QuickShape screening in the age of ultra-large libraries

Tim Knehans, Steve Dixon, Nick Boyles, Jianxin Duan, Chris von Bargen, Volodymyr Babin, Steve Jerome, Matt Repasky

Readily available chemical libraries continue to increase rapidly in size while the need for novel chemical matter for drug targets remains at an all time high. Although 3-dimensional Shape-based methodologies provide an effective approach to enrich actives by finding hits that are topologically different but similar in shape, their hardware requirements can become prohibitive for screening ultra-large chemical libraries. Here we compare the performance of a 1-dimensional (1D) pharmacophore-based fingerprint with GPU-based Shape Screening both in terms of enrichment of known actives as well as computational speed.

The 1D method not only retains enrichment at levels very similar to Shape Screening with high overlap of recovered actives but also provides distinct hits. Additionally, it allows screening of billions of molecules on a conventional laptop within a few days and requires 100 times less storage. By combining both 1D and 3D approaches, we developed an automated staged workflow called QuickShape enabling efficient 3D-Shape screenings for library sizes in the tens of billions of molecules. This workflow is an excellent ligand-based screening solution in the new era of ultra-large libraries.

Schrödinger

1D-fingerprint technology enables pre-screening of Enamine REAL in a few hours within QuickShape workflow

QuickShape is a cascaded screening workflow using 1D pharmacophoric fingerprint screening followed by Shape Screening. Pharmacophores are perceived for each molecule with appropriate protonation and tautomer states and subsequently projected into a fingerprint vector. The distance between the pharmacophoric points is the number of bonds between the pharmacophores.
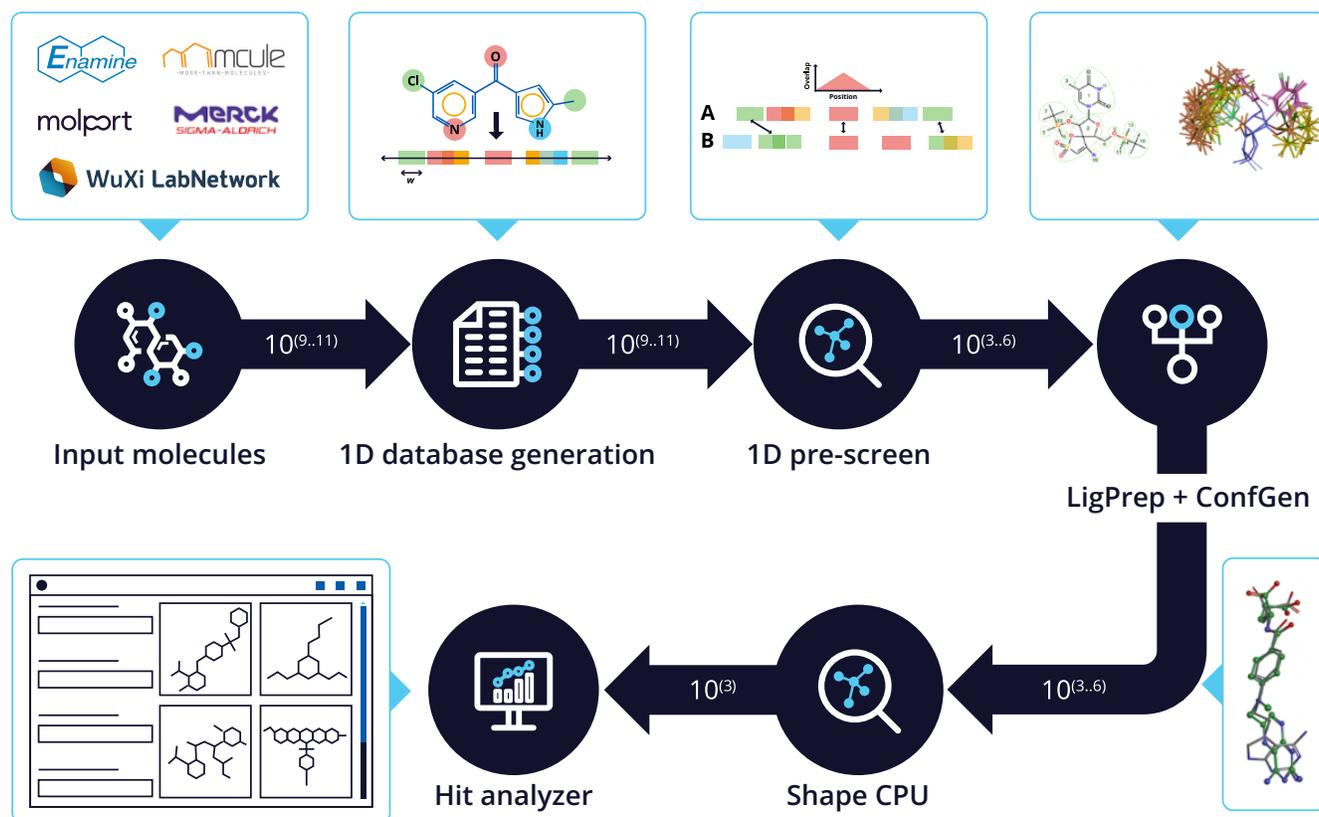
This process takes approximately 1 CPU hour per 1 million molecules. The output database requires 100 times less storage compared to equivalent Shape GPU databases. A 6.5 billion compound database takes up only 400 GB.

