GENERATIVE DESIGN USING ACTIVE LEARNING OVER SYNTHON SPACE

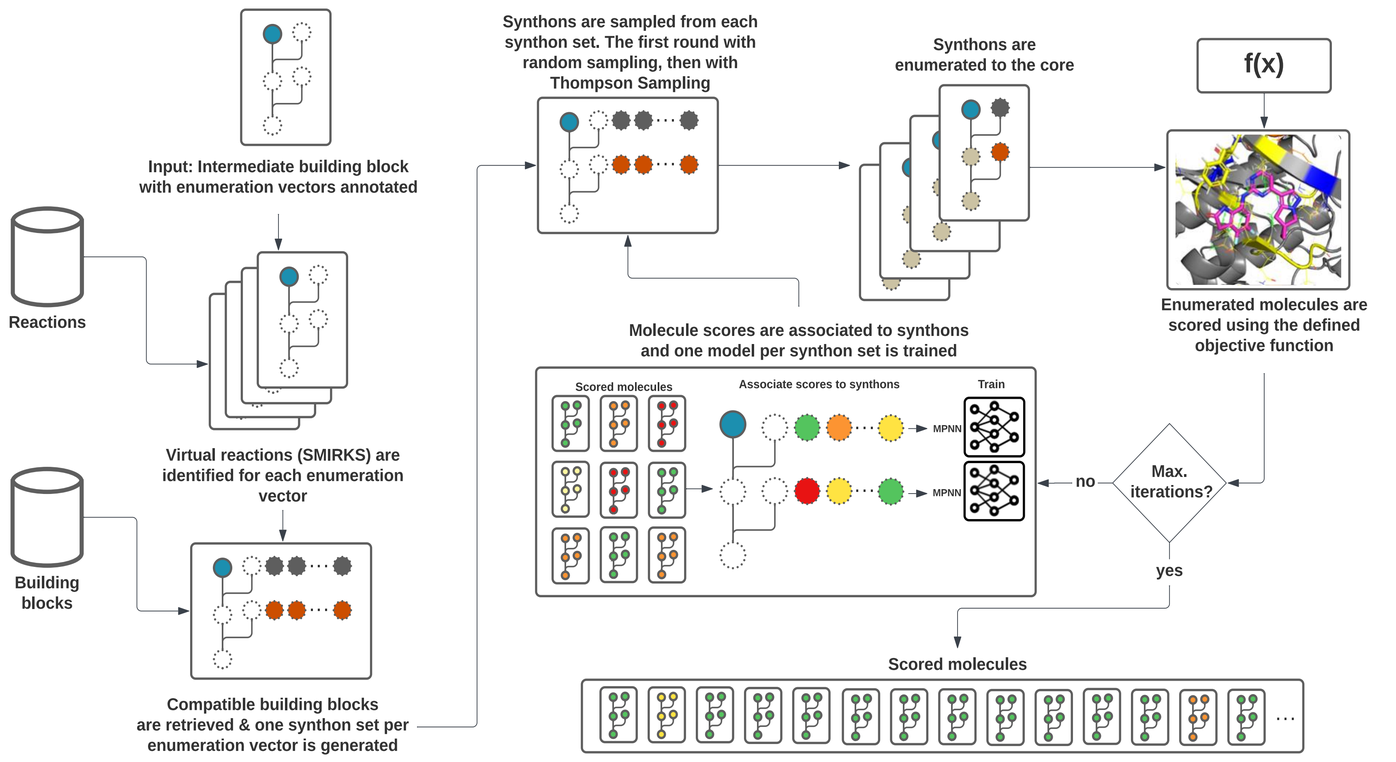
Mason C. Burlage, Tom Grigg, Adam Taouil, Oliver Scott, Dominique Sydow, Liam Wilbraham

Exscientia, Oxford Science Park, Oxford, UK

INTRODUCTION

- Current molecular generative design methods often lead to compounds that are challenging to synthesise due to the absence of explicit chemistry constraints
- Library enumeration strategies alleviate this by applying common medicinal chemistry reactions to commercially-available building blocks, creating a synthetically-tractable chemical space
- However, the combinatorial nature of library enumeration quickly leads to ultra-large spaces in which scoring beyond targeted enumeration schemes is computationally intractable
- To bridge the sample efficiency enjoyed by generative approaches with the synthetic tractability offered by library enumeration, we introduce scalable active learning via synthon acquisition (SALSA)

