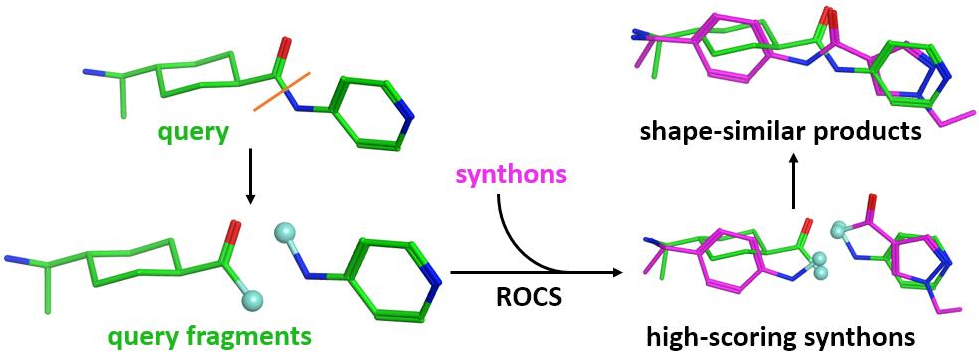
Shape-Aware Synthon Search (SASS) for virtual screening of synthon-based chemical spaces

Chen Cheng[1] and Paul Beroza[1]

[1]Discovery Chemistry, Genentech, South San Francisco, USA

TOC



Abstract

Virtual screening of large-scale chemical libraries has become increasingly useful for identifying high-quality candidates for drug discovery. While it is possible to exhaustively screen chemical spaces that number on the order of billions, indirect combinatorial approaches are needed to efficiently navigate larger, synthon-based virtual spaces. We describe Shape-Aware Synthon Search (SASS), a synthon-based virtual screening method that carries out shape similarity searches in the synthon space instead of the enumerated product space. SASS can replicate results from exhaustive searches in ultra-large, combinatorial spaces with high recall on a variety of query molecules while only scoring a small subspace of possible enumerated products, thereby significantly accelerating large-scale, shape-based virtual screening.

Introduction

In the past decade there have been significant advances in the scale and success of virtual screening campaigns.[1, 2, 11–14, 3–10] Traditionally, the scope of virtual screening was limited to compounds that were either present in local chemical inventory or available for purchase from vendors of chemical screening compounds. In either case, physical samples of compounds that were identified as desirable by computed metrics could be easily obtained and assayed. The scale of such efforts was millions of compounds, which puts a heavy, though manageable, resource burden on the computational methods.

The recent development of synthesis-on-demand libraries, such as those offered by Enamine,[15] Wuxi AppTec,[16] OTAVA,[17] ChemSpace,[18] and eMolecules,[19] has greatly expanded the scope of virtual screening. Instead of including only physically available compounds, chemical spaces now include compounds that are assumed to be readily synthesizable via validated reactions and reagents. This assumption greatly expanded the chemical space of available compounds that are relevant to drug discovery from millions to billions and beyond.[20, 21] Virtual screening methods have managed to keep pace, with several reports in the literature describing docking or shape similarity screening campaigns of up to a billion compounds.[6, 8, 22, 23] One important observation in these studies was that by increasing the size of the virtual screens, not only were more potent molecules found, but also that the novelty and quality of the hits improved. Thus, searching larger chemical spaces is worth the extra computational effort.

A significant roadblock to virtual screening on very large scales develops when the number of synthesis-on-demand compounds outpaces the ability to explicitly evaluate them computationally. Often the mere instantiation (i.e. creating a representation of the molecule in computer memory) and storage of an entire synthesis-on-demand library *in silico* is not trivial, and neither is the evaluation of the fully enumerated product space. To navigate such unenumerable libraries, one approach involves using generative models to propose in-library molecules based on learned distribution from a subset of the library.[24, 25] In another approach, similar to how the libraries are defined by the synthons, virtual screening methods evaluate the synthons instead of the full libraries and only instantiate and evaluate products formed from top synthons, thereby limiting the computational resources required to evaluate such large chemical spaces. Several 2D molecular graph-based virtual screening methods that rely on evaluation of synthons have been reported to search ultra-large chemical spaces efficiently.[26–30]

Only recently, however, has such synthon-based approach been adopted for 3D virtual screening. Two such methods are Chemical Space Docking[14] and V-SYNTHES[9], which first dock individual synthons to identify promising synthons and then instantiate and evaluate only products formed from those synthons. Those methods have hit rates well over 10%. In addition, a genetic algorithm-based approach has been reported for exploring synthon-based libraries using pharmacophore similarity as the scoring function.[31]

Ligand shape similarity is an additional important 3D method that, to our knowledge, has not been applied to synthon-based virtual screening. In this paper, we describe such a method that we term Shape-Aware Synthon Searching (SASS). As in the case of the synthon-based docking approaches, our shape-based approach identifies promising synthons and combines them to make products. Evaluation of only a subset of products formed from promising synthons is crucial to limiting the search space. This method uses ROCS[32] to calculate shape similarity,[33] but any shape-based similarity method[34] should be compatible with the approach.

