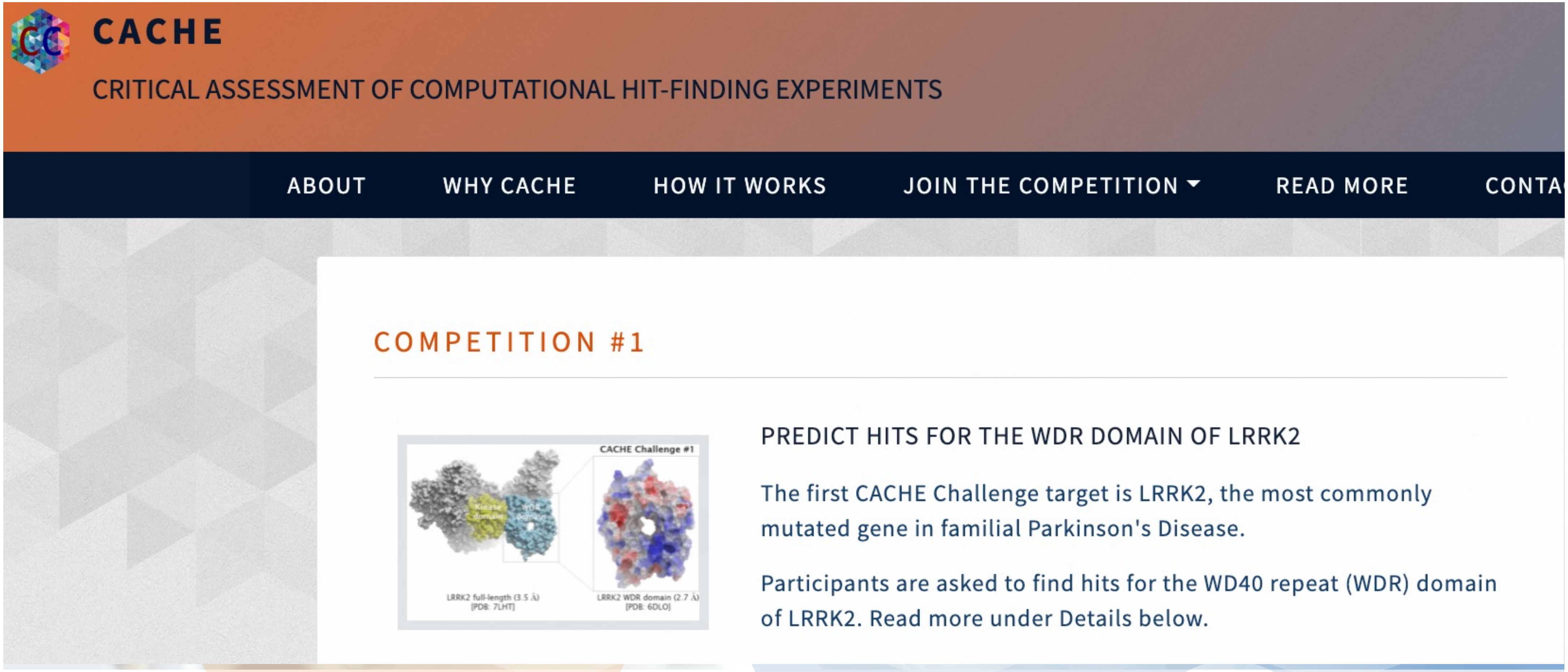
Synthon-GA

Searching make-on-demand libraries with genetic algorithms

Jan H. Jensen
Department of Chemistry,
University of Copenhagen

@janhjensen

Casper Steinmann
(Aalborg University)



CACHE

CRITICAL ASSESSMENT OF COMPUTATIONAL HIT-FINDING EXPERIMENTS

ABOUT WHY CACHE HOW IT WORKS JOIN THE COMPETITION READ MORE CONTA

COMPETITION #1

PREDICT HITS FOR THE WDR DOMAIN OF LRRK2

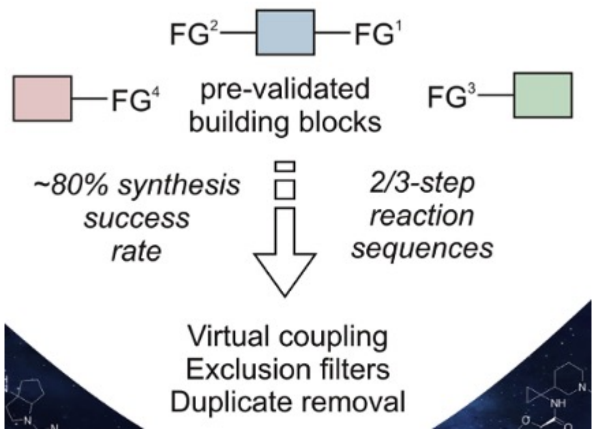
The first CACHE Challenge target is LRRK2, the most commonly mutated gene in familial Parkinson's Disease.

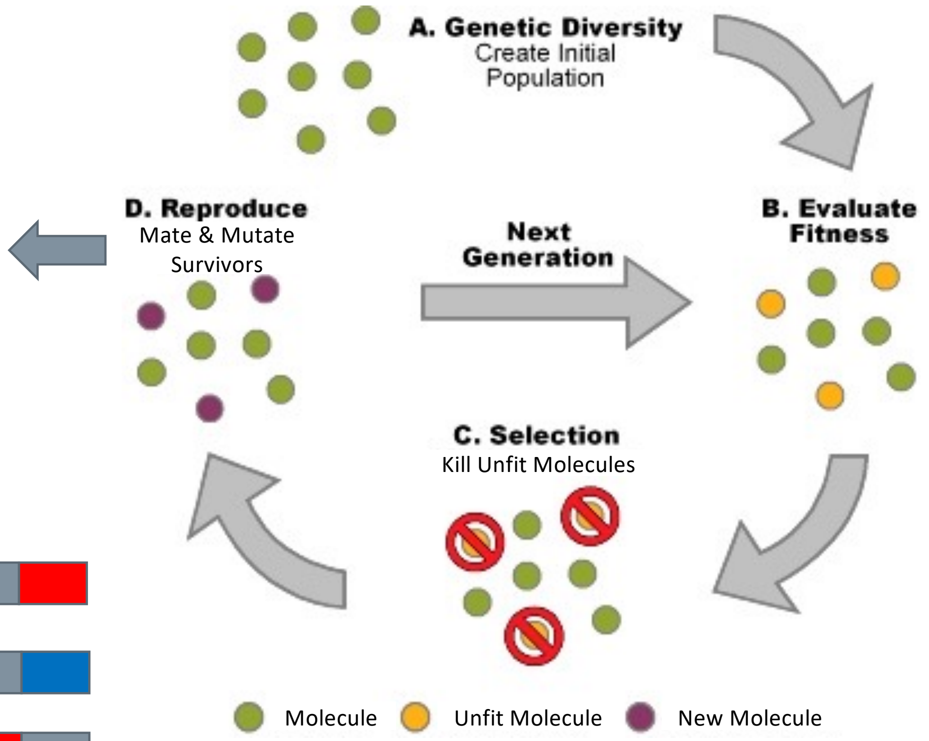
Participants are asked to find hits for the WD40 repeat (WDR) domain of LRRK2. Read more under Details below.