METHODS



RESULTS

Figure 1: SALSA is sample-efficient

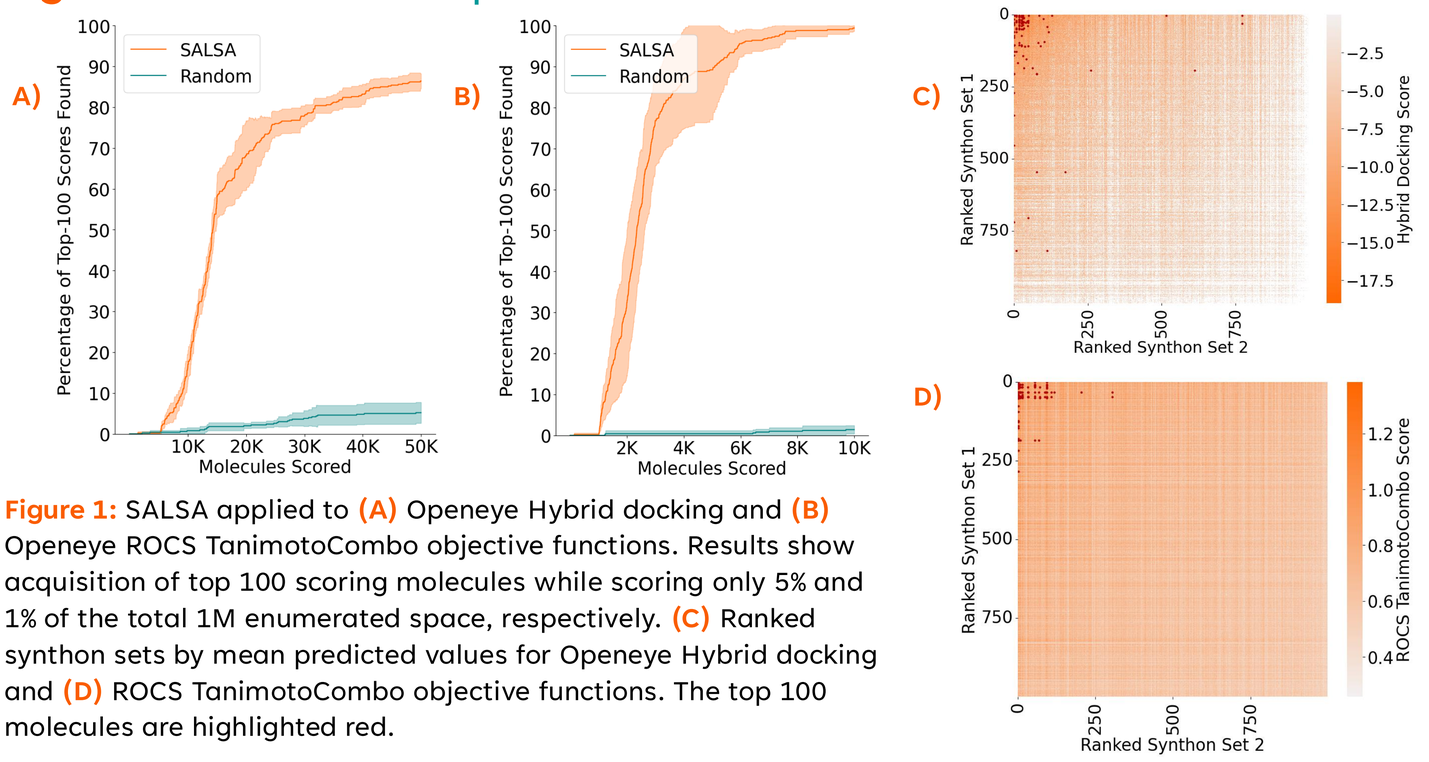


Figure 1: SALSA applied to **(A)** Openeye Hybrid docking and **(B)** Openeye ROCS TanimotoCombo objective functions. Results show acquisition of top 100 scoring molecules while scoring only 5% and 1% of the total 1M enumerated space, respectively. **(C)** Ranked synthon sets by mean predicted values for Openeye Hybrid docking and **(D)** ROCS TanimotoCombo objective functions. The top 100 molecules are highlighted red.

