



Handling large chemical spaces in Structure-Based Drug Design

Noel M. O'Boyle

Ninth Joint Sheffield Conference on Chemoinformatics

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.

© Sosei Group Corporation. Sosei Heptares is the corporate brand and trade mark of Sosei Group Corporation. Sosei, Heptares, the logo and StaR® are trade marks of Sosei Group companies.

Outline

1

Introduction

2

Generative Design of A2A ligands

3

Mol2Synthon: Make synthons from Enamine REAL

4

Gabby: Genetic Algorithm of building blocks

5

SynthonConnect: Join synthon poses

6

Generative Design of Enamine REAL molecules



1

Introduction

Sosei Heptares GPCR Structure-Based Drug Discovery Company

Delivered 25+ pre-clinical candidates, produced 10+ clinical candidates
>6 new pre-clinical candidates expected in the next 2 years for internal and collaboration programs

R&D CENTRE
CAMBRIDGE, UK (Heptares)

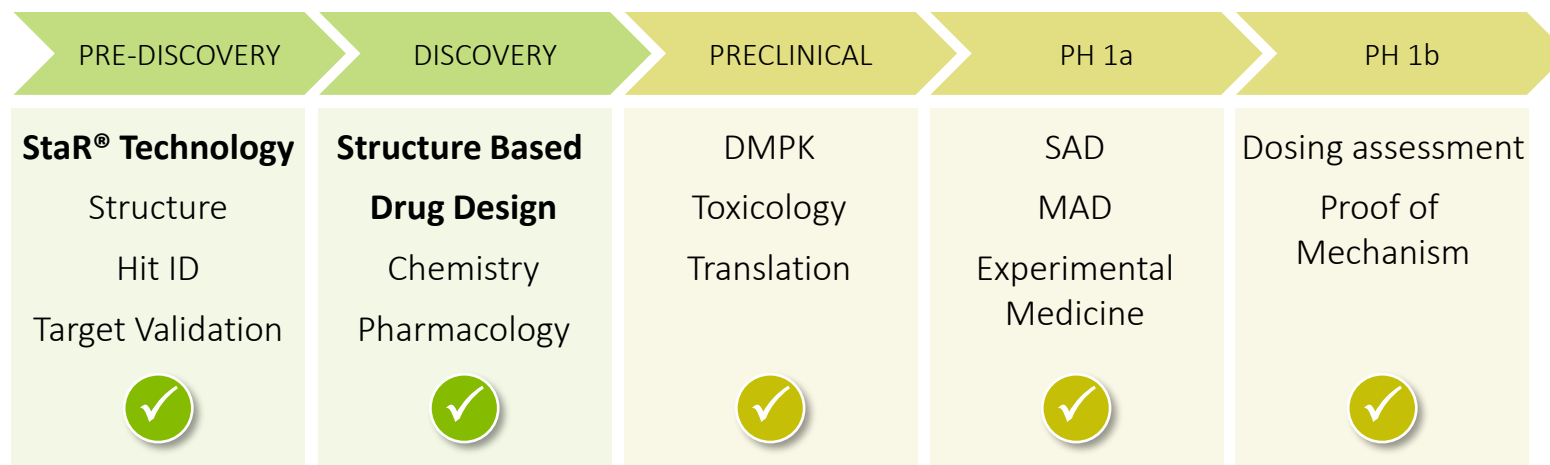
~170 EMPLOYEES



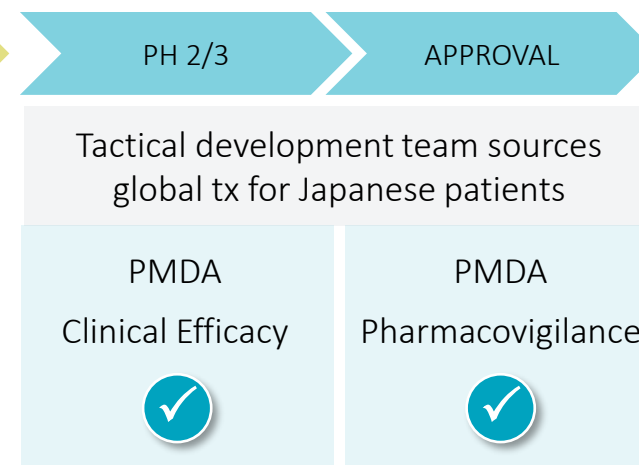
HEADQUARTERS
TOKYO, JAPAN (Sosei K.K.)

~25 EMPLOYEES

DRUG DISCOVERY/EARLY DEVELOPMENT DRIVEN BY STAR® / SBDD ENGINE



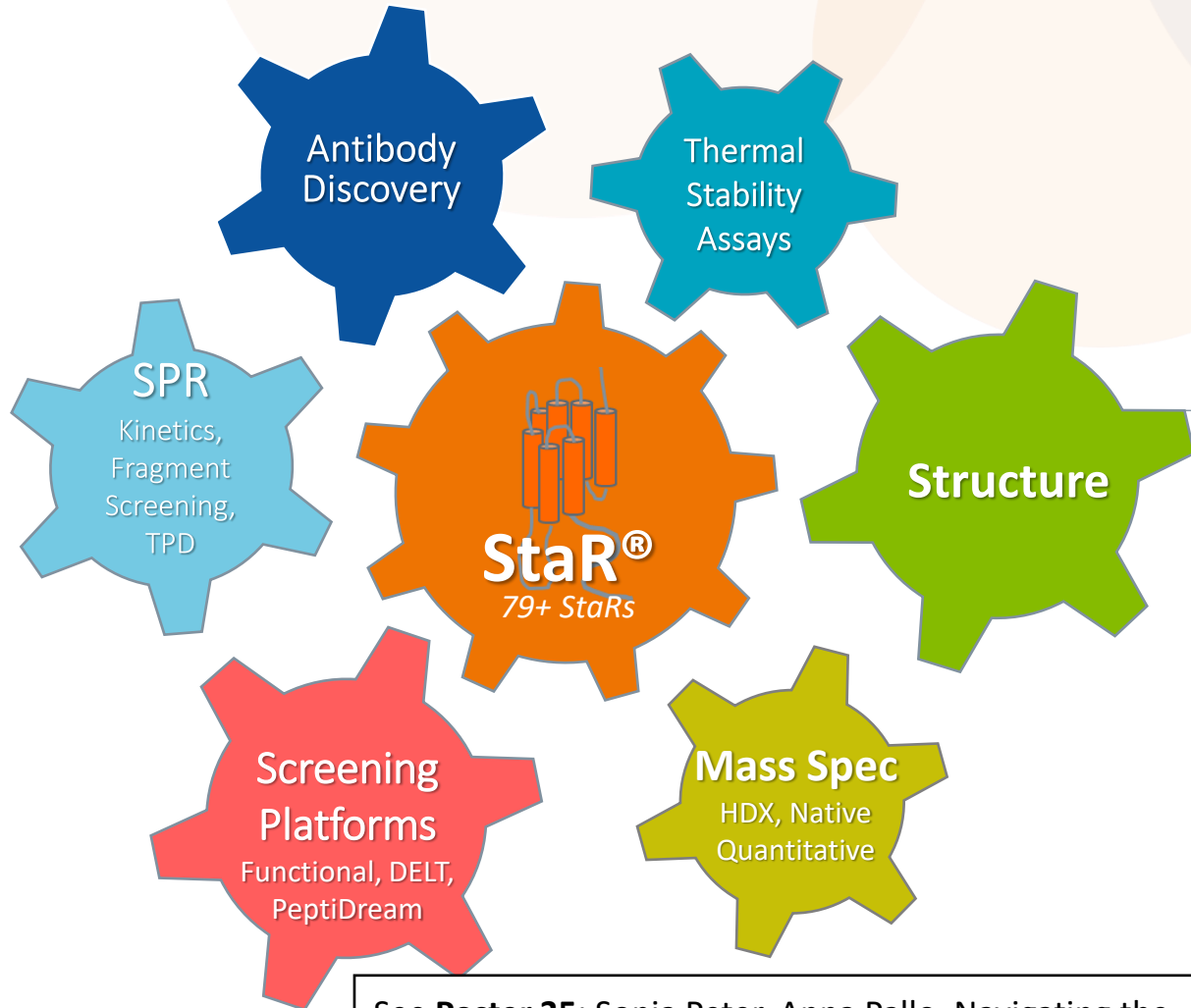
LATE STAGE DEVELOPMENT



Programs advanced to PoM or PoC before partnering / seeded into co-owned investment vehicles

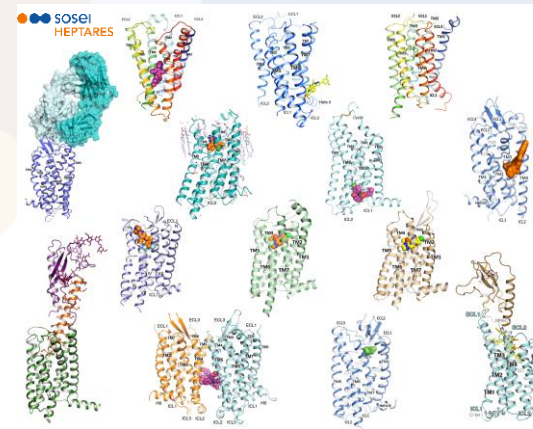
StaR® Technology enabling GPCR Structure-Based Drug Discovery

X-ray crystallography platform complemented by Cryo-EM structural enablement to expand scope of GPCR SBDD

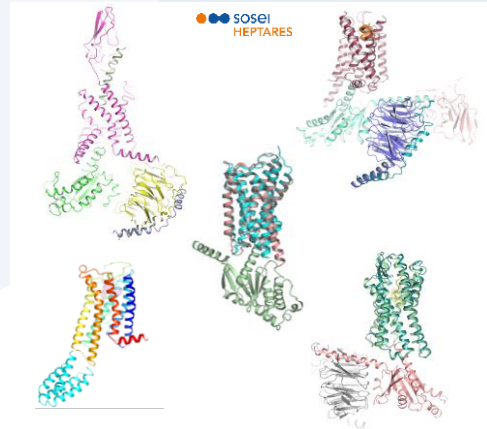


See **Poster 35**: Sonja Peter, Anna Pallo. Navigating the Orthosteric and Allosteric Structural GPCR Pocketome for Structure-Based Drug Discovery