We first describe SASS for two-component reactions, while drawing attention to some of the important considerations and parameters involved. We then apply the method to several shape queries to show that this method can achieve high recall on a diverse set of queries when compared to search results on fully enumerated libraries, while using a fraction of the computational cost. Finally, we demonstrate that this method can be scaled to a very large number of compounds while maintaining performance.

Methods

To search ultra-large chemical spaces, SASS evaluates synthons instead of the products they form. Synthon scores are combined, and only the top synthon combinations are selected to generate a subset of products to be scored that is much smaller than the fully enumerated library. This approach reduces the search space from n^k (size of the fully enumerated library) to n x k (total number of synthons) + m, where n is the number of synthons per reaction component, k is the number of reaction components, and m is the size of the subset of products to be scored in the second stage (after synthon scoring).

In this work, synthons are evaluated based on 3D shape similarity to a query molecule, which is complicated by the fact that synthons are usually much smaller than the query. In addition, we do not seek the top-scoring synthons when compared to the whole query molecule, but instead seek compatible synthons from the same reaction that score well when compared to complementary *substructures* of the whole query. When such synthons are combined according to the reaction rules, the resulting products will likely have high shape similarity to the whole query.

As a result, we split the query into a pair of query fragments, and score synthons against each query fragment. To ensure that the connector atoms on the synthons roughly point toward each other, we apply custom ROCS color features during the ROCS scoring (see *Synthon scoring* section for details). After scoring, we rank the synthon combinations by their aggregated scores. The top-m synthon combinations are used to generate m products, which are then scored with ROCS to generate their final ranking. The steps of SASS are summarized in Fig 1. Details of each step are described below.

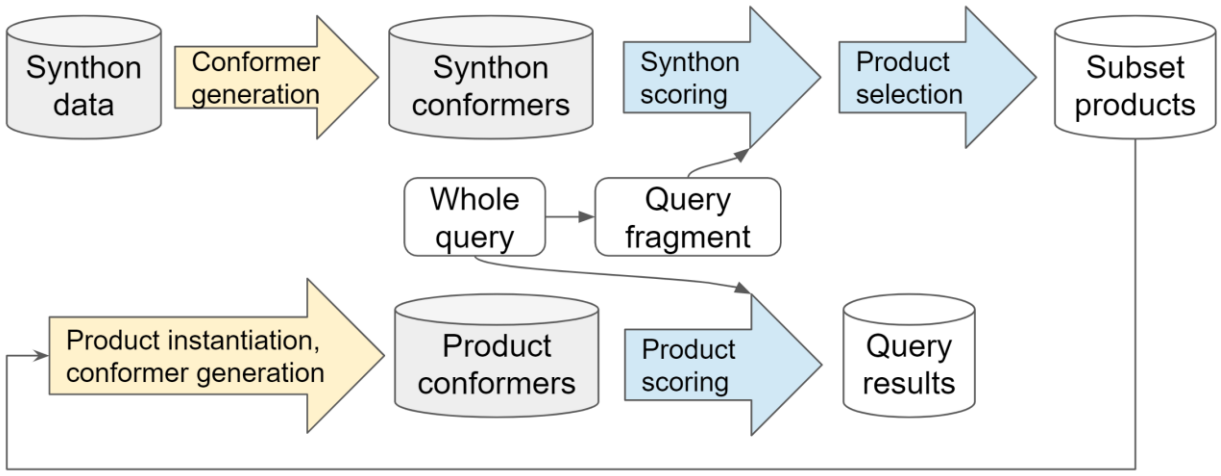


Fig 1. Overall schema of SASS. Synthon conformers are enumerated and scored against query fragments. Top synthon combinations are instantiated to form a subset of the products of the full chemical space. Conformers from these products are then enumerated and scored by their shape similarity to the complete 3D query molecule.