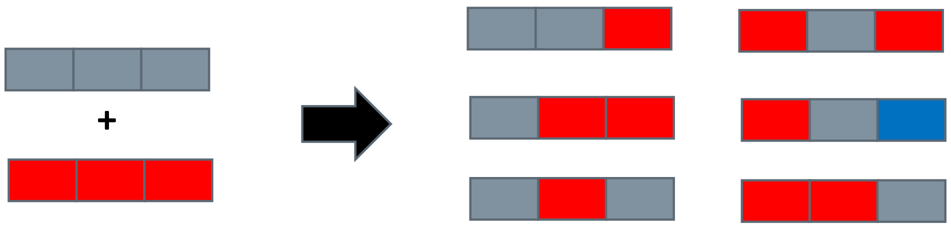
Organisers will purchase and test $10K worth of molecules from Enamine

Genetic Algorithms for make-on-demand libraries

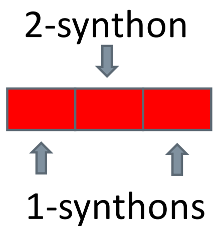
Mating


FG2
FG1
FG4
pre-validated building blocks
FG3
~80% synthesis success rate
2/3-step reaction sequences
Virtual coupling
Exclusion filters
Duplicate removal





REAL Space is only a small fraction of possible genes



2-synthon

1-synthons

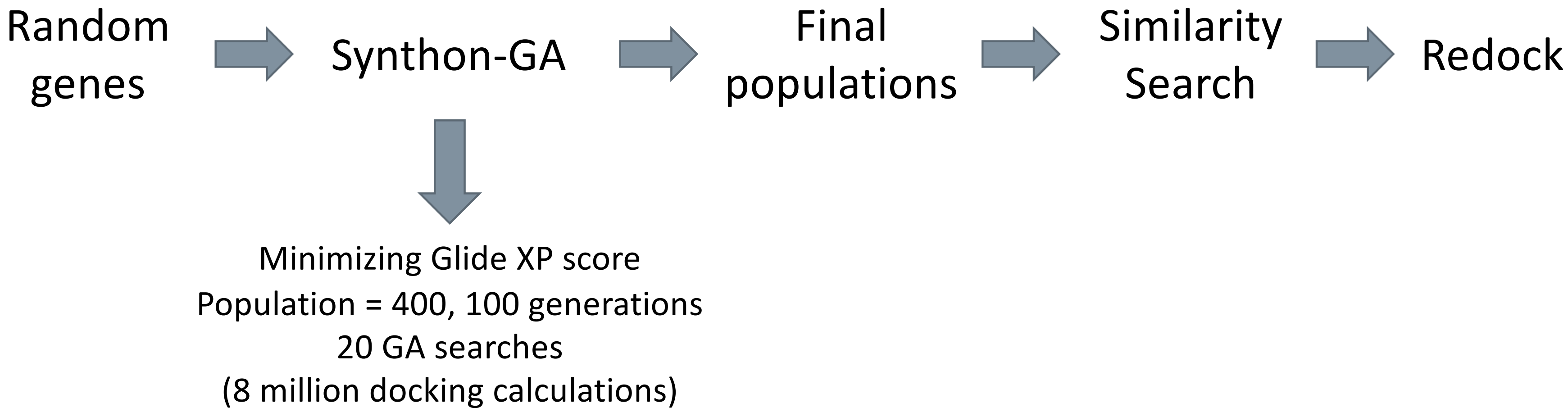
- Building Blocks used in the *REAL* database, 128.9K cpds, SDF

129K reagents
=> 91K and 41K 1-synthons and 2-synthons
+ 24 possible reactions
= 28 trillion genes

“The *REAL* Space comprises 21 billion make-on-demand molecules and is currently the largest offer of commercially available compounds.

The *REAL* compounds in the Space are assembled via more than 170 well-validated parallel synthesis protocols applied to over 112 000 qualified reagents and building blocks.”

Workflow:



Random genes → Synthon-GA → Final populations → Similarity Search → Redock

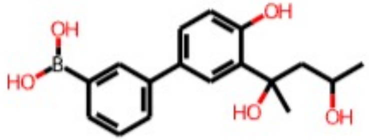
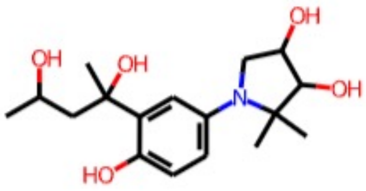
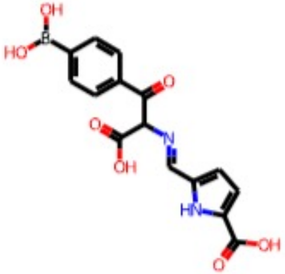
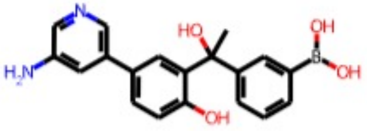
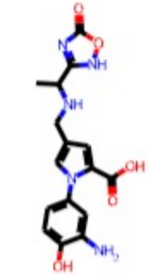
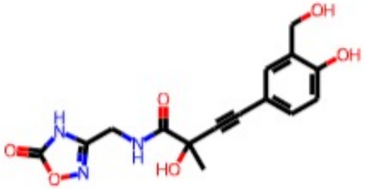
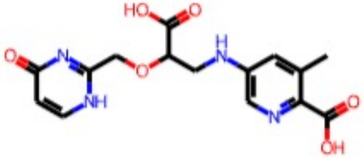
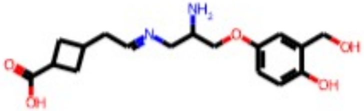
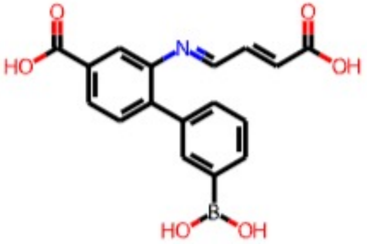
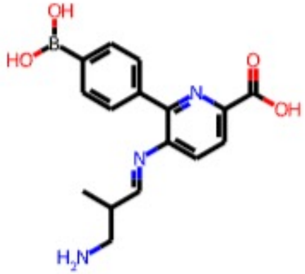
Synthon-GA ↓