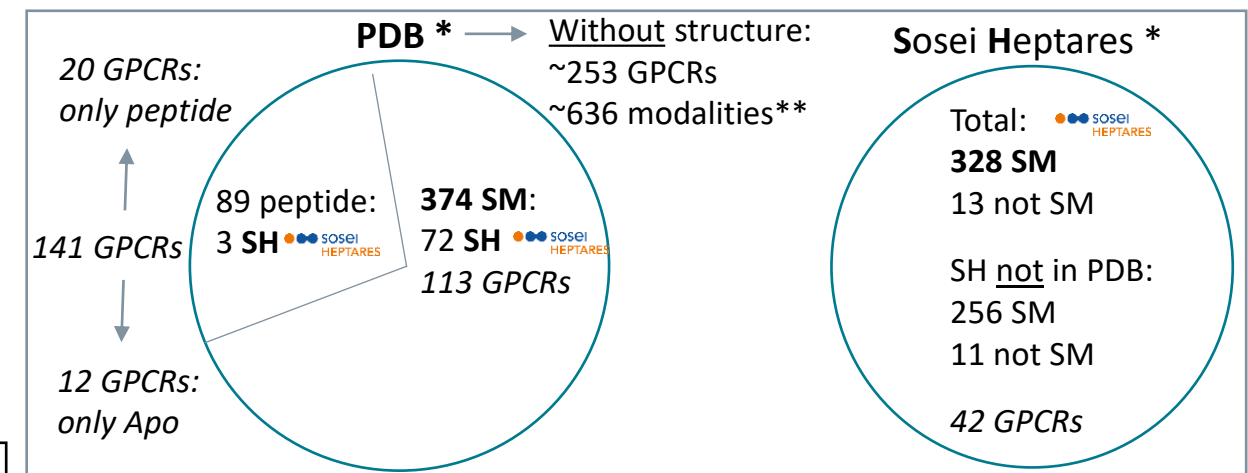
Examples of SH X-ray structures



Examples of SH cryoEM structures



StaRs double small molecule ligand bound GPCR structures for SH SBDD



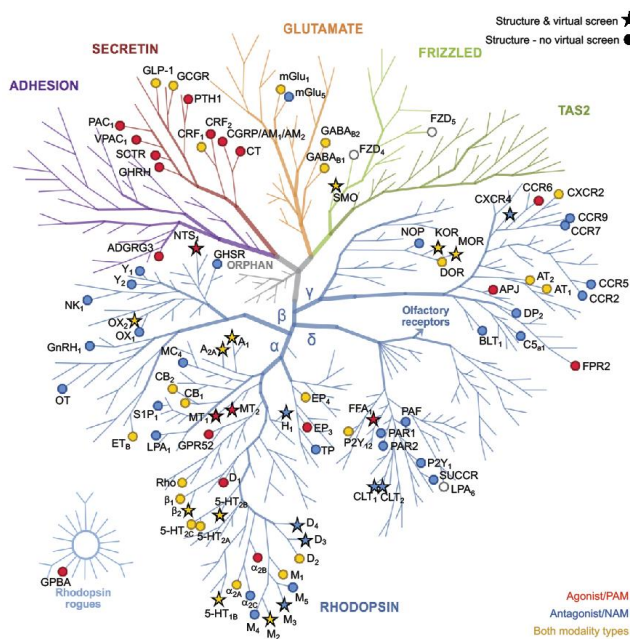
* September 2022, unique GPCR-ligand complexes

** Assuming 403 non-olfactory GPCRs x 2 modalities (agonist/PAM, antagonist/NAM)

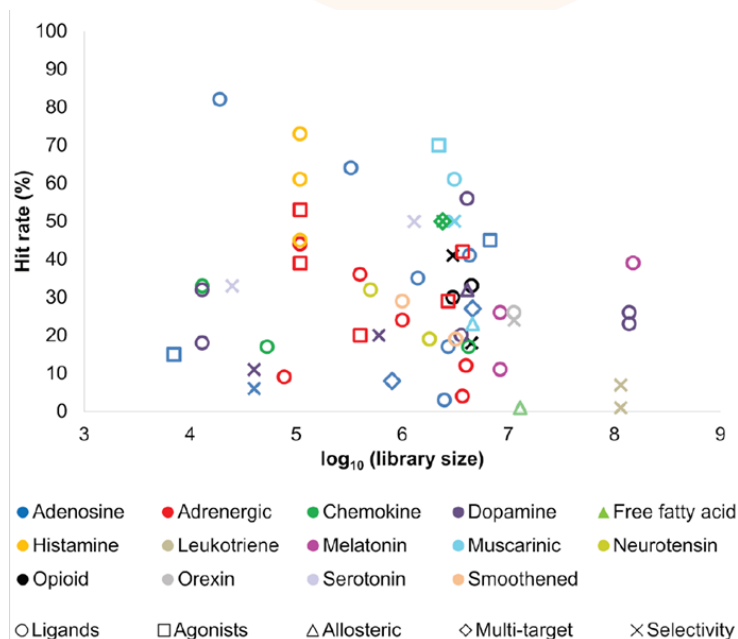
GPCR Structure-Based Virtual Screening – What can Docking do for you?

- 2021: 55 X-ray structure docking-based virtual screening studies for 22 GPCRs¹ – Increasing opportunities for X-ray/cryo-EM GPCR SBVS
- Hit rates of experimentally validated ligands >20% for e.g. aminergic, adenosine receptors – peptide GPCRs can be more challenging¹
- Recently, increasing vendor/virtual library sizes² are leading to challenges in throughput
- New/orthogonal approaches required to efficiently sample chemical space compatible with diverse GPCR binding sites

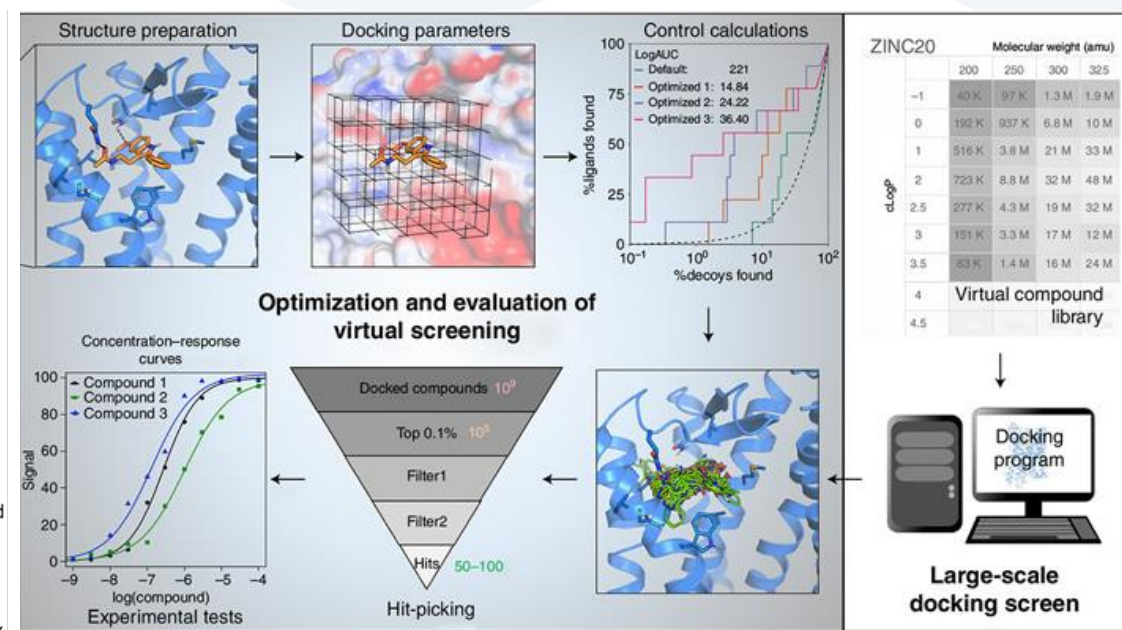
GPCR crystal structure docking-based VS studies¹



% Experimentally validated docking VS hits¹



Docking workflow – ultra-large libraries²



¹ Ballante, Kooistra, Kampen, de Graaf, Carlsson (2021) 73, 527

² B Bender et al. A practical guide to large-scale docking. *Nat. Protocols*. **2021**, 16, 4799

The background features a series of overlapping circles in shades of blue, orange, and light grey. A dark blue wavy line runs horizontally across the middle of the slide.

2

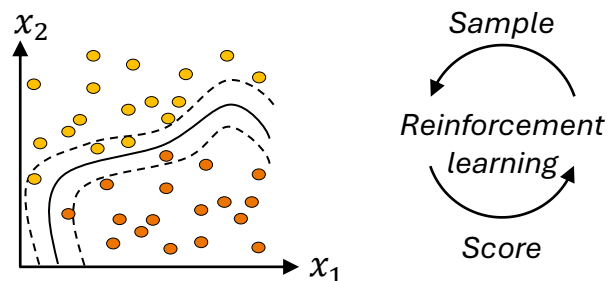
Generative Design of A2A ligands

Structure-based optimization improves de novo molecule generation

- Thomas, Smith, O'Boyle, de Graaf, Bender (2021) *J Cheminform* 13, 39
- Thomas, Bender, de Graaf (2023) *Curr Opin Struc Biol* 79, 102559

Ligand-based approach (SVM-Agent)

Machine learning model trained on D₂ ligands



MOSES [1]

~ 3M SMILES from ZINC

RNN

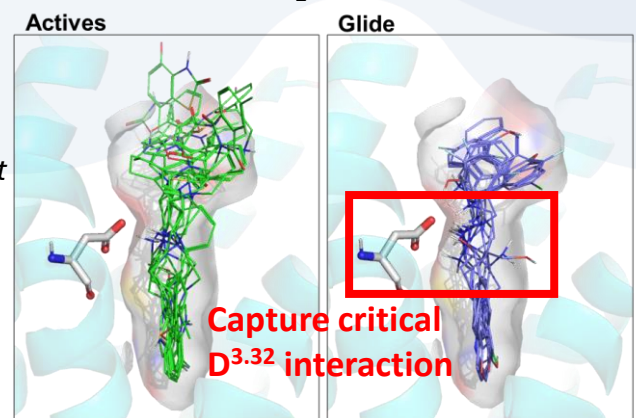
REINVENT [2]

RNN

RNN

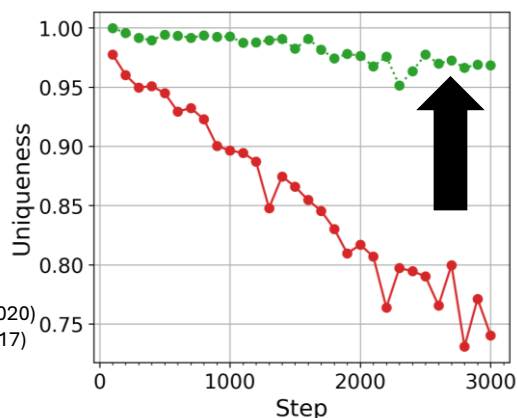
Structure-based approach (Glide-Agent)

Docking into D₂ crystal structure

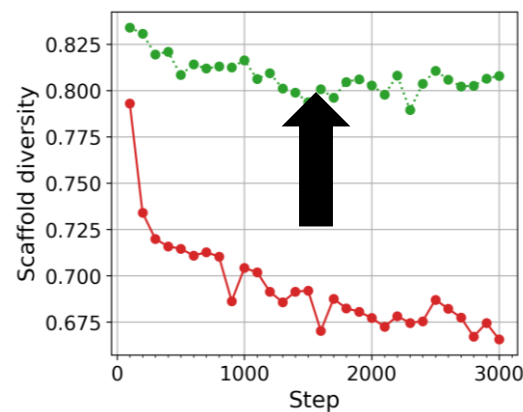


Glide-Agent has ...