Query fragmentation

We split a 3D query molecule into fragments by cleaving acyclic and cyclic bonds (Fig 2). To generate two fragments from a query, one acyclic bond or two cyclic bonds that are in the same cycle are cleaved. The cleaved bonds are capped with special connector atoms to preserve the information on query fragment connectivity. For these connector atoms, heavy isotopes ^{13}C and ^{14}C are used, because they are normally not present in synthons and are thus ideal for tracking connection points. To avoid generating overly small fragments that are unlikely to match well to any synthons, we remove any fragment pairs where one of the cleaved fragments had fewer than 5 heavy atoms, or cyclic fragments that had fewer than 5 heavy atoms outside the partial ring (i.e. not counting the former ring atoms).

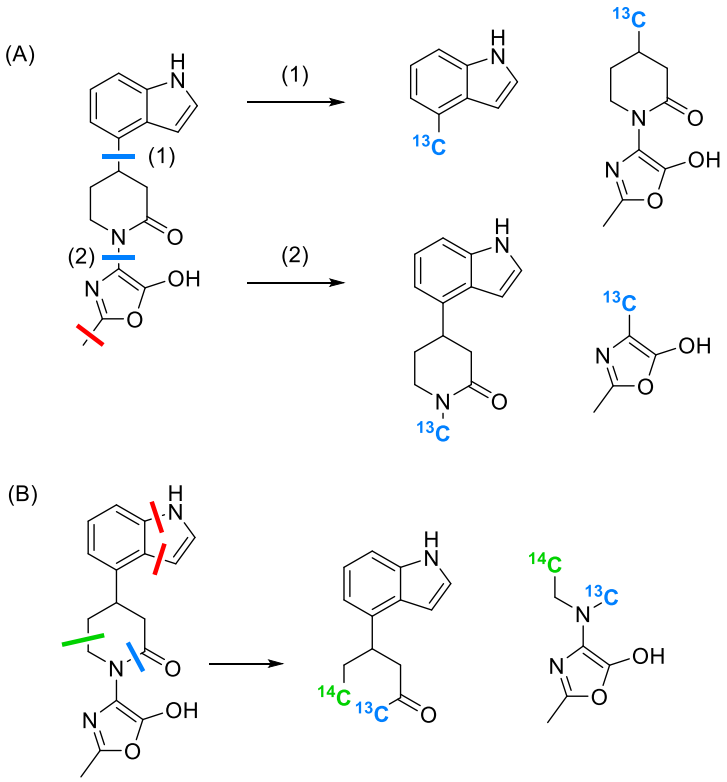


Fig 2. Query fragmentation. (A) Valid cuts (blue) of acyclic bonds in a query molecule to generate two query fragments. (B) One example among many possible ways to split a query molecule on two cyclic bonds (blue and green). Bonds with red bars are not chosen to fragment the query, because cleaving those bonds would result in overly small fragments. Structures are shown in 2D for clarity. In practice, the query is a 3D molecular conformer, and each query fragment retains the conformation of the parent query molecule after the split.

For this paper, we examine two-component reactions in the Enamine REAL Space and thus split the query molecules into two fragments. Future work will extend to 3 and 4 component reactions, which will require splitting the query molecule into 3 and 4 fragments.

Synthon conformation sampling

In order to score synthons against query fragments with ROCS, synthon conformations are required. For each non-ring-forming synthon, we use OMEGA directly to sample its low-energy conformers. For ring-forming synthons, there are two connector atoms, and care must be taken to ensure the conformations enumerated for the synthon are comparable to the ring formed upon product creation. With no constraints, it is likely that the atoms in the partial ring of the synthon will rotate out of the ring plane because of steric clashes. To prevent this, we apply a series of operations in which we first complete the partial ring by adding the minimal set of atoms from a complementary synthon (i.e., atoms that fall on the shortest path between the two connector atoms in a ring-forming synthon). Conformers are then generated for this minimal product. After conformers are generated, the minimal set of atoms that has been added to complete the ring is removed and the connector atoms are returned to their positions in each enumerated pose. This ensures that atoms in the partial ring of synthons are properly oriented so that the synthons can be accurately compared to cyclic query fragments (Fig 3).