Minimizing Glide XP score
Population = 400, 100 generations
20 GA searches
(8 million docking calculations)

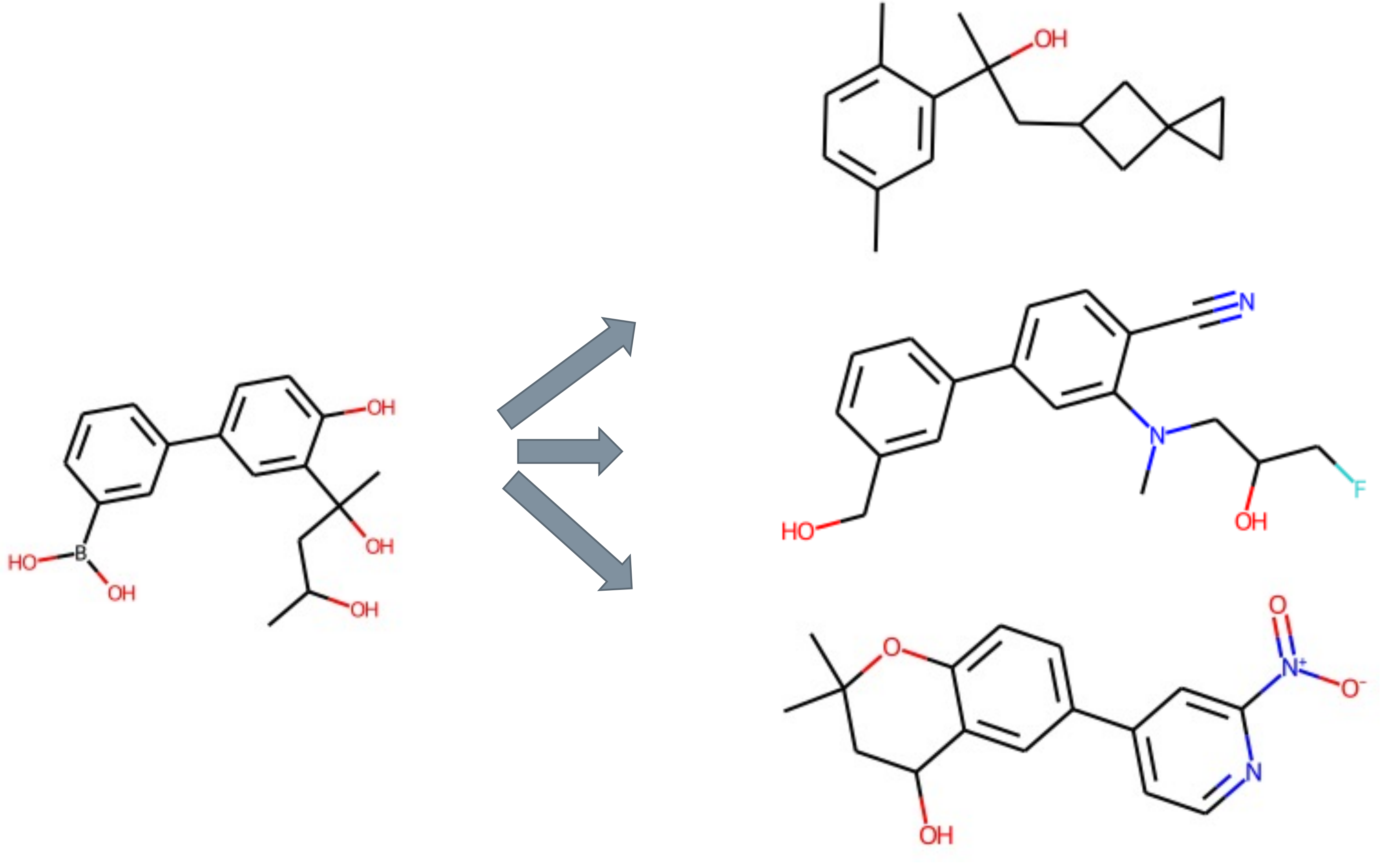
GA-2

MW < 350, logP < 3.5

Final pop 8000 → 90 unique molecules

□  -12.8	□  -12.5	□  -12.0	□  -12.0	□  -12.0
□  -11.8	□  -11.7	□  -11.7	□  -11.7	□  -11.6

