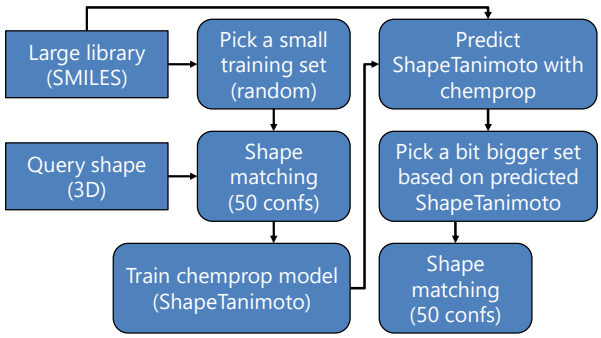
MACHINE LEARNING-BOOSTED SHAPE MATCHING (HASTESM): SEARCHING ENUMERATED GIGA-SCALE VIRTUAL LIBRARIES WITH LARGE CONFORMER ENSEMBLES

Tuomo Kalliokoski, Samuli Näppi and Ainoleena Turku
Orion Pharma, Espoo, Finland. E-mail: tuomo.kalliokoski@orionpharma.com

HASTESM process



Testing HASTESM with targets A and B

Query	Heavy atom count	MW	Rotatable bonds	ShapeTanimoto to the other query
A	36	483.571	7	0.62
B	31	413.481	3	0.61

Query	Enamine REAL subset size	Heavy atom range	Virtual hit Shape Tanimoto cutoff	Estimated number of virtual hits in the subset (based on random sampling of 1M)
A	77,322,795	30 - 38	0.75	21,109
B	984,006,846	29 - 33	0.85	113,161

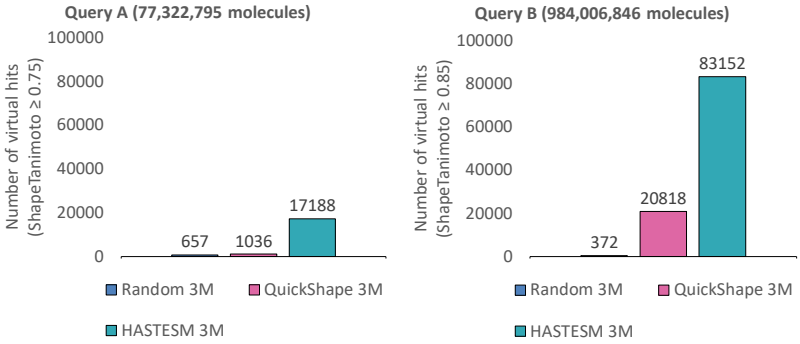
Training set of 500k-1M compounds is suitable for billion-scale libraries

Training set size	Fraction of the training set from DB (%) (A)	Fraction of the training set from DB (%) (B)	Virtual hits (A)	Virtual hits (B)	Estimated recall of virtual hits (A)	Estimated recall of virtual hits (B)
10k	0.013	0.001	11,278	45,677	0.53	0.40
100k	0.129	0.010	15,712	64,382	0.74	0.57
500k	0.647	0.051	16,647	77,041	0.79	0.68
1M	1.293	0.102	16,996	83,349	0.81	0.74

Comparison to the current state-of-art

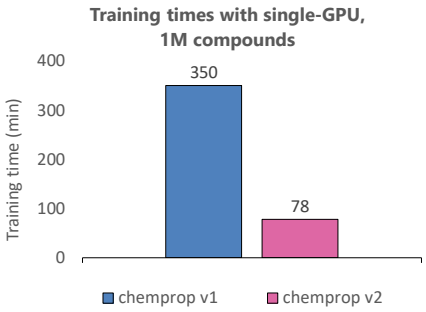
For baseline method comparison, we used the Dixon and Merz 1D similarity method [1] implemented in Schrödinger Suite as a part of the QuickShape workflow [2].

The machine learning utilized in HASTESM outperforms the simple 1D descriptor method



Runtime for HASTESM in our AWS cluster

Task	# cores	Query A time (min)	Query B time (min)
Shape matching for training set	500 for confgen, 40 for shape matching	74	91
Training chemprop model	NVIDIA Tesla T4 (8 workers)	45	54
Predict ShapeTanimoto	500 cores	29	187
Final Shape matching	500 for confgen, 40 for shape matching	212	202
TOTAL PROCESS	500 cores, NVIDIA Tesla T4	360	534



Chemprop v2 is notably faster than chemprop v1

No multi-GPUs required for training

HASTESM software implementation

- Main program creates slurm jobs and processes intermediate results for efficient data transfers with low bandwidth
- Distributes the CPU workloads of conformer generation, shape screening and chemprop prediction on slurm cluster
- Chemprop training uses a single GPU node with multiple dataloader workers to speed up batch construction
- Memory usage is controlled by chunking the inputs for CPUs
- Stores intermediate results in SQLite databases
- Checkpoint system for error recovery

Samuli Näppi was funded by Orion Phase 1 Trainee program. Phase 1 program offers possibility for students in the fields of natural sciences, pharmacy, technology and economics to show their talent in an international pharmaceutical company. For more information, please see:
https://www.orion.fi/en/careers/students/phase1-trainee-program/

[1] Dixon SL and Merz KM: *J. Med. Chem.* 44 (23), 3795-3809, 2001. DOI: 10.1021/jm010137f
[2] QuickShape Screening: https://www.schrodinger.com/wp-content/uploads/2023/12/23_491_QuickShape-Screening_White-Paper_Mkt_R3-2-1.pdf