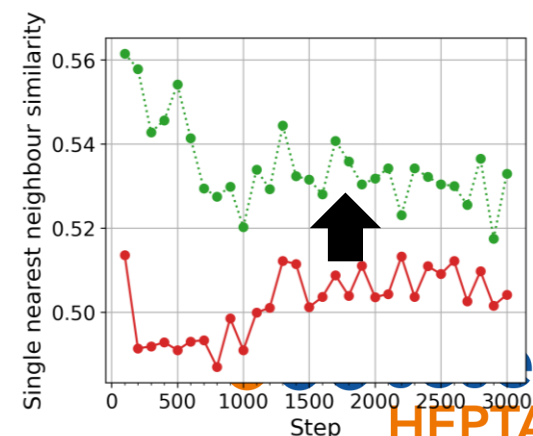
... more **unique** molecules



... higher scaffold **diversity**



... higher **similarity** to the training set



[1] PolyKovskiy, D., et al. *Front Pharmacol* 11 (2020)
 [2] Olivecrona, M., et al. *J Cheminform* 9, 48 (2017)

Structure-based optimization improves de novo molecule generation

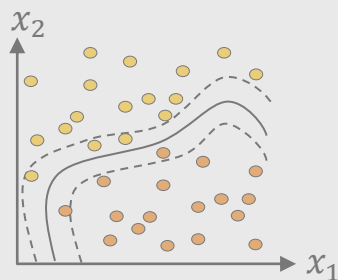
MOSES [1]

~ 3M SMILES from ZINC

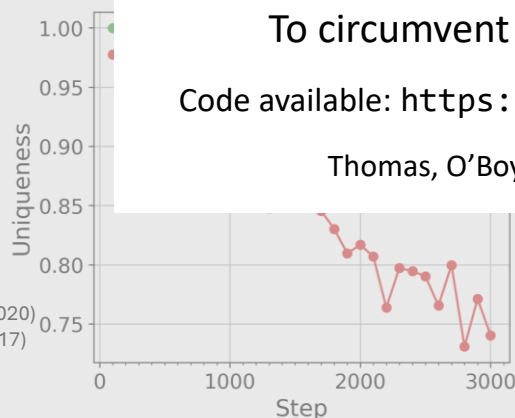
- Thomas, Smith, O'Boyle, de Graaf, Bender (2021) *J Cheminform* 13, 39
- Thomas, Bender, de Graaf (2023) *Curr Opin Struc Biol* 79, 102559

Ligand-based approach (SVM)

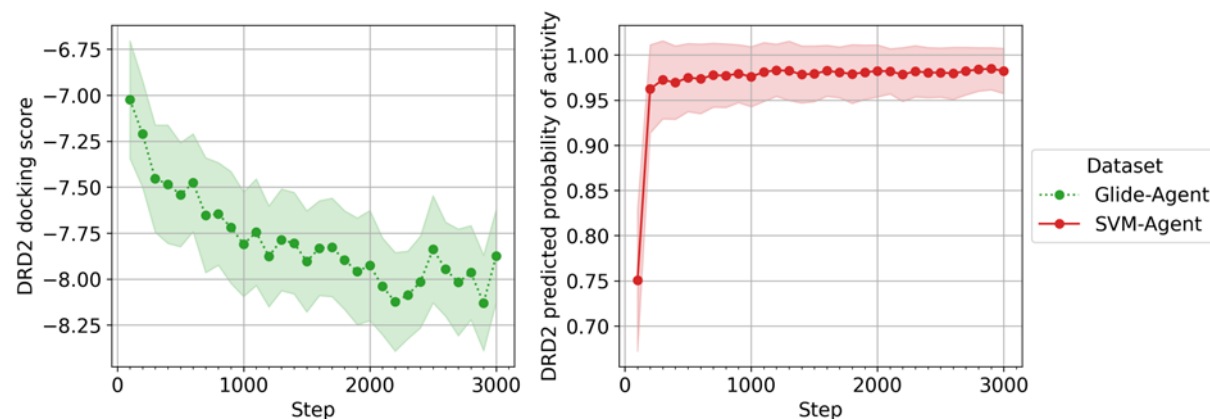
Machine learning model trained on 1



Glide-Agent has mo



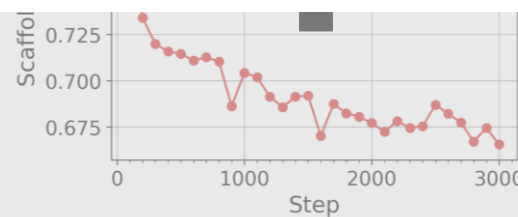
However, structure-based reinforcement learning suffers from sampling efficiency



To circumvent speed losses, use an **Augmented Hill Climb**

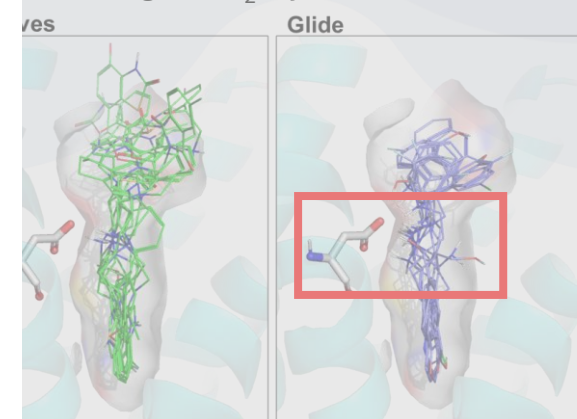
Code available: <https://github.com/MorganCThomas/SMILES-RNN>

Thomas, O'Boyle, Bender, de Graaf (2022) *J Cheminformatics* 14: 68

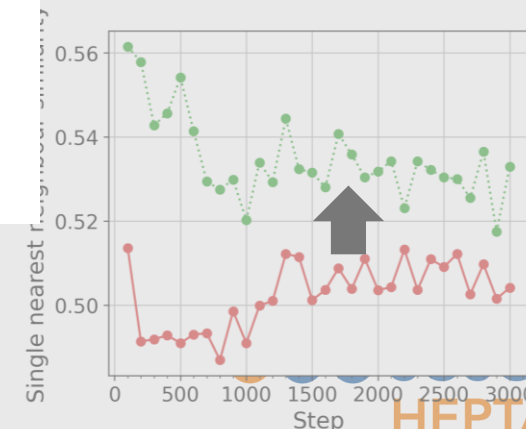


Structure-based approach (Glide-Agent)

Docking into D₂ crystal structure



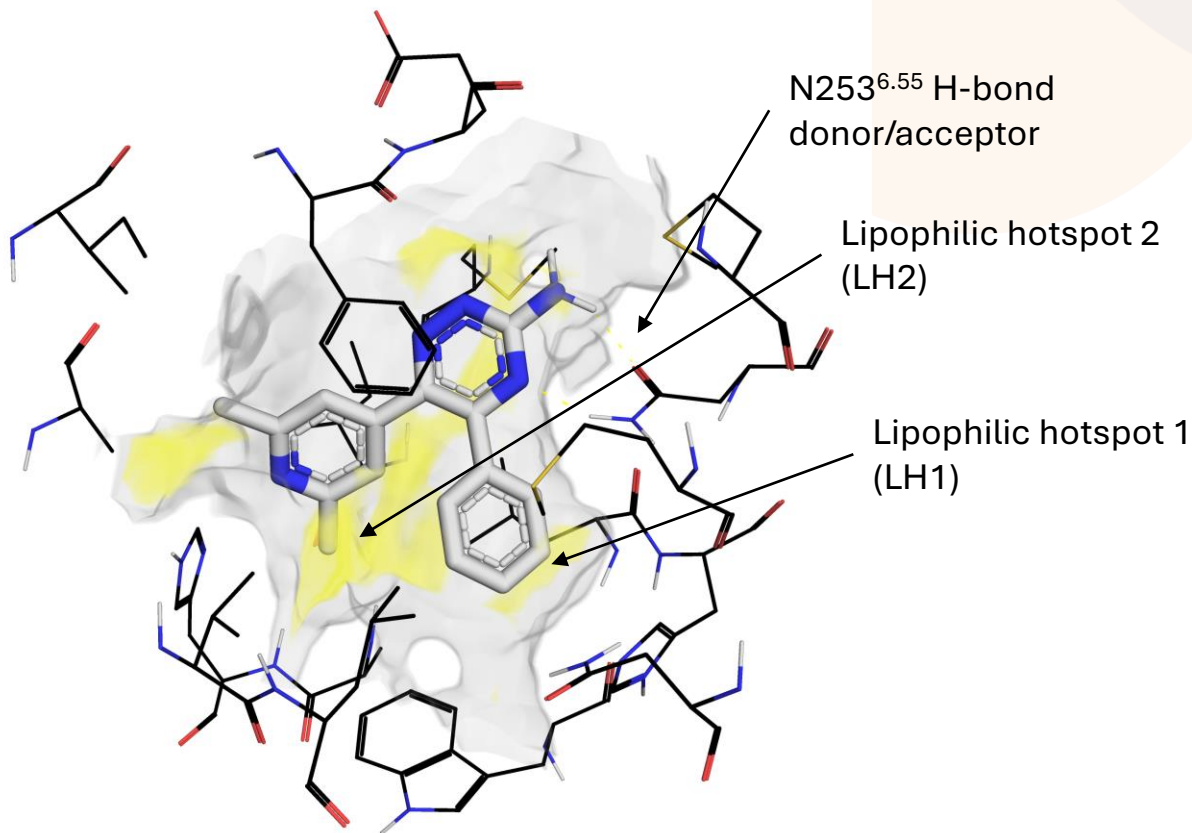
similarity to the training set



[1] PolyKovskiy, D., et al. *Front Pharmacol* 11 (2020)
[2] Olivecrona, M., et al. *J Cheminform* 9, 48 (2017)

Generative Structure-Based A_{2A} Adenosine Receptor Antagonist Design*

A_{2a} StaR[®] receptor in complex with antagonist [1]



A_{2a} is a well-liganded G protein-coupled receptor with many known chemotypes and crystal structures



Curated set of 79 known A_{2a} chemotypes [2]

Multi objective optimization

Structure-based:

Glide-SP docking

N253^{6.55} H-bond acceptor/donor

[optional]

Lipophilic hotspots LH1 & LH2 [optional]

Property-based:

1 ≤ CLogP ≤ 3

1 ≤ CRotBonds ≤ 3

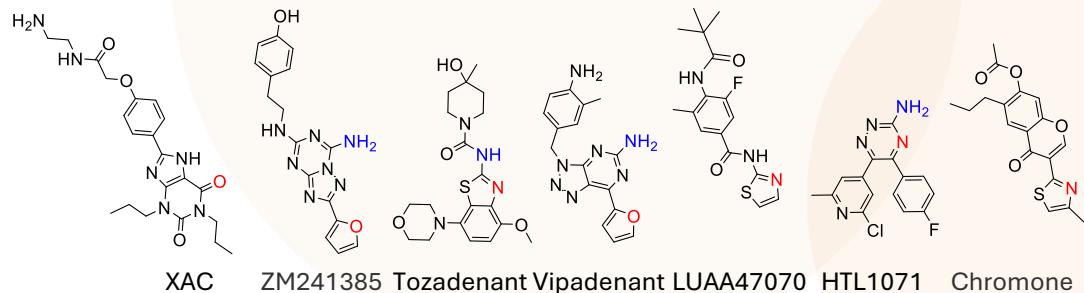
HBond donors ≤ 3

Synthesizability-based:

RAScore [3]

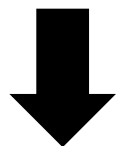
- [1] Congreve, M., *et al.* J Med Chem 55, 1898-1903 (2012)
- [2] Weiss, D.R., *et al.* J Chem Inf Model 54, 642-651 (2016)
- [3] Thakkar, A., *et al.* Chem Sci 12, 3339-3349 (2021)

A_{2A} antagonist co-crystal structure influences chemotypes generated*

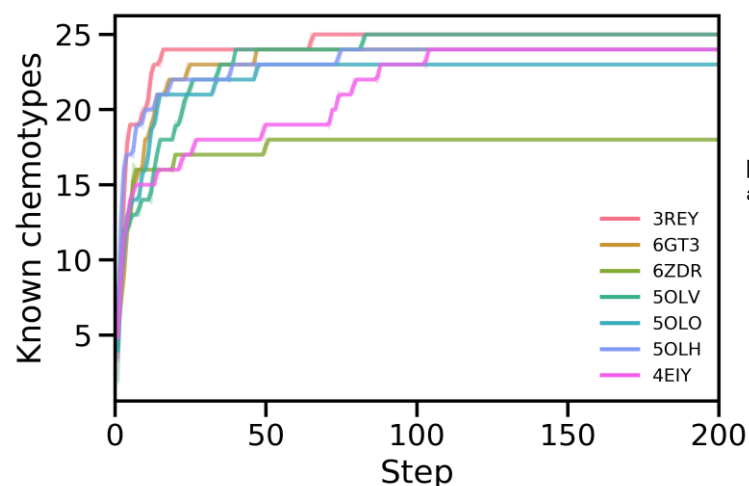


3REY ^[1] StaR®	4IEY ^[2]	5OLO ^[3] StaR®	5OLH ^[3] StaR®	5OLV ^[3] StaR®	6GT3 ^[4] StaR®	6ZDR ^[5] StaR®
------------------------------	---------------------	------------------------------	------------------------------	------------------------------	------------------------------	------------------------------

Conduct de novo molecule
generation optimizing the
docking score with different
xstal structures^[1-5]

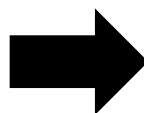


- [1] Dore, *et al.* Structure 19, 1283 (2011)
- [2] Liu *et al.* Science 337, 232 (2012)
- [3] Rucktooa *et al.* Sci Rep 8, 41 (2018)
- [4] Borodovsky *et al.* J Immunother Cancer 8, e000417 (2020)
- [5] Jespers *et al.* Angew Chem Int Ed Engl 59, 16536 (2020)



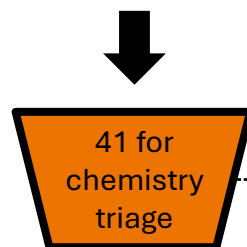
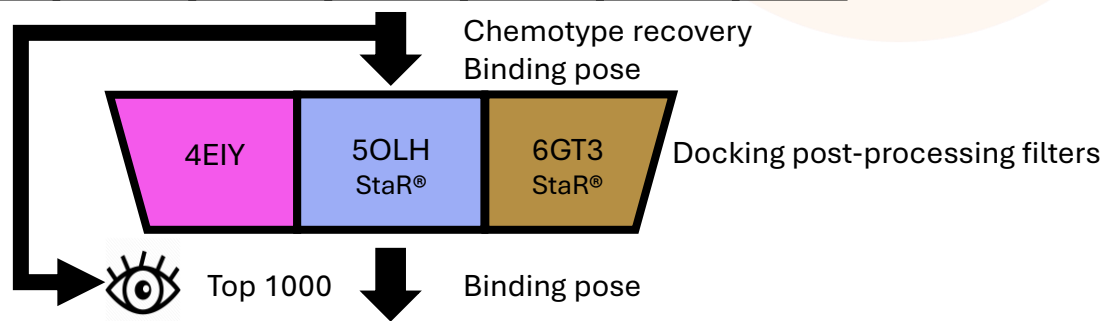
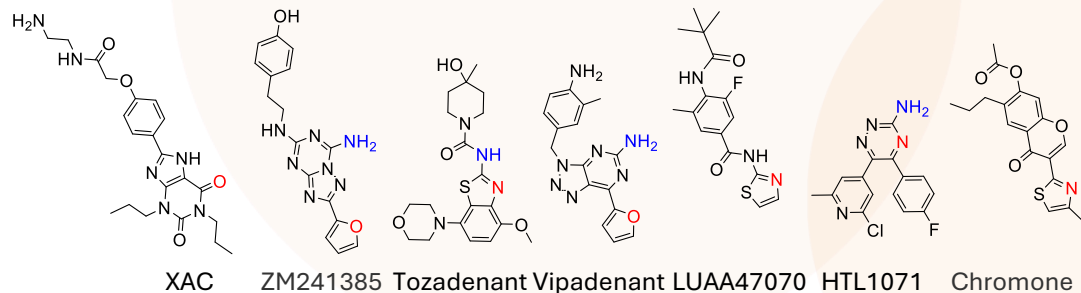
30 known chemotypes^[6]
recovered in total – some
more than others

[6] Weiss, *et al.* J Chem Inf Model 54, 642 (2016)
appended with A_{2A} chemotypes since 2016

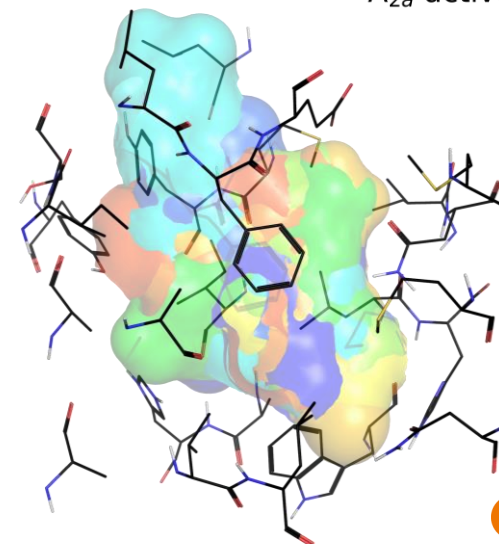
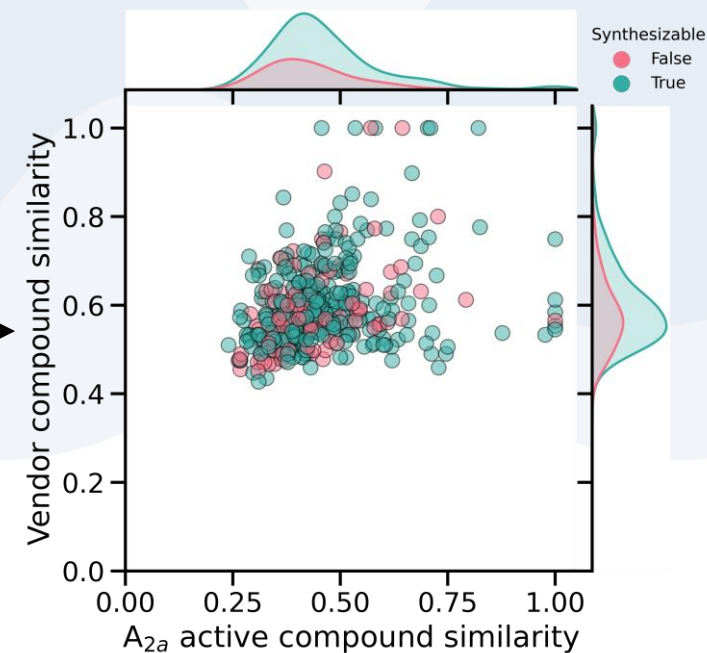


Some A_{2A} xtal structures recover more chemotypes than others

Prospective Generative Structure-Based A_{2A} Antagonist Screening*



Diverse array of predicted binding modes



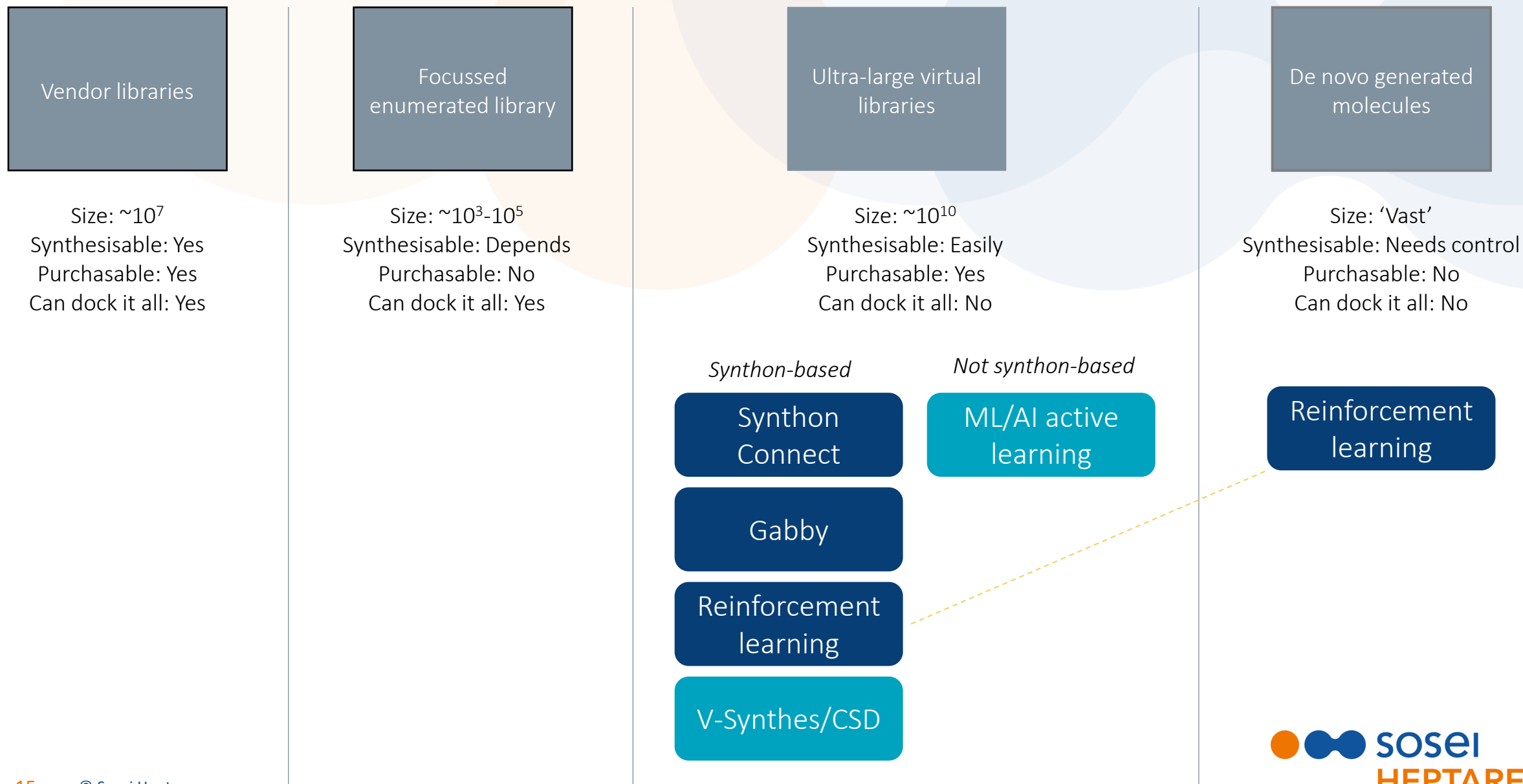
[1] Thakkar, A., et al. Chem Sci 12, 3339-3349 (2021)
[2] Mason, J. et al. TIPS 33, 249-260 (2012)