Random Example

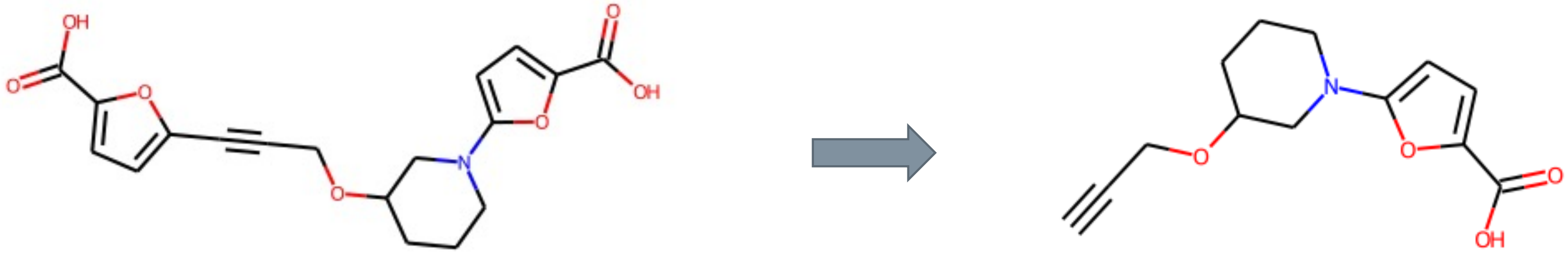


Postera

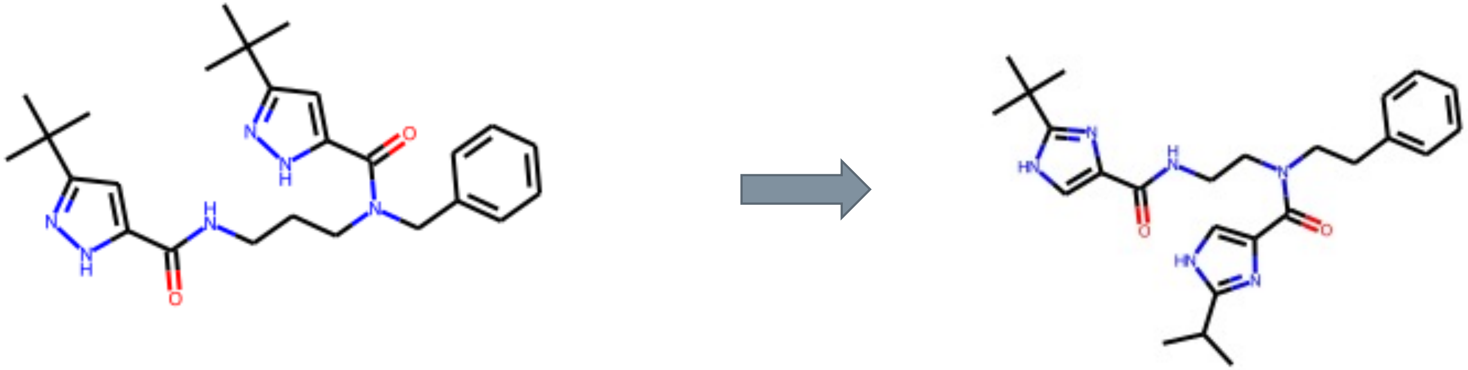
FTrees

SmallWorld

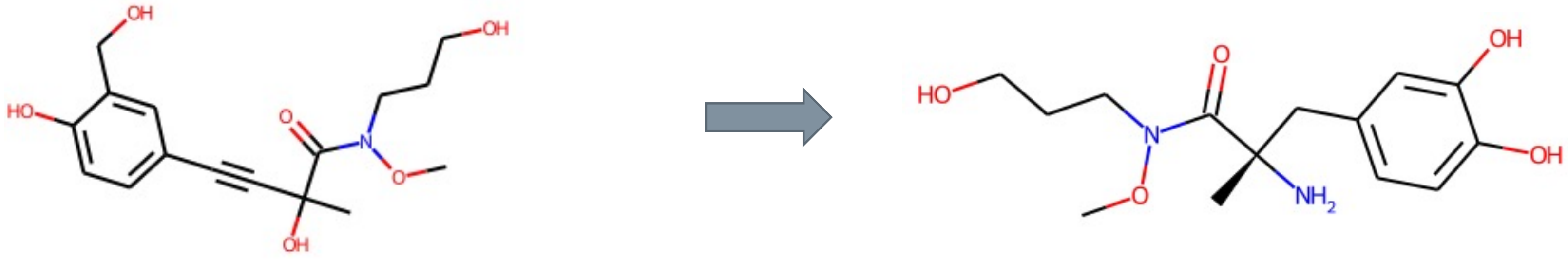
Best Case



Postera



FTrees



SmallWorld

Ultra-large Chemical Libraries

10 August 2022 10:00-17:00, London, United Kingdom

Chris Swain
Cambridge MedChem Consulting, United Kingdom
https://www.cambridgemedchemconsulting.com

Ilaria Proietti Silvestri
Liverpool Chirochem, United Kingdom
https://www.liverpoolchirochem.com

Over to you Noel!

NON-CONFIDENTIAL

Gabby – A genetic algorithm of building blocks

Noel M O’Boyle, Jan H Jensen

Disclaimer

The material that follows is a presentation of general background information about Sosei Group Corporation and its subsidiaries (collectively, the “Company”) as of the date of this presentation. This material has been prepared solely for informational purposes and is not to be construed as a solicitation or an offer to buy or sell any securities and should not be treated as giving investment advice to recipients. It is not targeted to the specific investment objectives, financial situation or particular needs of any recipient. It is not intended to provide the basis for any third party evaluation of any securities or any offering of them and should not be considered as a recommendation that any recipient should subscribe for or purchase any securities.

The information contained herein is in summary form and does not purport to be complete. Certain information has been obtained from public sources. No representation or warranty, either express or implied, by the Company is made as to the accuracy, fairness, or completeness of the information presented herein and no reliance should be placed on the accuracy, fairness, or completeness of such information. The Company takes no responsibility or liability to update the contents of this presentation in the light of new information and/or future events. In addition, the Company may alter, modify or otherwise change in any manner the contents of this presentation, in its own discretion without the obligation to notify any person of such revision or changes.

This presentation contains “forward-looking statements,” as that term is defined in Section 27A of the U.S. Securities Act of 1933, as amended, and Section 21E of the U.S. Securities Exchange Act of 1934, as amended. The words “believe”, “expect”, “anticipate”, “intend”, “plan”, “seeks”, “estimates”, “will” and “may” and similar expressions identify forward looking statements. All statements other than statements of historical facts included in this presentation, including, without limitation, those regarding our financial position, business strategy, plans and objectives of management for future operations (including development plans and objectives relating to our products), are forward looking statements. Such forward looking statements involve known and unknown risks, uncertainties and other factors which may cause our actual results, performance or achievements to be materially different from any future results, performance or achievements expressed or implied by such forward looking statements. Such forward looking statements are based on numerous assumptions regarding our present and future business strategies and the environment in which we will operate in the future. The important factors that could cause our actual results, performance or achievements to differ materially from those in the forward looking statements include, among others, risks associated with product discovery and development, uncertainties related to the outcome of clinical trials, slower than expected rates of patient recruitment, unforeseen safety issues resulting from the administration of our products in patients, uncertainties related to product manufacturing, the lack of market acceptance of our products, our inability to manage growth, the competitive environment in relation to our business area and markets, our inability to attract and retain suitably qualified personnel, the unenforceability or lack of protection of our patents and proprietary rights, our relationships with affiliated entities, changes and developments in technology which may render our products obsolete, and other factors. These factors include, without limitation, those discussed in our public reports filed with the Tokyo Stock Exchange and the Financial Services Agency of Japan. Although the Company believes that the expectations and assumptions reflected in the forward-looking statements are reasonably based on information currently available to the Company's management, certain forward looking statements are based upon assumptions of future events which may not prove to be accurate. The forward looking statements in this document speak only as at the date of this presentation and the company does not assume any obligations to update or revise any of these forward statements, even if new information becomes available in the future.