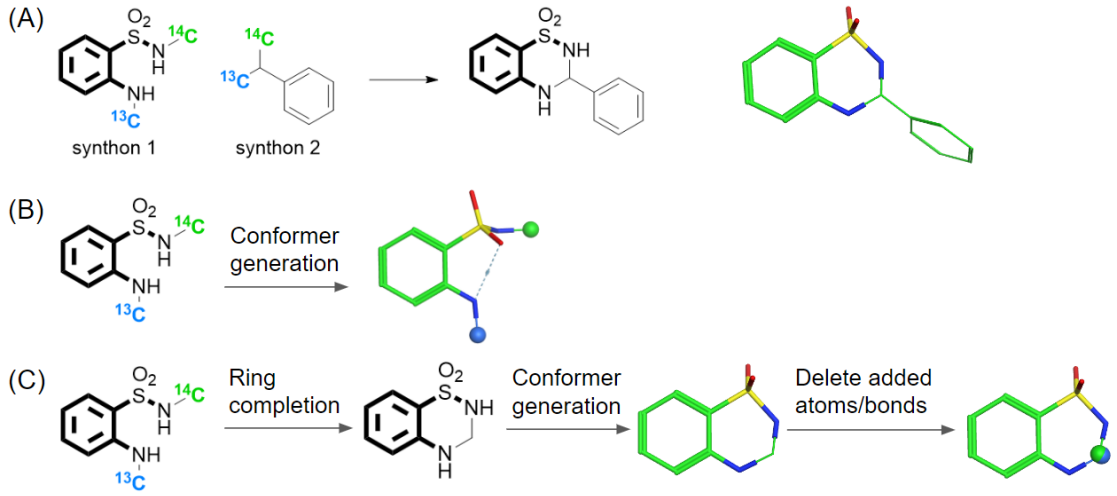


Fig 3. Conformer sampling of ring-forming-synthons. (A) An example of two synthons that form a ring in their product. The low-energy conformer of the product is shown in 3D. (B) Direct conformer sampling of a ring-forming synthon leads to conformers that do not resemble the conformation of the synthon in the product. (C) Workflow that consists of ring completion, conformer sampling, and removal of added atoms and bonds leads to synthon conformers that match the conformation of the synthon in the product.

Synthon scoring

Once synthon conformers have been enumerated, they need to be scored for shape similarity to each of the valid query fragment pairs. Each reaction type in the chemical space combines a set of S1 synthons with a set of compatible S2 synthons, and these must be compared with each pair of query fragments, F1 and F2. Because we do not know which synthon set should match which of the two query fragments, we score both synthon sets against both query fragments (i.e. S1-F1, S1-F2, S2-F1, and S2-F2)

To enable proper scoring of the synthons, not only should the best overlay pose between a synthon and a query fragment have high volume and chemical feature overlap, but the connector atoms on the synthon and the connector atoms on the query fragment must also be in close proximity (Fig 4), to ensure that the orientation of this best overlay pose of the synthon is similar to the pose it adopts in the instantiated product.

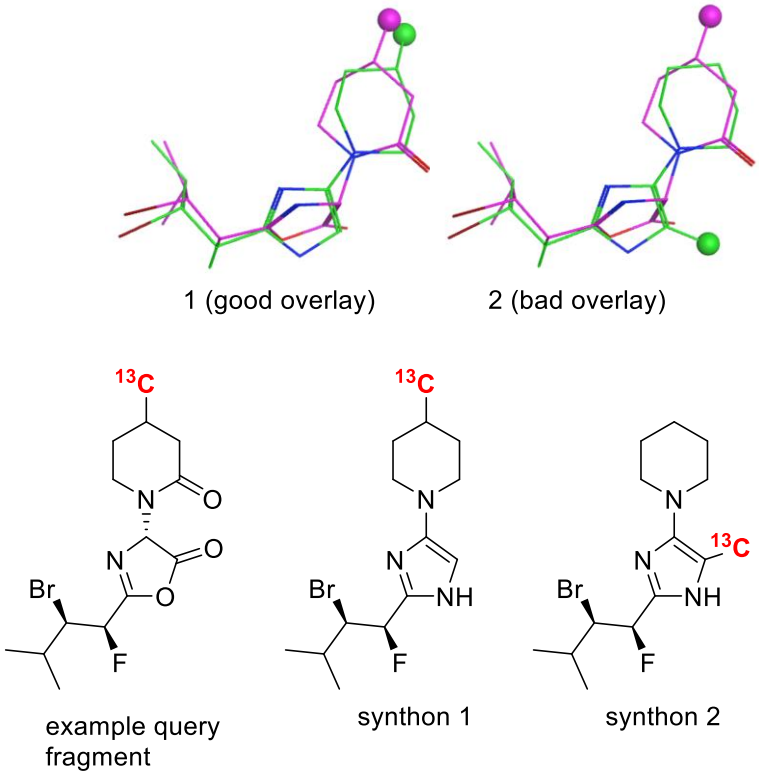


Fig 4. Example of overlays between two synthons 1 and 2 (green) and a query fragment (purple). Connector atoms are shown as spheres. Synthon 1 shows an example of a good overlay: there is high shape similarity and close proximity between the connector atoms of the query fragment and the synthon. Synthon 2 shows an example of a poor overlay: there is high shape similarity but poor proximity between the connector atoms. Products generated from synthon 2 will be unlikely to have good shape similarity to the whole query molecule.