The similarity is defined by comparing the overlap of equivalent pharmacophores between the molecules A and B:

**Sim(A, B) = Overlap(A, B)/max[Overlap(A, A), Overlap(B, B)]**

The similarity comparison is extremely fast and parallelizable. It is possible to screen 6.5 billion compounds on a computer with 10 cores in 14 hours.

The hits from 1D screening are triaged using Shape Screening and the final hit list can be analyzed based on diversity, specific pharmacophore criteria and property filters using the Hit Analyzer panel in Maestro. The complete workflow can be run on a laptop in a few days.

Input molecules $\quad 10^{(9..11)} \quad$ 1D database generation $\quad 10^{(9..11)} \quad$ 1D pre-screen $\quad 10^{(3..6)} \quad$ LigPrep + ConfGen

$10^{(3..6)} \quad$ Shape CPU $\quad 10^{(3)} \quad$ Hit analyzer

1D Screening provides similar enrichment compared to Shape screens

First we validated the enrichment of 1D similarity against that of Shape using the actives from the DUD-E dataset seeded in 1.4 billion molecules from Enamine REAL. The five most diverse actives were used as probes. Both methods provide very similar enrichment of known actives. Among top 200,000 hits, actives were identified for all targets, indicating 1D screening alone can be useful. It is important to note that real enrichment may be much higher because there are likely additional unknown actives in the Enamine REAL database.
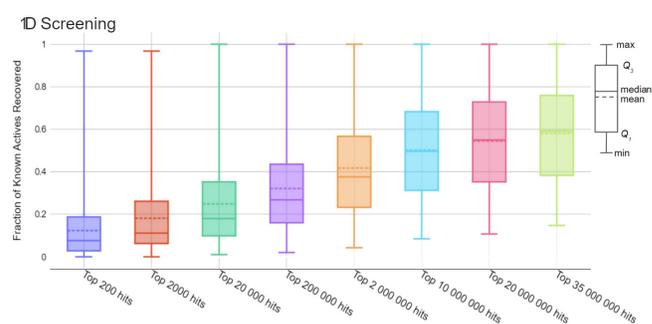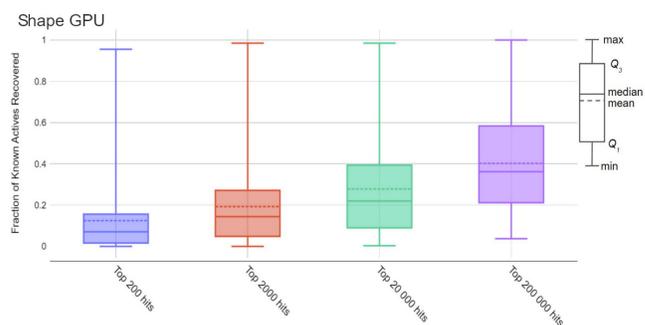