This presentation does not constitute an offer, or invitation, or solicitation of an offer, to subscribe for or purchase any securities. Neither this presentation nor anything contained herein shall form the basis of any contract or commitment whatsoever. Recipients of this presentation are not to construe the contents of this summary as legal, tax or investment advice and recipients should consult their own advisors in this regard.

This presentation and its contents are proprietary confidential information and may not be reproduced, published or otherwise disseminated in whole or in part without the Company’s prior written consent. These materials are not intended for distribution to, or use by, any person or entity in any jurisdiction or country where such distribution or use would be contrary to local law or regulation.

This presentation contains non-GAAP financial measures. The non-GAAP financial measures contained in this presentation are not measures of financial performance calculated in accordance with IFRS and should not be considered as replacements or alternatives profit, or operating profit, as an indicator of operating performance or as replacements or alternatives to cash flow provided by operating activities or as a measure of liquidity (in each case, as determined in accordance with IFRS). Non-GAAP financial measures should be viewed in addition to, and not as a substitute for, analysis of the Company's results reported in accordance with IFRS.

References to "FY" in this presentation for periods prior to 1 January 2018 are to the 12-month periods commencing in each case on April 1 of the year indicated and ending on March 31 of the following year, and the 9 month period from April 1 2017 to December 31 2017. From January 1 2018 the Company changed its fiscal year to the 12-month period commencing in each case on January 1. References to "FY" in this presentation should be construed accordingly.

What we will do about Enamine REAL?

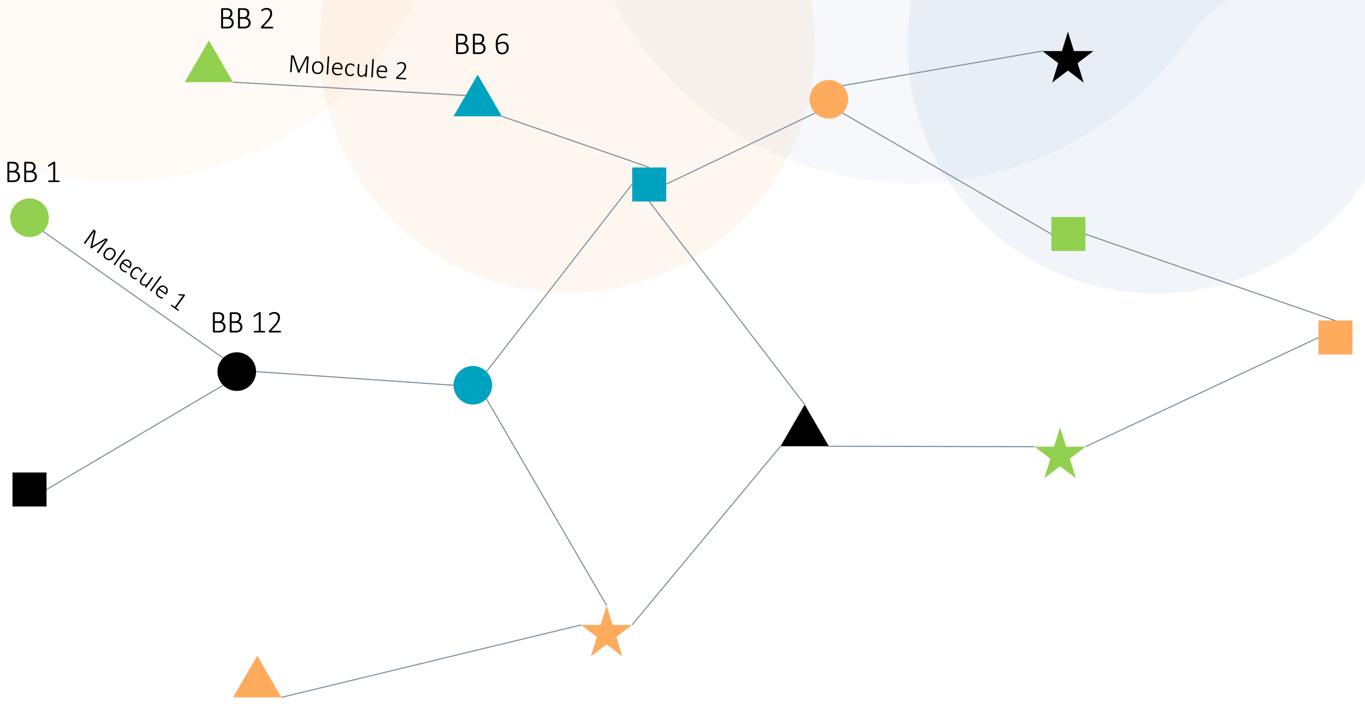
- Enamine REAL:
  - 5.5B compounds available for €45-70 each and delivered in 3-4 weeks (with >80% success rate)
- Sosei Heptares is a structure-based drug discovery (SBDD) biotech focussing on GPCRs
  - In Comp Chem, we are interested in using protein-ligand docking to find hits
  - But…how do we perform a virtual screen on a database of 5.5B?
- Brute force?
  - Assuming 20s per docking, and a 1024-core machine, this would take 3.4 years
  - Maybe with 100 such machines? Indeed, just 12.4 days
  - Amazon! 5600 x 16 CPU machines is $1.2M/mo so maybe $480K for 12 days
- And then Enamine REAL becomes 10x bigger….?
- Is it possible to instead screen a fraction of the database with a high probability of finding those with good scores?
  - A genetic algorithm is a stochastic search algorithm that is used to find near-optimal solutions without exploring the entire search space
  - Gabby – Genetic algorithm (GA) of building blocks (BB)