To prioritize synthon poses where the connector atoms on the synthon are in close proximity with the connector atoms on the query fragment, we add custom attractive interactions between the connector atoms to the ROCS color force field. The best type, weight, and radius of such interactions can be determined empirically (see below).

Typically, we score non-ring-forming synthons against acyclic query fragments (i.e. those that were generated by cleaving acyclic bonds in the query molecule), and ring-forming synthons

against cyclic query fragments (i.e., those that were generated by cleaving pairs of cyclic bonds in the query molecule). We can also score non-ring-forming synthons against cyclic query fragments, and *vice versa* (cross-scoring). This requires a small modification because the non-ring-forming synthons (of a two-component reaction) have only one connector atom, while the cyclic query fragment to be scored against has two (and *vice versa*). For cross-scoring, we replace the two connector atoms in the cyclic query fragment with one centroid connector atom, so that during shape comparison, the custom attractive interaction is applied between the connector atom on the acyclic synthons and the connector centroid on the cyclic query fragment.

Product selection

After scoring all synthons against all query fragments, we combine the scores of compatible synthons for all reactions in the chemical space (Fig 5). Because we score each of the query fragments (F1 and F2) against each of the synthons (S1 and S2), we combine the synthon scores from S1-F1 with S2-F2 and S2-F1 with S1-F2. For duplicate synthon combinations, only the top-scoring combination is kept.

Scores can be aggregated by means of a simple average or a weighted average, where each synthon score is weighted by the number of its heavy atoms (to approximate the size of the synthon).

For practical purposes, if a synthon list is very large (e.g. >10000 synthons), we can reduce the number of pairwise combinations and sorting time by taking only a fraction of the top scores from each synthon-fragment score list after combining the scores (S1', S2' in Fig 5, also see SI).

Score combinations from all reactions are concatenated and sorted. Finally, the products from the top-scoring synthon combinations are instantiated, and their conformations are generated and scored against the full (unfragmented) query. The products are then ranked by their scores to give the final product list.