* Unpublished case study - Sosei Heptares, Univ Cambridge

Conclusions from A_{2A} generative modelling

- Generative modelling can successfully be applied to SBDD of GPCR structures
- In comparison to a ligand-based approach, structure-guided models allow capturing of key interactions in the binding pocket
 - However, a multi-parameter optimisation is needed to prevent the model exploiting inaccuracies in the scoring function
- Using multiple crystal structures identifies a broader range of known chemotypes
- The adenosine A_{2A} system is a good test bed for new structure-based virtual screening (SBVS) techniques:
 - A_{2A} antagonist SBVS hit was identified and optimised using A_{2A} StaR® X-ray SBDD, leading to a Ph2 clinical candidate, HTL1071/AZD4635
 - Langmead *et al. J. Med. Chem.* **2012**, 1904; Congreve *et al. J. Med. Chem.* **2012**, 1898, Borodovsky *et al. J Immunther Cancer* **2020**, 8: e000417
- Another recent example of successful application of SBVS to a GPCR is identification of an M₁ partial agonist:
 - M₁ SBVS fragment hit was identified and optimised using M₁ StaR® X-ray SBDD, leading to a PoM, as part of extensive SBDD program incl. multiple M₁/M₄ structures enabling design of further generations of selective agonists
 - Brown *et al. Cell.* **2021**, 184, 5886

Structure-based drug discovery in large chemical spaces



Searching REAL: A Readily Accessible ultra-large virtual space

- Enamine REAL
 - Publicly available – 5.5B (2022q1-2), 6.0B (2022q3-4)
 - 1.1B two-component reactions in 2022q1-2 Enamine REAL
 - Remainder are almost exclusively three-component reactions
- Enamine REAL Space
 - Available under NDA – 31.5B (2022q1-2), 36B (2022q3-4)
- 167 synthesis protocols, 137K BBs, 14M Bemis-Murcko scaffolds
- How can we efficiently screen/explore these ultra-large synthetically accessible virtual spaces in the context of SBDD?
- Let's look at three approaches
 - Each will have a budget of 1M protein-ligand dockings using Glide
 - Here, for simplicity we focus on finding good docking scores
 - In production we adopt a more sophisticated approach including post-processing, additional constraints, MPO



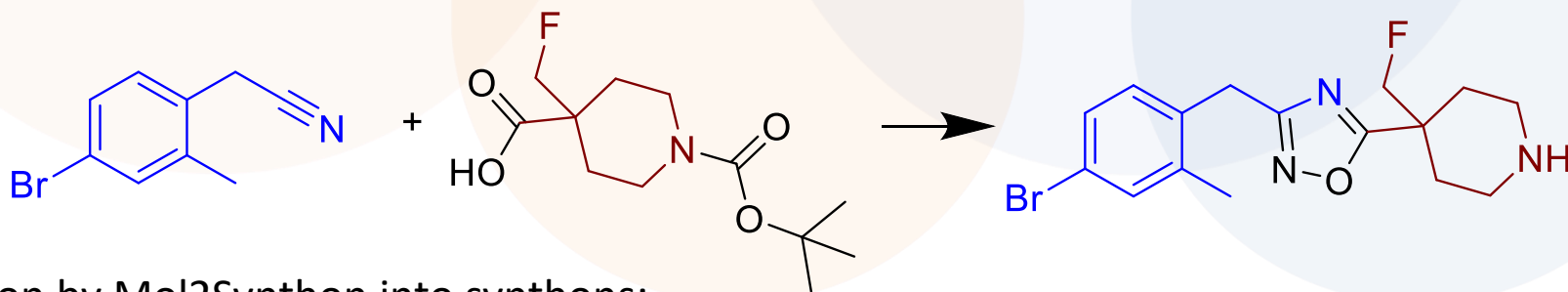
3

Mol2Synthon

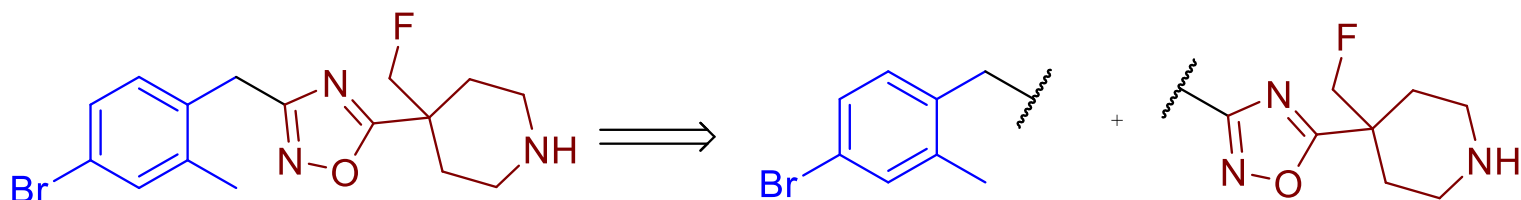
Make synthons from Enamine REAL

Mol2Synthon: example

- A 2-component reaction from Enamine REAL: note leaving groups, loss of atoms, gain of atoms



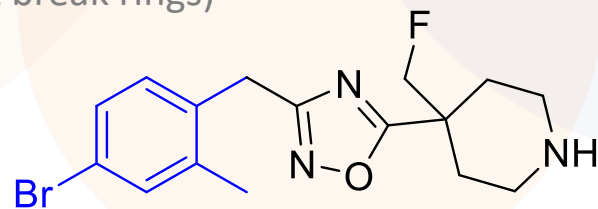
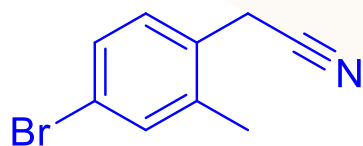
- Decomposition by Mol2Synthon into synthons:



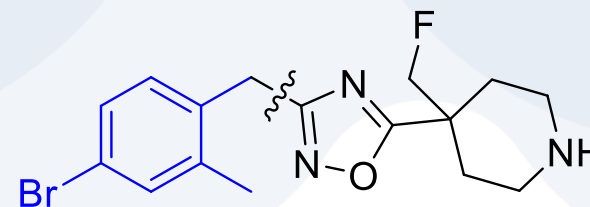
- Compared to using the original reactants:
 - Synthons incorporate the structural features of the final structure (e.g. in this case, the ring)
 - Synthons do not incorporate extraneous features (e.g. the protecting group, carboxylic acid, nitrile)
- => Synthons are better suited for fragment docking and other applications than the original reactants

Mol2Synthon: basic algorithm

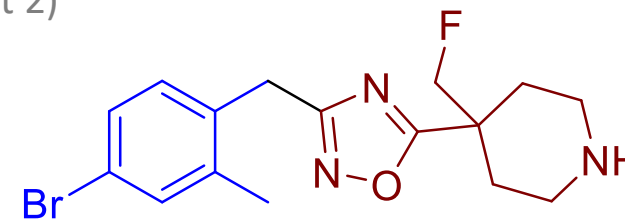
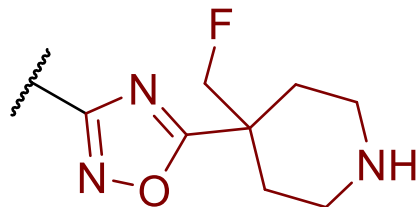
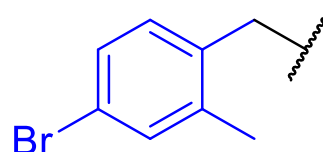
Map reactant 1 onto the product using MCS
(ignore bond orders, don't break rings)



Check whether only a single bond will be broken



If so, assign all of the remaining atoms to reactant 2
(If not, repeat from start trying reactant 2)