Let’s get REAL, and take a look

smiles	idnumber	reagent1	reagent2	reagent3	reagent4	reaction	MW	HAC	sLogP	HBA	HBD	RotBonds	FSP3	TPSA	...
C[C@H](NC(=O)N1CC2(CCC2)C1C1=CC=C(F)C=C1)C1CC1	Z2699586657	3278147	24271632			512	302.393	22	3.861	1	1	3	0.611	32.34	...
CC(NC(=O)N1CC2(CCC2)C1C1=CC=C(F)C=C1)C1CC1	Z2697338668	221666	24271632			512	302.393	22	3.861	1	1	3	0.611	32.34	...
C[C@@H](NC(=O)N1CC2(CCC2)C1C1=CC=C(F)C=C1)C1CC1	Z2699585259	3274379	24271632			512	302.393	22	3.861	1	1	3	0.611	32.34	...
CC(C1=CC=CC=C1)C(O)C(=O)N(C)C1CCN2CCOC1C2	Z2909270631	15329544	2128492			22	318.417	23	1.083	4	1	4	0.611	53.01	...
CCOCCCCNC(=O)N1CCC2=CC=CC(CO)=C2C1	Z4407745067	150381	26617790			2708	306.406	22	2.063	3	2	7	0.588	61.8	...
NCC1=C(CS(=O)(=O)NCC2=CC=CC(C(F)(F)F)=C2)N=CO1	Z4188234679	10467	26587962			270084	349.334	23	1.772	5	2	6	0.308	98.22	...
CC1=C(OCC(=O)NC(C)C2=CN(C)N=N2)C(Cl)=CC(Cl)=C1	Z4642401991	15686252	4138576			22	343.214	22	2.687	5	1	5	0.357	69.04	...
CC(O)C1=NC=C(C(=O)NCC2COCC3(CCNCC3)O2)S1	Z4422454943	24850355	11251268			240690	341.433	23	0.464	7	3	4	0.733	92.71	...
NC(=O)C1=C(N)N(CCOC(=O)C2=CC=C(Cl)C(F)=C2)N=C1	Z1552910396	3098381	74284			276436	326.715	22	1.214	6	2	5	0.154	113.23	...

Consider all two-component molecules as forming a graph

Building blocks (BBs) as nodes, molecules as edges



- Number of nodes = 125K, number of edges = 1.0B
- Gabby will walk through this graph searching for a global optimum

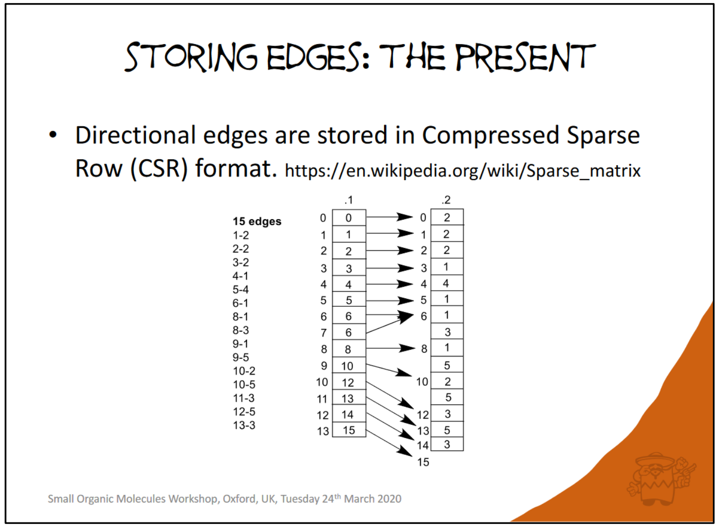
sosei HEPTARES

Why is a graph representation useful?

- By restricting ourselves to the edges in the graph, we stay entirely within Enamine REAL
  - These are the ‘allowed’ combinations of BBs
- Neighbouring edges have an increased likelihood to have similar structures and similar docking scores

```
from gabby.graph import Graph
from gabby.molecule import Molecule
g = Graph(index1_fname, index2_fname, enamine_real_gz)
bb = 3
nbrs = g.get_nbrs(bb)
mol = Molecule(bb, nbrs[0])
smiles = next(g.get_smiles(mol))
```

- Managing this large graph can be done efficiently if we ~~steal~~ borrow approaches used by others for storing and querying ultra-large graph databases
  - SmallWorld from NextMove Software
    - 10.0T edges, 790B nodes
    - https://www.nextmovesoftware.com/talks/Sayle_SmallWorld_Oxford_202003.pdf



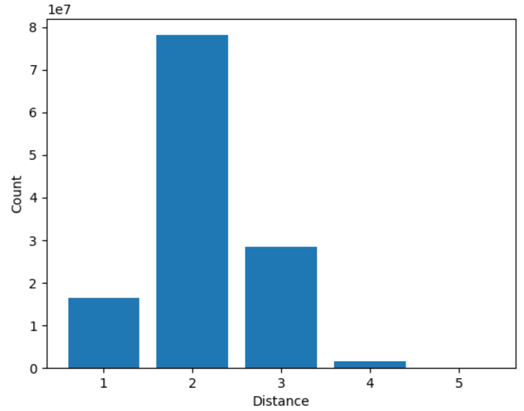
STORING EDGES: THE PRESENT

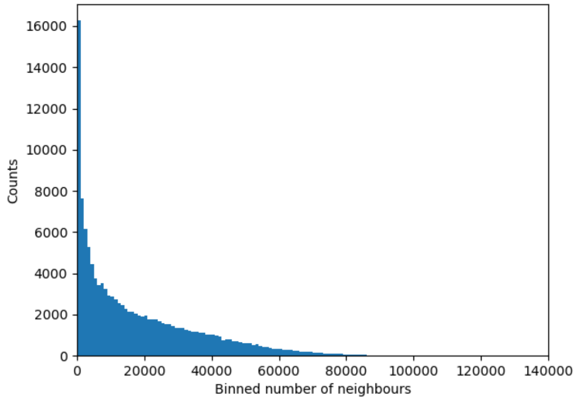
- Directional edges are stored in Compressed Sparse Row (CSR) format. https://en.wikipedia.org/wiki/Sparse_matrix