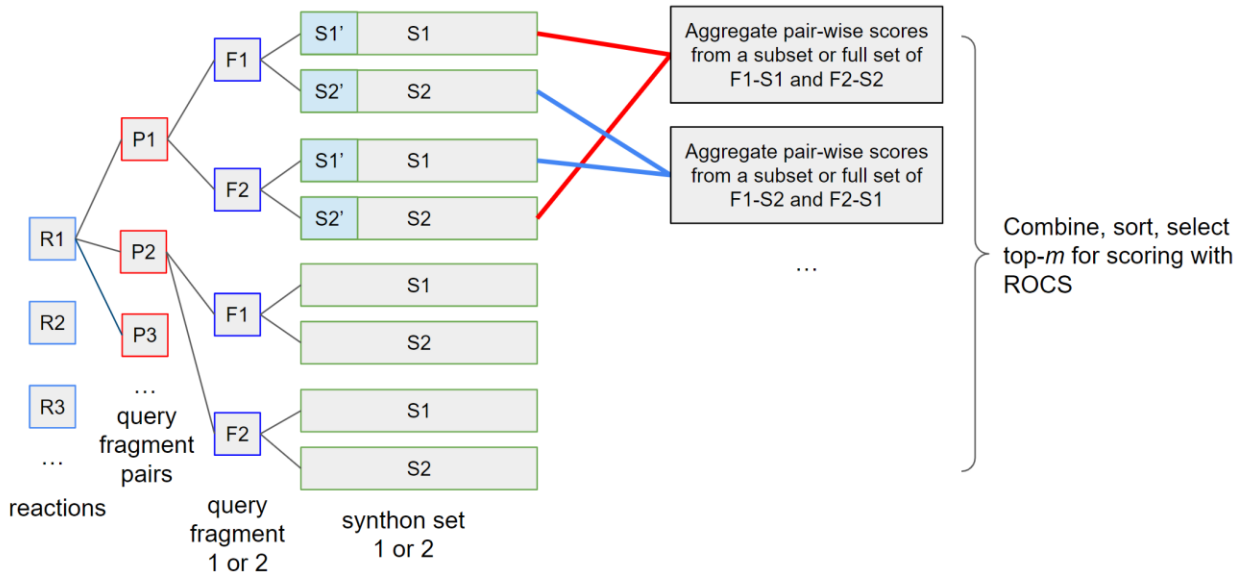


Fig 5. Synthon score combination and product selection. The scores are organized by reactions (R1, R2, …), then query fragment pair (P1, P2, …), then the four sets of scores (F1-S1, F1-S2, F2-S1, F2-S2). During synthon aggregation, either the full sets (e.g. S1, S2) or only subsets (e.g. S1', S2') are pair-wise aggregated.

Evaluation

Because our aim is to reproduce the virtual screening results for the fully enumerated library (ground truth), we evaluate SASS by how well it can recall the top-ranked compounds in the ground truth. For example, if 80 out of the top x = 100 compounds from the results of SASS are in the top 100 of the ground truth, the recall at the top x = 100 results is 80%. Similarly, we can calculate the recall rate at any value of x. To reflect the performance of this method over a range of x values, we calculate the AUC of the recall rate for all values of x up to 1000 (AUC@1000). We choose 1000 because this is a reasonable scale of compounds to be selected and post-processed (manual inspection, clustering, etc.) after a search.

Results

Method optimization

The major parameters of SASS to tune are those influencing the alignment between connector atoms on the synthons and those on the query fragments (i.e., the weight (w) and radius (r) of the interactions between connector atoms that are defined by the custom interactions in the ROCS color force field). To optimize the w/r parameters using a representative set of reactions and synthons, we selected all two-component reactions from the Enamine REAL Space[33] and randomly selected up to 100 synthons from each synthon set from each reaction, which totaled ~40k synthons and 4.2M enumerated products (including all isomers of products).

After defining the chemical space, we selected six molecules (including some well-known drug molecules, Fig S1) as the queries for parameter optimization. For molecules with known crystal structures, their PDB ligand conformations were used. Otherwise, the conformation of the molecules was optimized with energy minimization in MOE. To generate the ground truth ranking for evaluation, conformers were generated for each of the 4.2M products resulting from the procedure above and scored against each query molecule using ROCS.

To run SASS, we fragmented the query molecules as described above and scored the synthons against query fragments. For simplicity, we initially scored ring-forming synthons against only ring query fragments, and non-ring-forming synthons against only non-ring query fragments. After synthon scoring, the top 20k synthon combinations were selected (top-m=20k, which corresponds to 0.5% of the full library size), the corresponding subset of products was instantiated, conformers generated, and products ranked by ROCS scores.

Initially, with the default color force field, running SASS on query 1 gave AUC@1000 of 0.56 (Fig S2). To reward connector atom alignment, we added connector atom interactions to the ROCS color force field. Because some synthons contain two connector atoms while others contain one, the weights of the interactions were normalized, such that the maximum reward from connector atom alignment that every synthon can obtain would be the same, regardless of the number of connector atoms in the synthon. With the default w/r parameters (w = -1, r = 1) for the custom connector atom interaction, the AUC@1000 increased to 0.73.

We examined various values of w and r with the default Gaussian color features[35] for the connector atoms using queries 1-6, and found that w = -10 and r = 10 gave the best results on average (Fig S3, S4). The seemingly large weight likely reflects the importance of connector atom alignment, and using a large weight tends to favor synthons with a slightly worse shape-match but a good connector atom alignment over synthons with better shape-match but worse connector atom alignment. The benefit of a large radius could be twofold: 1) it could allow for the "detection" of connector atom "interaction" when the ROCS initial alignment places connector atoms far apart. During the overlay optimization, the interaction between connector atoms moves them closer; 2) it might not require the connector atoms to be very close to each other to contribute significantly to the color score, and this fuzziness appears to be beneficial. Since the connector bonds can move when forming the product compared to their positions in individual synthons, exact superposition of the connector atoms during overlay of synthons with query fragments is neither needed nor desired.

For aggregating synthon scores, both simple average and weighted averages gave similar AUC results for queries 1-6 (Fig S5), so simple averages were used for ranking the synthon combinations.