Break the connecting bond



4

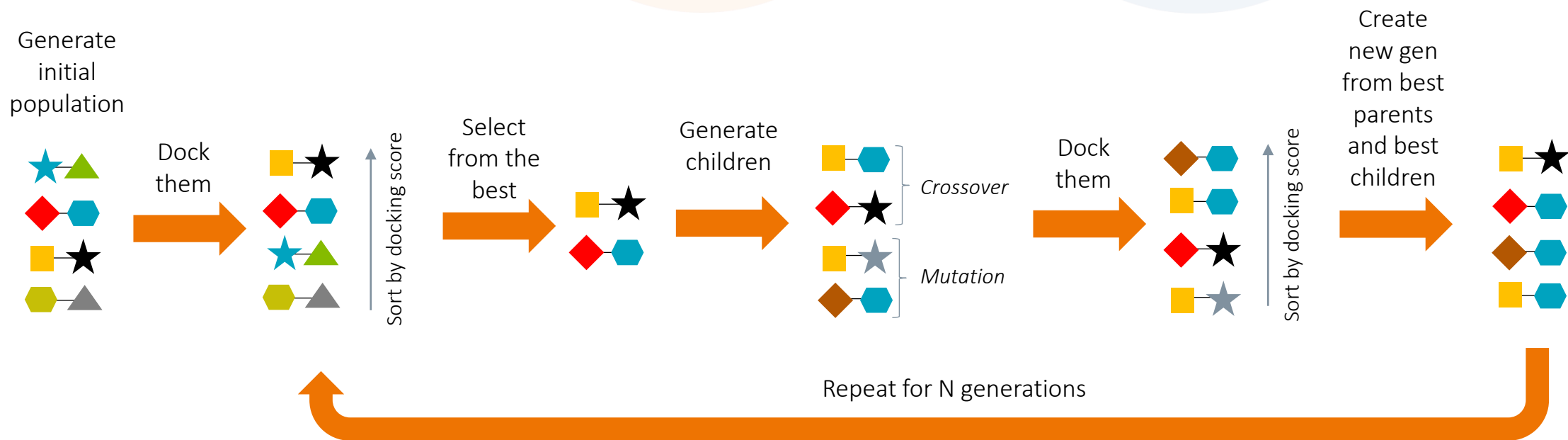
Gabby

Genetic Algorithm of building blocks

Gabby: a genetic algorithm to search across synthon-space

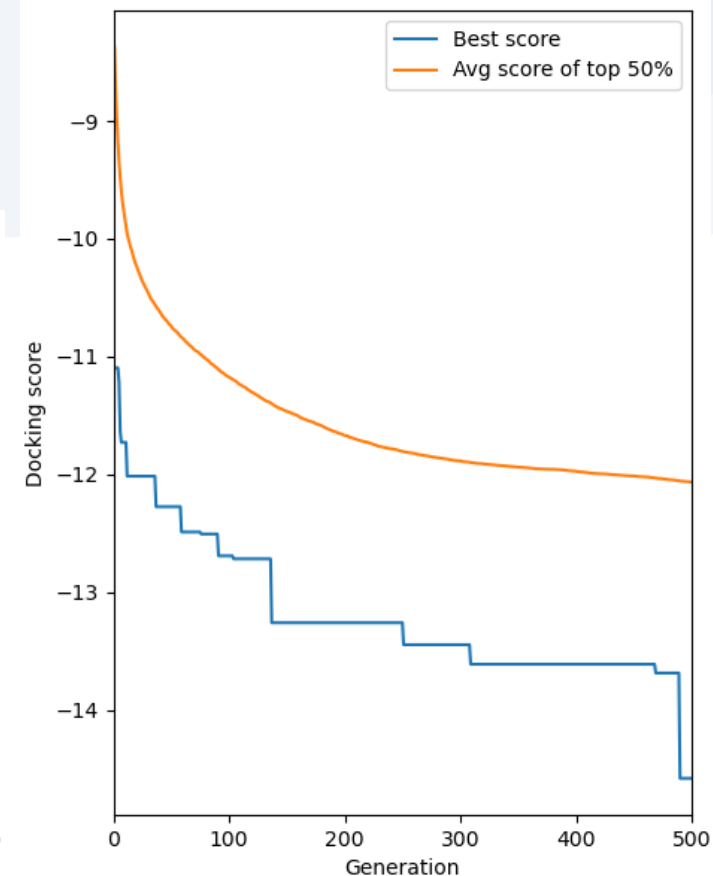
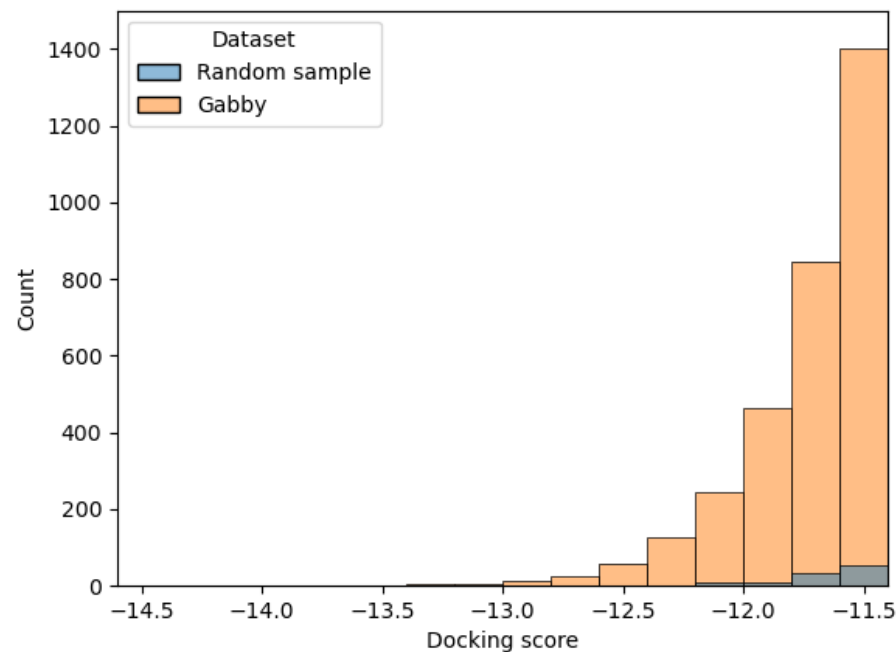
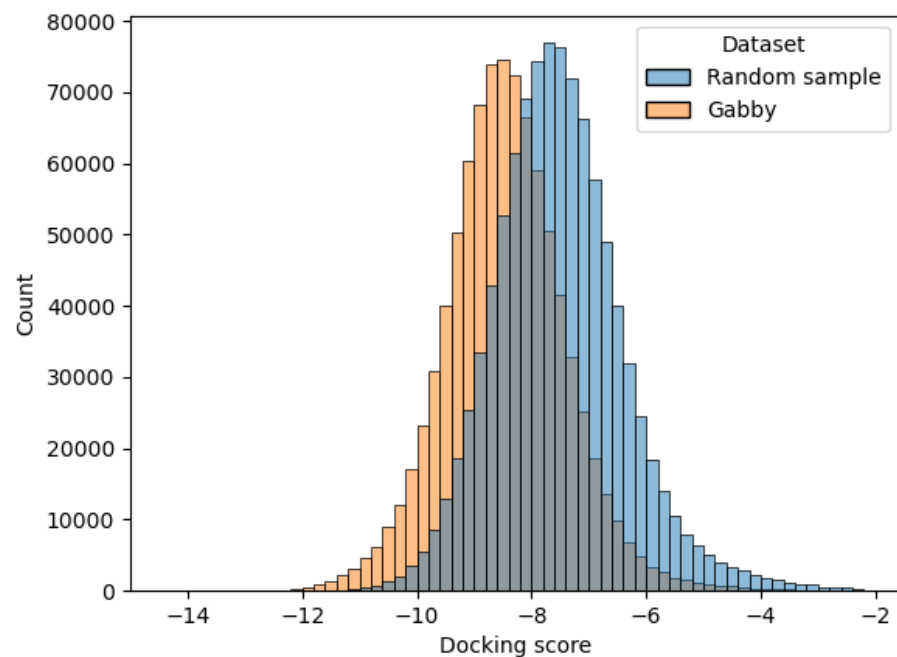
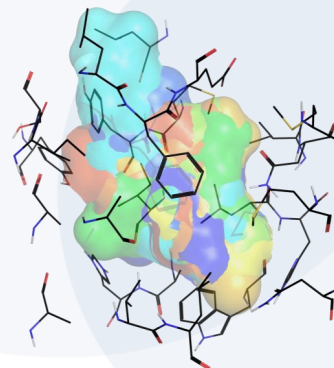
- Preparation:

- Distinguish the synthons from Mol2Synthon based on their SMILES, attachment point, and reactant position
 - This gave 313K synthons
- Measure the pairwise similarity of the synthons, after replacing the attachment point with Xe
- Describe each molecule in terms of the two synthons that compose it



Gabby results

- 902K dockings from 500 generations x 2000 molecules
- Best/1000th best score: -14.6/-11.8 (vs -13.2/-10.8 for random)
- Number of molecules with scores ≤ -12 : 477 (vs 12 for random)