Small Organic Molecules Workshop, Oxford, UK, Tuesday 24th March 2020

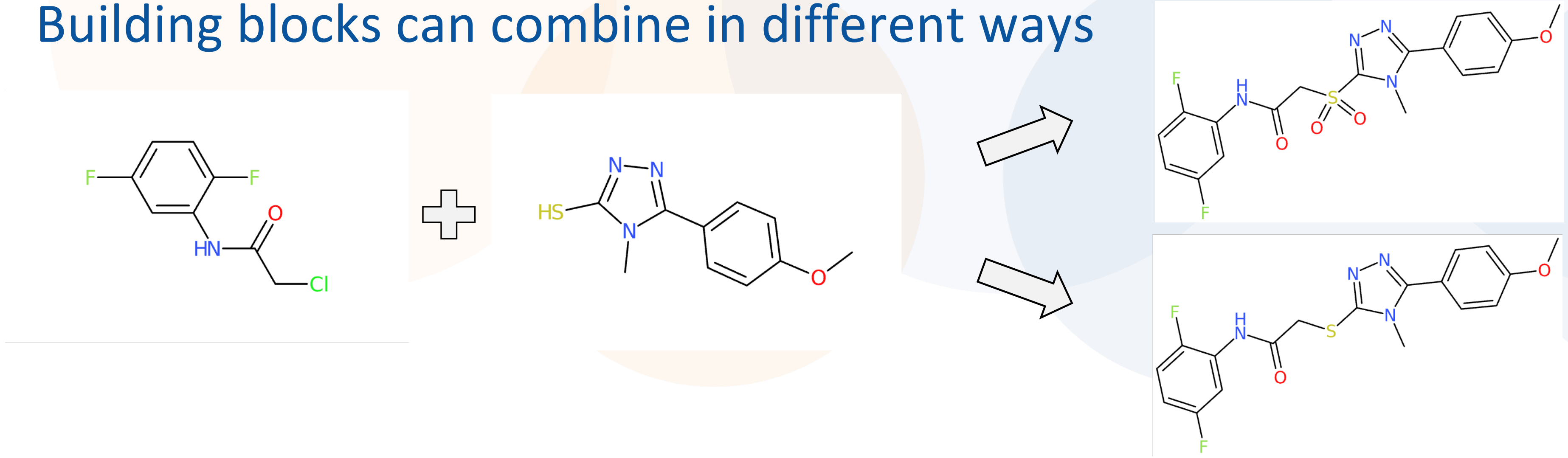
What are the characteristics of the Enamine REAL graph?

- Num of nodes = 124509, num of edges = 1.0B
- It forms a single connected component
  - This simplifies things, as it means we don't have to consider the problem of disconnected graphs
- It is highly connected
  - On average, the shortest distance between any two BBs is on average 2 and at maximum 5 (based on a sample of 1000 BBs)
  - Getting from point A to point B within N generations will not be a problem
- While the median number of neighbours is 11995, it is a typical long-tailed distribution
  - One BB has 134958 neighbours
  - 16266 BBs have only a single neighbour (are only part of a single molecule)