Testing with "self-recall"

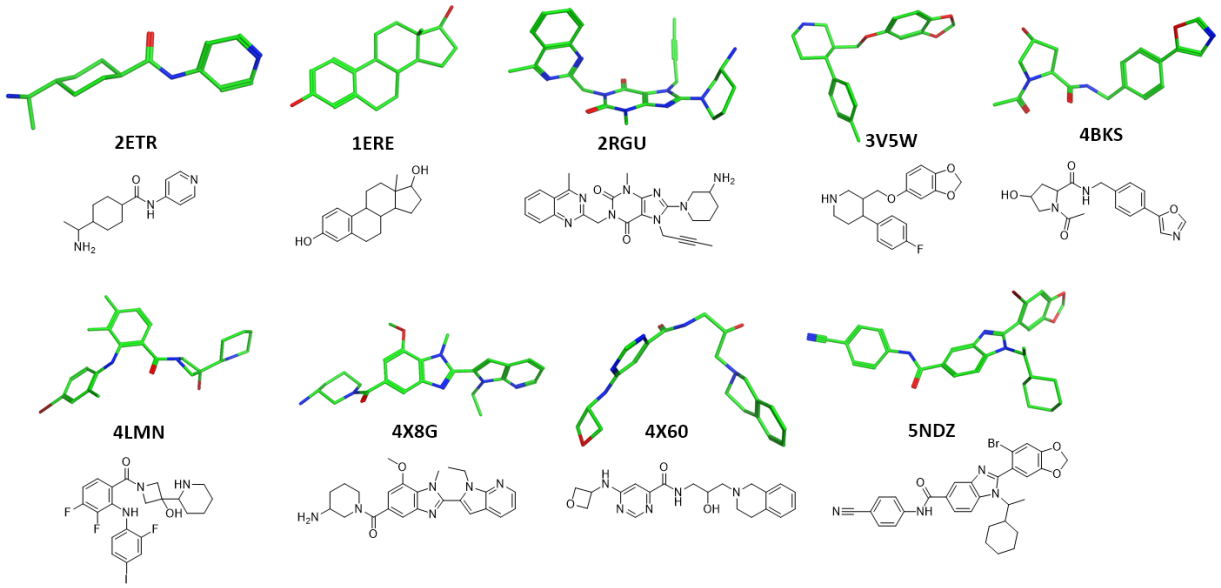
A simple test of the validity of SASS is to verify that a molecule can "find itself": if a low-energy 3D conformer of a synthon-derived product is used as a query, will that same compound be found? Because the synthons that make up these queries are contained within the synthon lists (i.e. positive controls), a good synthon-based search method should rank the aggregated scores of synthons that constitute these queries very highly. We ran SASS using 96 randomly selected products from the full library space (4.2M) as the queries[36] and ranked the full library of compounds by the combined score of each compound's constituent synthons. The query molecules were indeed close to the top of all products: out of 96 queries, 89 synthon combinations were ranked #1, 2 ranked #2, 4 ranked #3, and one ranked #29 (Fig S6), which means that even with a very stringent top-*m* of 100 (number of products selected for re-scoring), all those exact query molecules would have been recalled by this method.

Testing on additional queries

We examined the generalizability of our method by applying it to 9 query molecules from the literature,[8, 14] which are diverse and represent a broad class of molecules and targets. The results are summarized in Fig 6. This method showed strong performance for 6 out of 9 (1ERE, 2ETR, 3V5W, 4BKS, 4X8G, 5NDZ) queries, with AUC@1000 > 0.85 (top-*m* = 20k, 0.5% of fully

enumerated space). The AUC@100 values were even higher, as expected (see Fig S7). In addition, aggregation of the top 1000 ground truth compounds across all queries showed that the recall is the best for the top-ranked compounds, i.e., the method performs the best when it counts the most (Fig 7).

For 2ETR, we examined some top-ranked ground truth compounds, and found several examples of scaffold-hopping (Fig 8), in which the amide bond is reversed, or extended to a urea moiety, but the product is still similar to the original query molecule in overall shape. Gratifyingly, this method was able to recall those compounds (i.e. reproduce the results of a full library search).