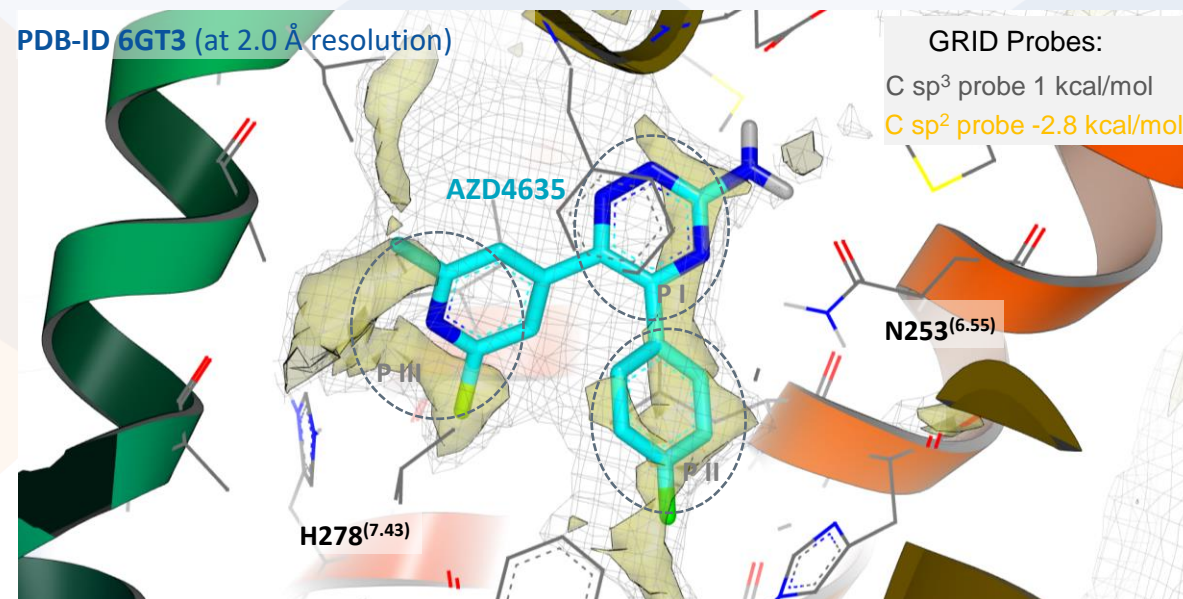
5

SynthonConnect

Join high-scoring synthon poses

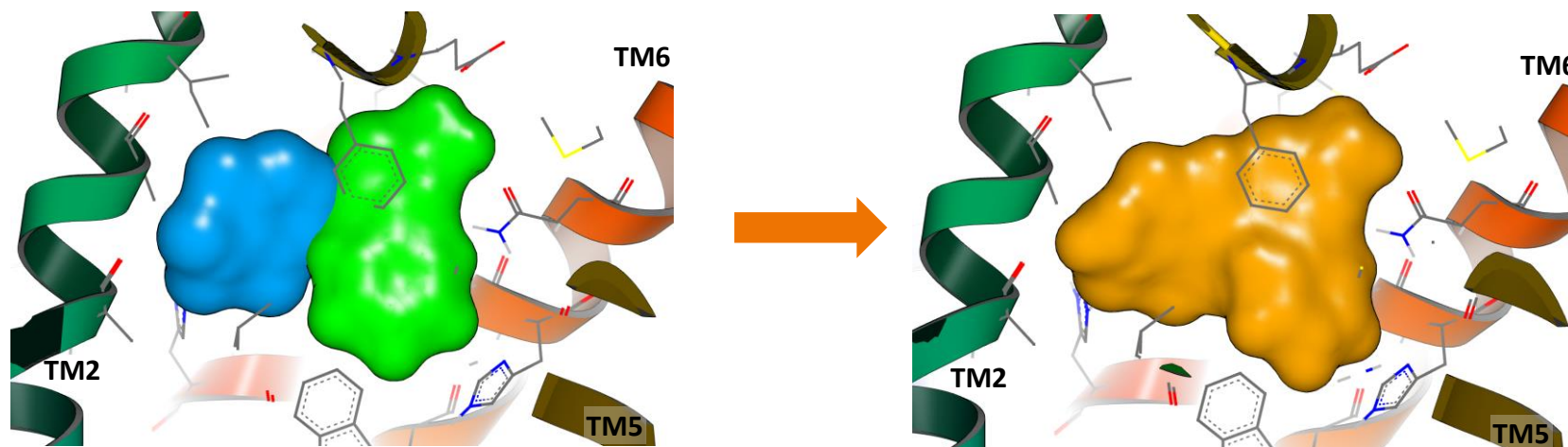
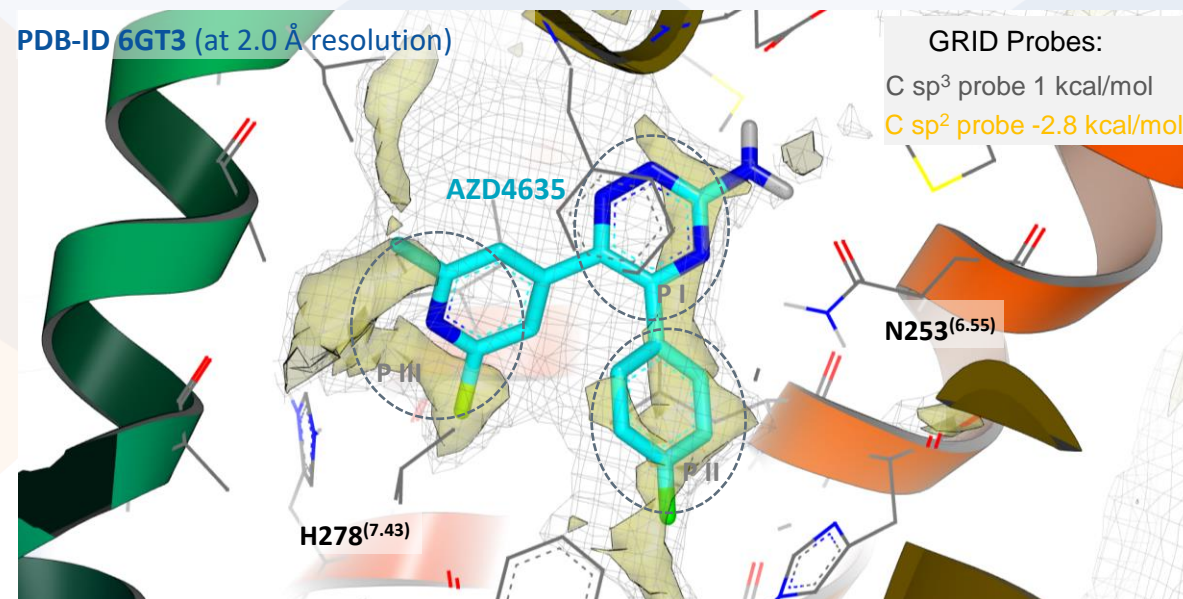
SynthonConnect algorithm

- Synthon attachment points are replaced with H
 - Reduces the number to dock to 265K
- Synthons are docked with Glide
 - Expanded sampling, max number poses of 100
- Iterate over the docked poses to find compatible synthon pairs, combinations where:
 - Attachment points (for pairs involved in the same reaction) are within range for a plausible bond to be made
 - Local orientation around the connection point is complementary
- Select from those with best sum of synthon scores



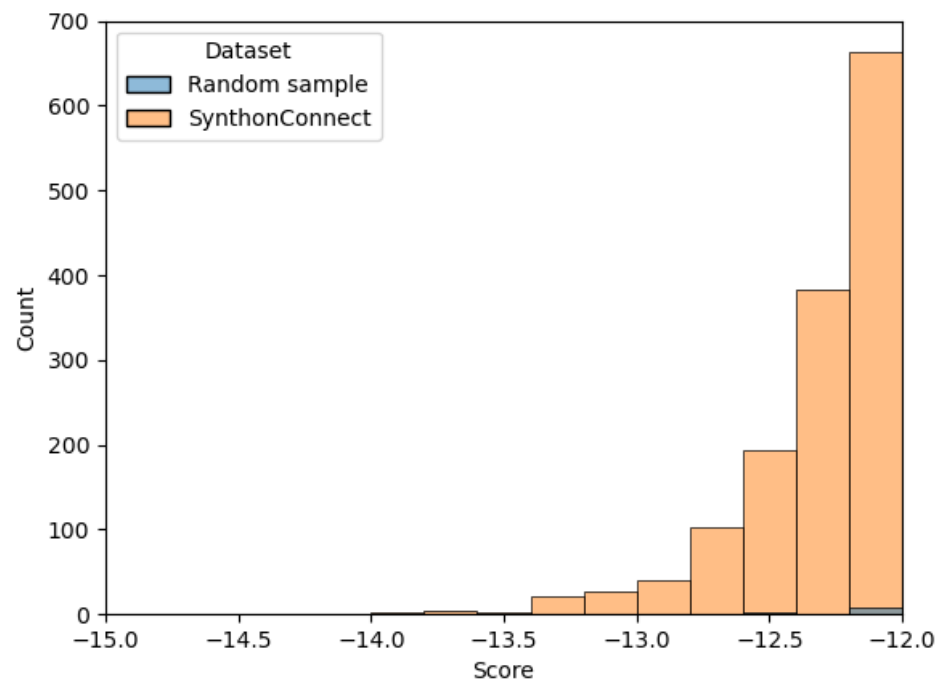
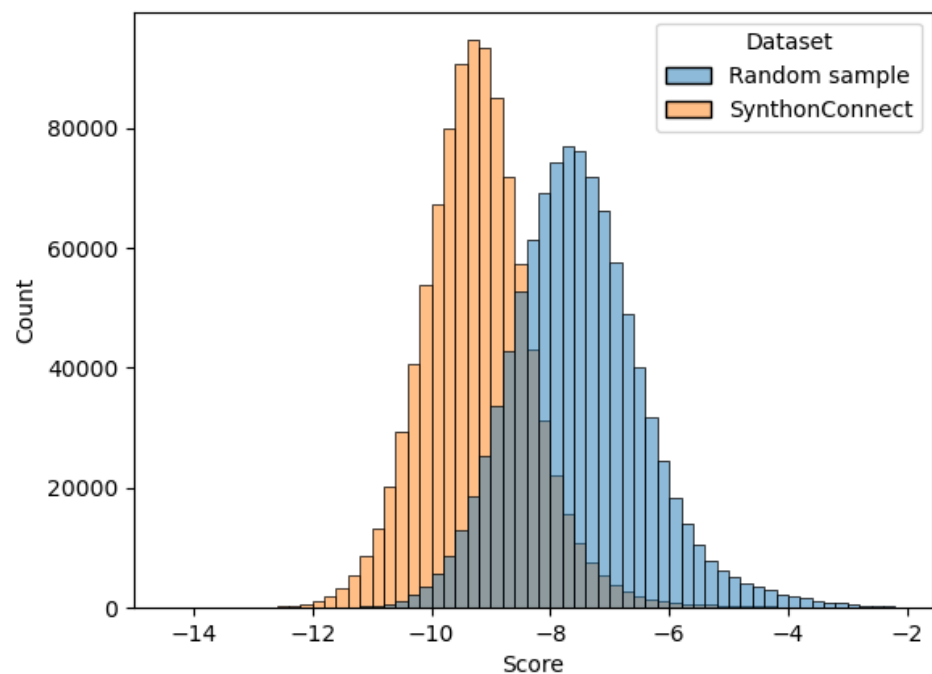
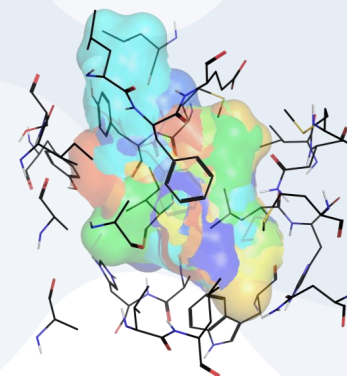
SynthonConnect algorithm

- Synthon attachment points are replaced with H
 - Reduces the number to dock to 265K
- Synthons are docked with Glide
 - Expanded sampling, max number poses of 100
- Iterate over the docked poses to find compatible synthon pairs, combinations where:
 - Attachment points (for pairs involved in the same reaction) are within range for a plausible bond to be made
 - Local orientation around the connection point is complementary
- Select from those with best sum of synthon scores



SynthonConnect results

- 1M dockings of Enamine REAL molecules
 - Plus additional dockings of synthons
- Best/1000th best score: -14.5/-12.1 (vs -13.2/-10.8 for random)
- Number of molecules with scores ≤ -12 : 1448 (vs 12 for random, 477 for Gabby)