Figure 2: SALSA finds better solutions in larger chemical spaces

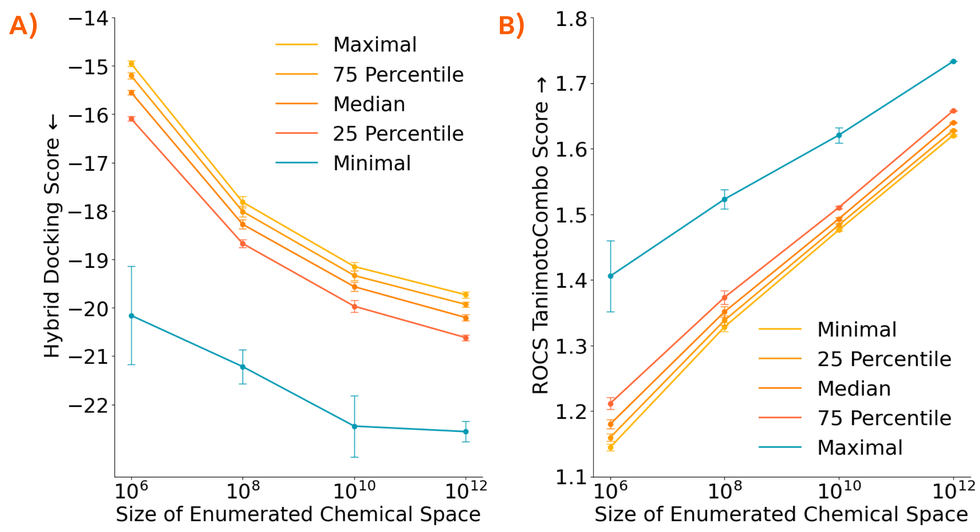


Figure 2: Scores of best 1K compounds acquired by SALSA in various size chemical spaces (budget of 100K objective function calls) with **(A)** Openeye Hybrid docking and **(B)** ROCS TanimotoCombo scores.

Figure 3: SALSA applied to multiparameter objectives across three targets

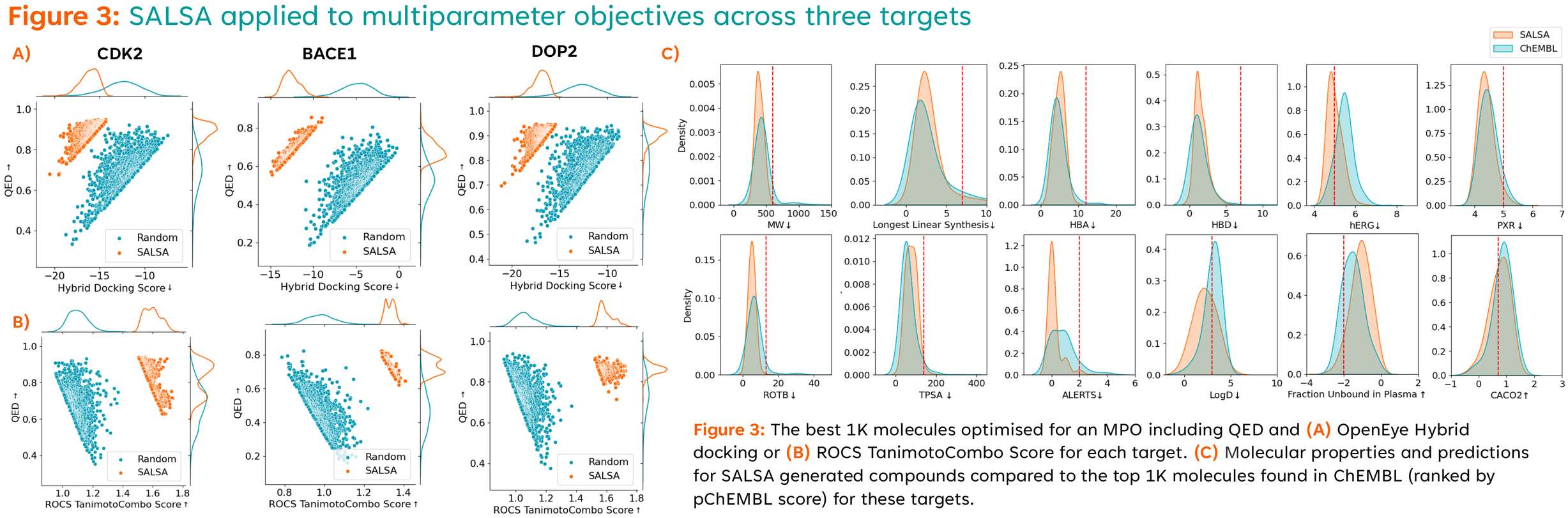


Figure 3: The best 1K molecules optimised for an MPO including QED and **(A)** OpenEye Hybrid docking or **(B)** ROCS TanimotoCombo Score for each target. **(C)** Molecular properties and predictions for SALSA generated compounds compared to the top 1K molecules found in ChEMBL (ranked by pChEMBL score) for these targets.

CONCLUSION

- We introduce a novel approach for constructing and navigating virtual synthesisable chemical space. SALSA leverages active learning to efficiently explore constituent synthons rather than entire molecules, enabling tractable search for optimal molecules in ultra-large virtual spaces
- By evaluating just 5% and 1% of a 1M molecule space, we can successfully obtain 85% and 99% of the top-performing molecules for docking and ROCS objective functions, respectively
- Experiments demonstrated that SALSA's efficiency extends to much larger chemical spaces, consistently discovering superior molecules as the size of the chemical space expanded
- SALSA optimises molecules for multiparameter objective functions, with the resulting molecules exhibiting similar ADME and synthesisability metrics to their ChEMBL counterparts

REFERENCES

1. Gao et al. J Chem Inf Model, 2020
2. Graff et al. Chem Sci, 2021
3. Mendez et al. Nucleic Acids Res, 2019

For more information
mburlage@exscientia.co.uk

Exscientia