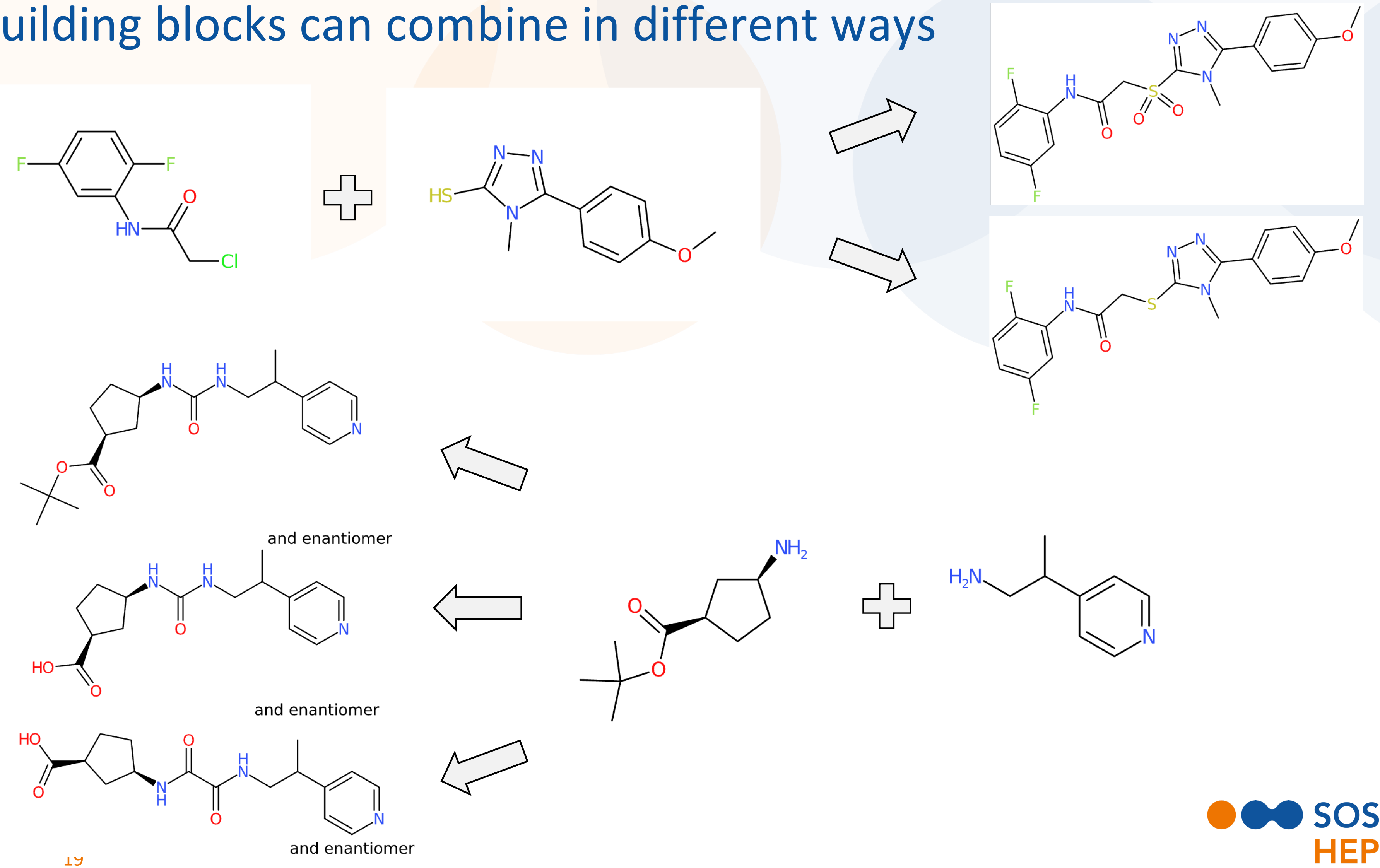




Building blocks can combine in different ways

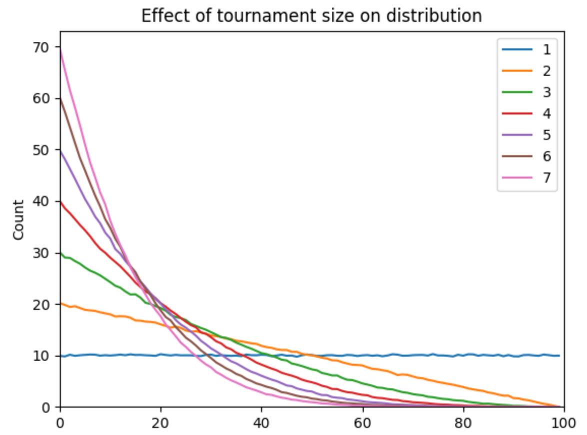


Building blocks can combine in different ways



Gabby

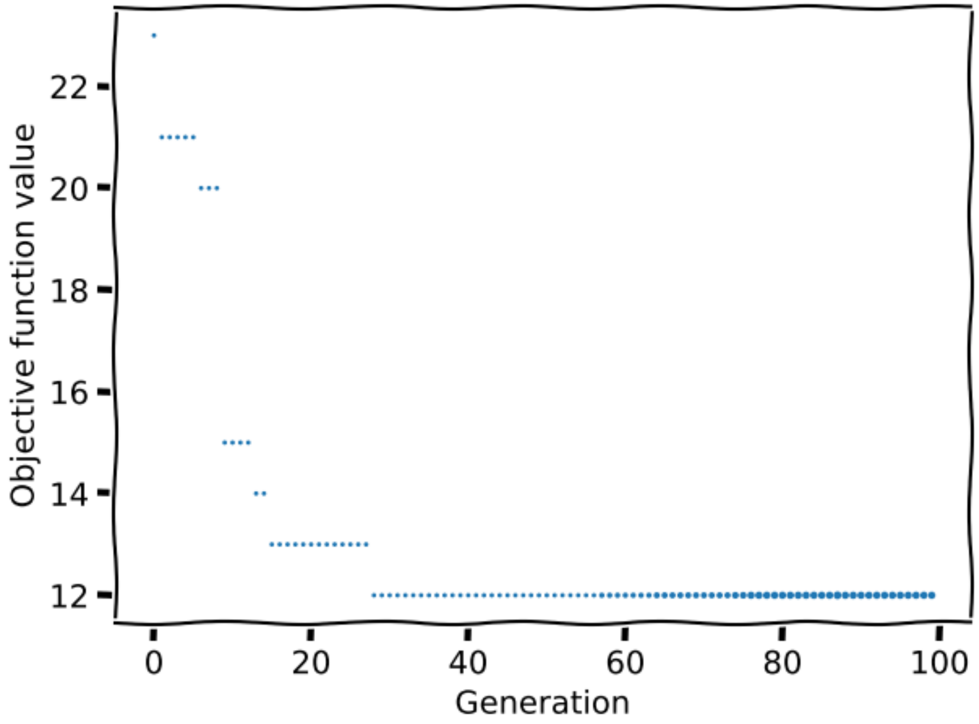
- Create initial population of size N:
  - Randomly select N edges of the graph (i.e. molecules)
- Generate N children
  - Select N/5 parents via tournament selection
  - Repeatedly select 2 of these for cross-over and mutation
- Score children
- Create new generation
  - Combine N/2 of the best scoring parents with N/2 of the best-scoring children

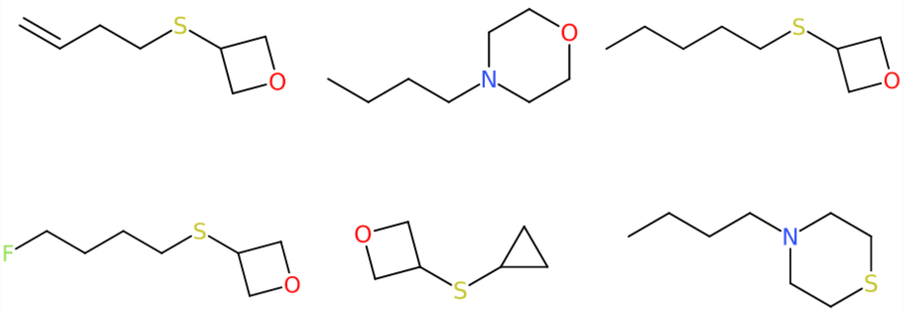


https://baoilleach.blogspot.com/2022/08/tournament-sizes-and-their-effect-on.html

- Important to balance exploitation of existing knowledge (cross-over) with exploration (mutation)
- The details of the balance are still being worked out!!

Example result with toy objective function





Alternatives to this approach

- A. Build an ML model to reproduce docking scores (e.g. Active Learning Glide [1], Hasten [2], Deep Docking [3], Lean-Docking [4])
- B. Dock the building blocks (BB), select those that have high scores and are oriented suitably, then dock the molecules having those building blocks (V-SYNTHES [5], Chemical Space Docking [6])

- Comparison to Gabby:
  - Gabby and the ML model (A) dock the full molecule from the start, compared to the V-SYNTHES approach (B)
    - May or may not make much difference
  - Gabby can use any scoring function, e.g. a MPO that includes the docking score but also additional desired properties
    - Could be used to guide the search towards preferred poses
  - Gabby as described is focussed on two-component molecules – an extension to additional components is planned
  - Could use a ML model to guide Gabby, rather than the simplistic iteration loop described in the original papers
  - The graph module of Gabby could be of use to implement other algorithms, e.g. B above
- Would be interesting to compare the results of all three

[1] Yang et al. *J. Chem. Theory Comput.* **2021**, *17*, 7106
[2] Kalliokoski. *Mol. Inf.* **2021**, *40*, 2100089
[3] Gentile et al. *ACS Cent. Sci.* **2020**, *6*, 939
[4] Berenger et al. *J. Chem. Inf. Model.* **2021**, *61*, 2341
[5] Sadybekov et al. *Nature* **2022**, *601*, 452
[6] https://www.biosolveit.de/application-academy/chemical-space-docking/. Paper to appear soon.

Acknowledgements

- Chris de Graaf

Availability

- Will be available on GitHub as soon as possible

Any questions? Or if we’ve run out of time…

- noel.oboyle@soseiheptares.com, jhjensen@chem.ku.dk