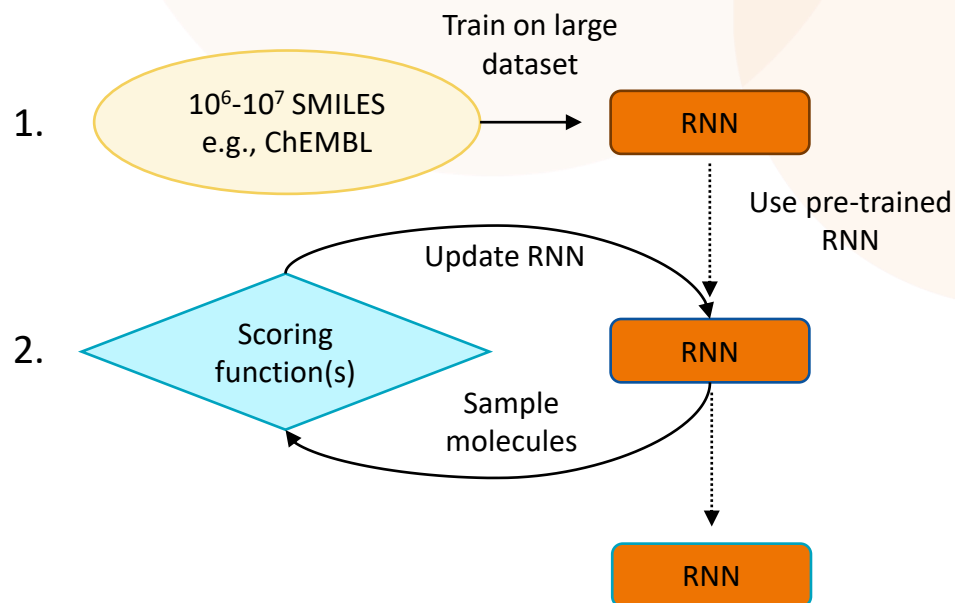


6

Generative design of Enamine REAL molecules

Generating molecules that look like Enamine REAL molecules

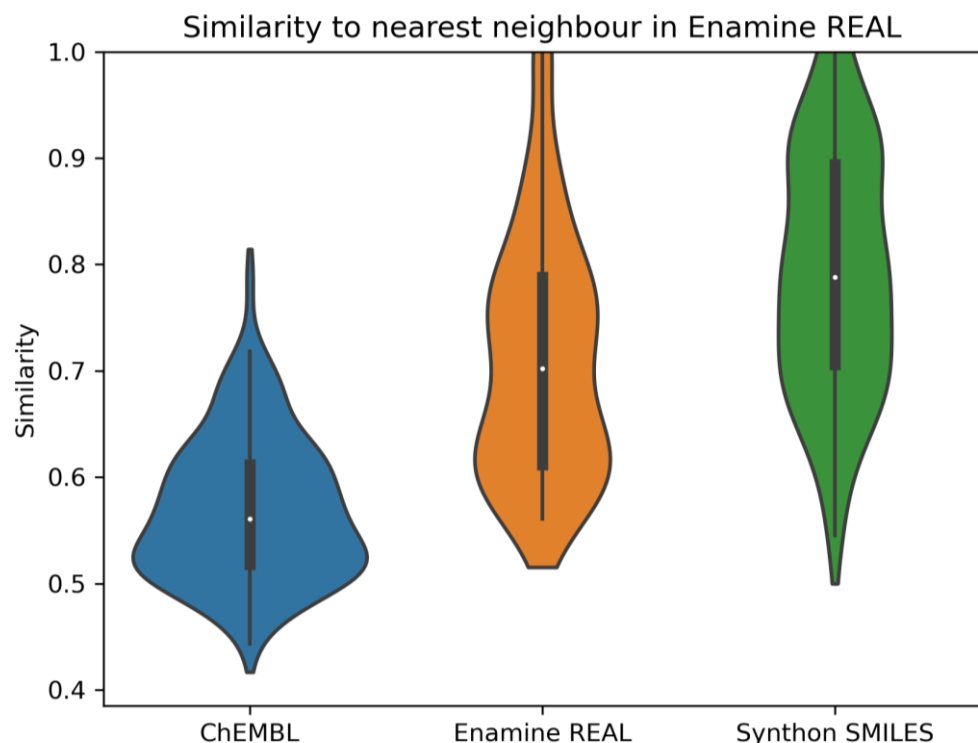
Reinforcement learning of large language models



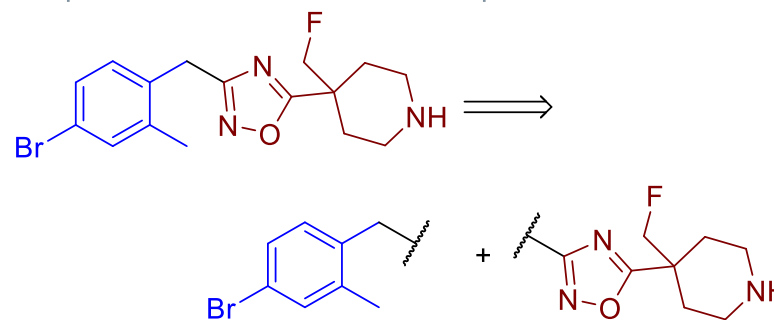
- During the reinforcement learning (RL) iterations, deviations from the original learned dataset probabilities are penalised
 - That is, if you train on ChEMBL, the generated molecules will resemble ChEMBL entries
- What if we instead train on Enamine REAL entries?
 - We can then use reinforcement learning over a virtual space that remains close to entries in Enamine REAL
 - Generated molecules or a close neighbour may be purchasable from Enamine REAL, rather than requiring synthesis

- M Olivecrona, T Blaschke, O Engkvist, H Chen. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, 9, 48
- M Thomas, RT Smith, NM O'Boyle, C de Graaf, A Bender. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. *J. Cheminform.* **2021**, 13, 39.
- M Thomas, NM O'Boyle, A Bender, C de Graaf. Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. *J. Cheminform.* **2022**, 14, 68.
- <https://github.com/MorganCThomas/SMILES-RNN>
- <https://github.com/MorganCThomas/MolScore>

Generating molecules that look like Enamine REAL molecules



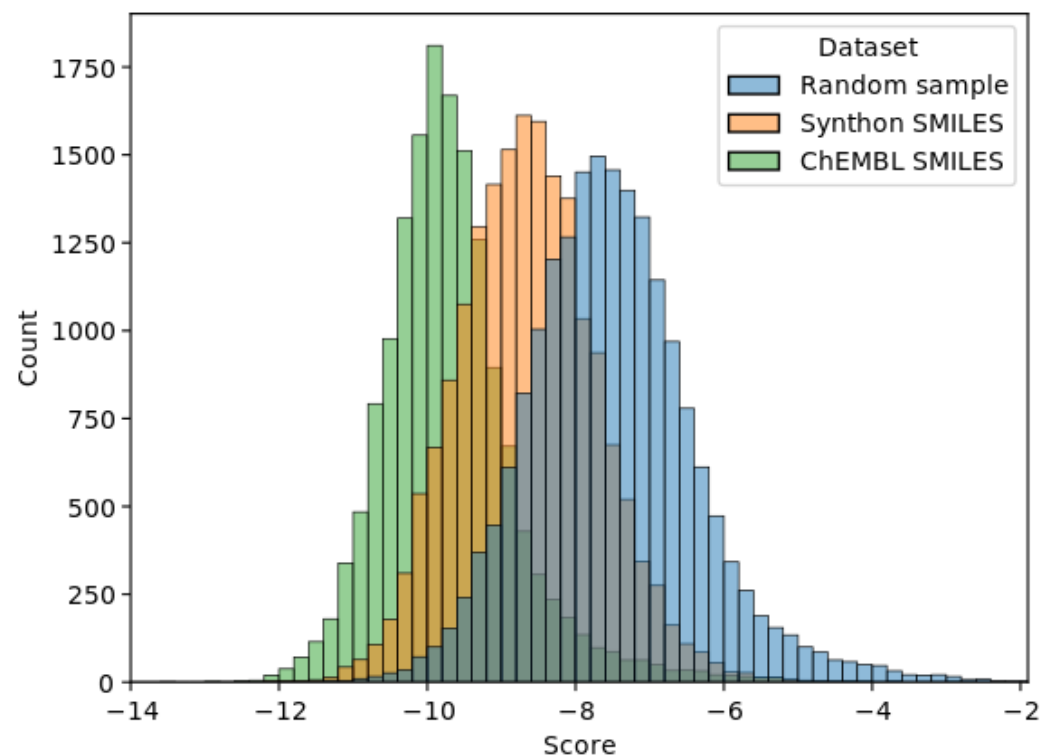
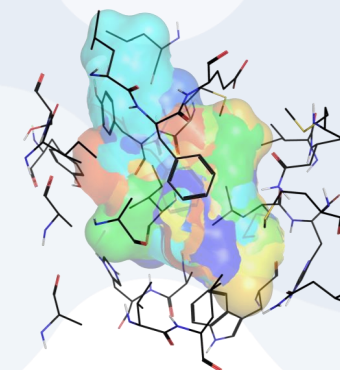
- During the reinforcement learning (RL) iterations, deviations from the original learned dataset probabilities are penalised
 - That is, if you train on ChEMBL, the generated molecules will resemble ChEMBL entries
- What if we instead train on Enamine REAL entries?
 - We can then use reinforcement learning over a virtual space that remains close to entries in Enamine REAL
 - Generated molecules or a close neighbour may be purchasable from Enamine REAL, rather than requiring synthesis
- Can we do better? Introducing 'Synthon SMILES'
 - Write each entry in terms of its synthons, both written with the attachment point first and separated by a dot disconnection
 - C9c1ccc(Br)cc1C.c91nc(C2(CF)CCNCC2)on1



- M Olivecrona, T Blaschke, O Engkvist, H Chen. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, 9, 48
- M Thomas, RT Smith, NM O'Boyle, C de Graaf, A Bender. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. *J. Cheminform.* **2021**, 13, 39.
- M Thomas, NM O'Boyle, A Bender, C de Graaf. Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. *J. Cheminform.* **2022**, 14, 68.
- <https://github.com/MorganCThomas/SMILES-RNN>
- <https://github.com/MorganCThomas/MolScore>

Generative model results

- 19K dockings from 300 generations x 64 molecules
- MPO involving docking score, RAScore, CLogP, consecutive rotatable bonds, HBD
- Best/1000th best score: -12.1/-10.1 (vs -11.6/-9.2 for 19K random, -14.6/-10.9 for ChEMBL SMILES)
- Number of molecules with scores ≤ -11.5 : 13 (vs 2 for 19K random, 197 for ChEMBL SMILES)



- M Thomas, NM O'Boyle, A Bender, C de Graaf. Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. *J. Cheminform.* **2022**, 14, 68.



7

Conclusion

Conclusions

	Requires synthons	Involves docking synthons	Can be restricted to REAL	Not just docking	Converges over iterations	Molecules with scores \leq -12	Progress
SynthonConnect						1448/1M	<div><div></div></div>
Gabby						477/1M	<div><div></div></div>
Generative model						4/19K	<div><div></div></div>
Active Learning							<div><div></div></div>
V-Synthes/CSD							<div><div></div></div>

- Mol2Synthon generates synthons that can be used as the basis for several approaches
- Both SynthonConnect and Gabby can identify molecules that dock with high scores
 - Caveat for Gabby is that the large number of iterations used here is inefficient
 - An advantage of SynthonConnect is that it only involves a single docking of synthons and then one of full molecules
- *De novo* design based around Enamine REAL is promising
- Work is continuing on all fronts to improve performance further, also finishing active learning and CSD
- Next step is to go beyond optimising docking score, include interactions and physicochemical properties.
 - For example, incorporating QSAR models (see [Poster 15](#)).

Acknowledgements

- **Sosei Heptares Computational Chemistry**
 - Chris de Graaf – Head of Group
 - Jon Tyzack – Mol2Synthon, SynthonConnect
 - Daniel Santos-Stone – Active Learning
 - Pierre Matricon, Francesca Deflorian, Jon Mason – A2A SBDD
- **Sosei Heptares Medicinal Chemistry**
 - Charlotte Fieldhouse – A_{2A} AI/SBDD MedChem
 - Robert Gillespie – A_{2A} AI/SBDD MedChem
- **University of Cambridge**
 - Andreas Bender – A2A generative modelling
 - Morgan Thomas – A2A generative modelling, MolScore, SMILES-RNN, AHC

Posters

- **Poster 15: Jon Tyzack**
 - Development of QSAR Models for SBDD of GPCRs
- **Poster 28: Morgan Thomas**
 - MolScore: A Semi-automated Platform for Generative Model Molecule Scoring and Evaluation in Drug Design
- **Poster 35: Sonja Peter, Anna Pallo**
 - Navigating the Orthosteric and Allosteric Structural GPCR Pocketome for Structure-Based Drug Discovery



**Current Job
Openings!!**