1D Screening & Shape have good correlation and high overlap

Secondly, we ensure that 1D screening is able to retrieve the same molecules as Shape, which is represented by the percent overlap among the top hits. Among the top 20 million hits from 1D screening we find a minimum 50% and on average 90% of the top 200,000 hits from Shape screening. Already at 200,000 returned hits, the average overlap is >60% indicating that it may be sufficient to run 100 times less Shape calculations.
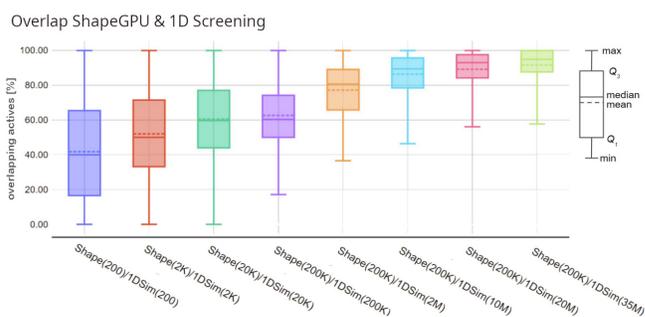
The 1D similarity score is highly correlated with Shape similarity with an R^2 of 0.65. This demonstrates that 1D screening is able to reduce the number of Shape comparisons by 84% while maintaining similar performance.
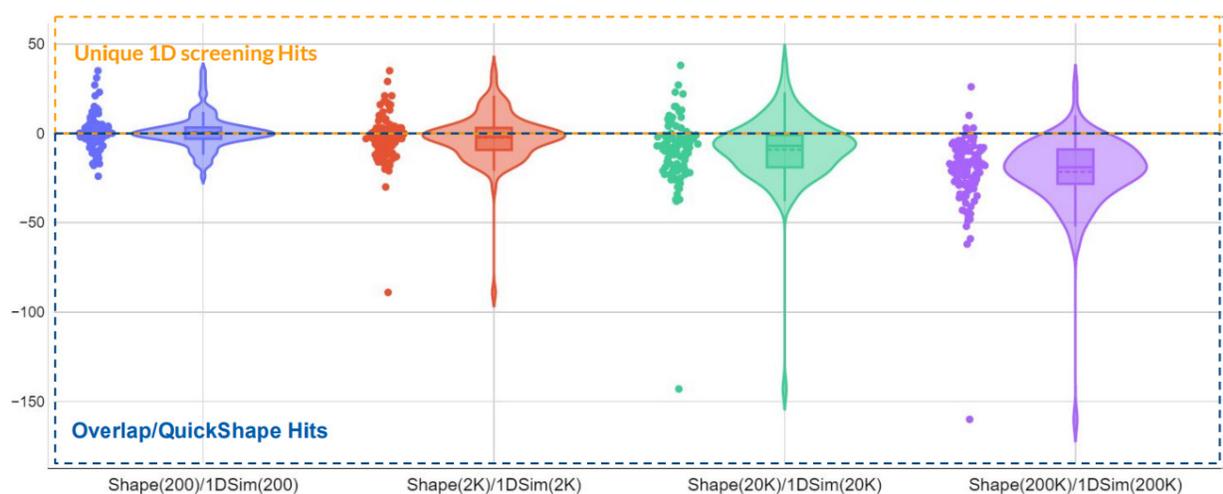

3
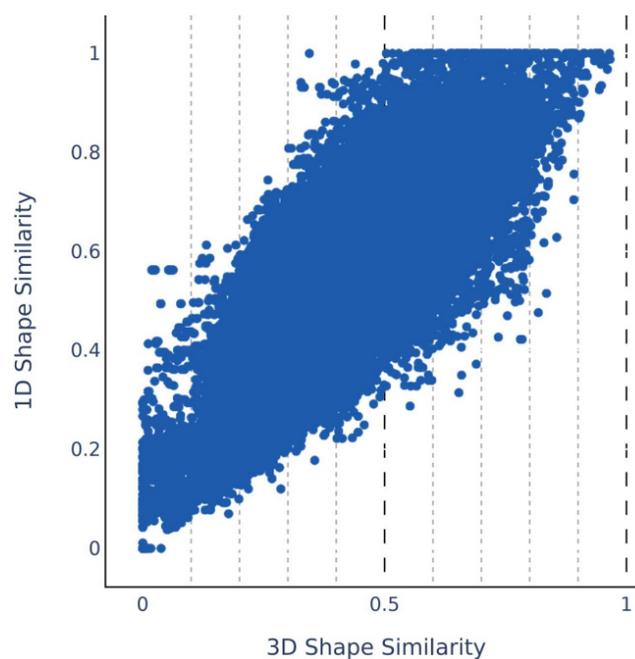
QuickShape and its constituent technologies should be used in a combined fashion

Since the enrichment between Shape and 1D screening (1DSIM) are similar but neither correlation nor overlap are 100%, we visualized the trends in unique contribution to the recovered actives by 1D Screening via calculating:

#actives1DSIM - 2 x #actives1DSIM ∩ Shape

In this representation, all values above 0 show targets where at least twice as many actives uniquely returned by 1D screening compared to overlapping hits.

As with previous figures the more molecules are returned by each methodology the bigger the overlap gets and the distributions are shifted below 0 which makes these molecules accessible to the QuickShape workflow. However, in all bins, significant amounts of molecules are above and close to 0. This indicates the ability of 1D screening to contribute valuable hit matter on its own hence it is generally recommended to investigate the top-ranked molecules by either methodology in order to increase the likelihood of novel hit identification.