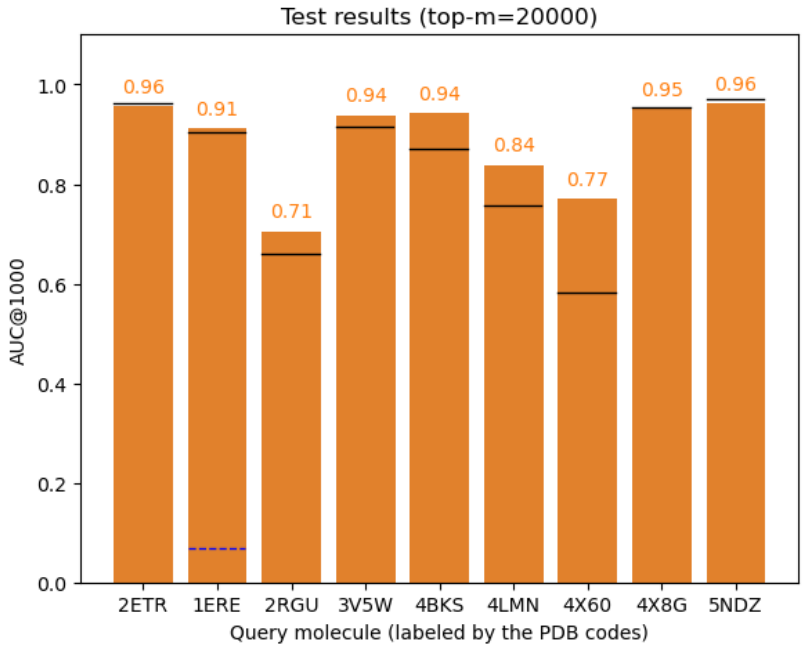


Fig 6. Structures of the query molecules (top) and the search performance (bottom). Orange bars in the graph represent the results when the ground truth is calculated by running ROCS with up to 500 conformers for each product. Black lines represent the results when ground truth is calculated with up to 50 conformers for each product. For 1ERE, the results are with cross-scoring. Without cross-scoring, the AUC@1000 is 0.07 (blue dashed line).

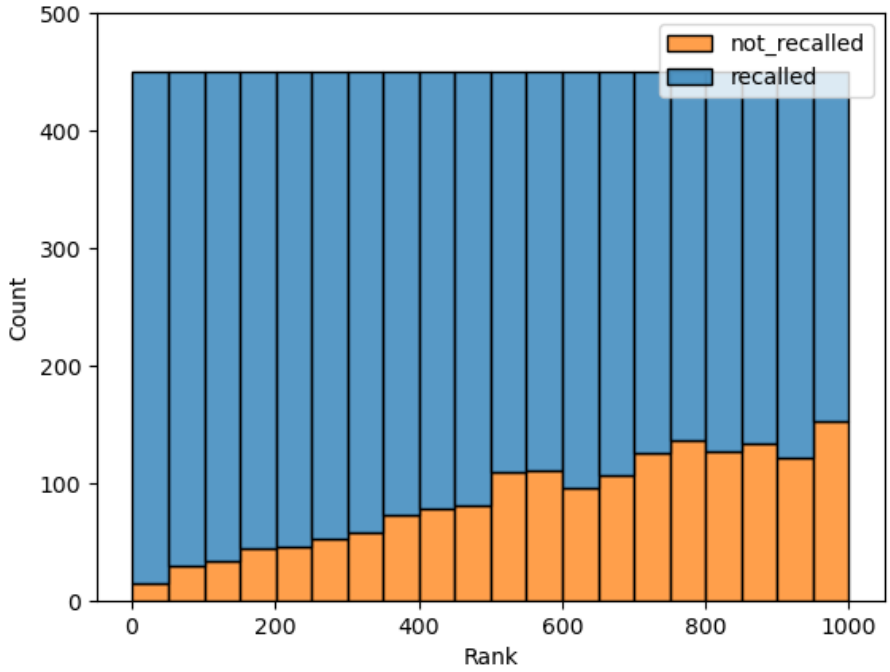


Fig 7. Stacked histograms of the ranks of ground truth molecules across all 9 queries that were recalled and not recalled.

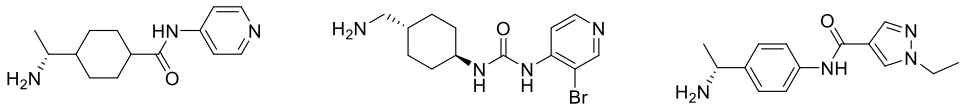


Fig 8. Examples of top-ranked products in ground truth that have different scaffolds compared to the query. SASS ranked these products within top 20k among 4.2M products.

The AUC@1000 values were lower for 2RGU (0.66) and 4X60 (0.58) (Fig 6, black lines), and while those lower values could be attributed to 3D shape similarities not always being additive (i.e. the optimal conformations of synthons in products can be different from their optimal conformations during ROCS overlay with query fragments), we also examined the possibility of insufficient conformer sampling when calculating the ground truth results. By default, OMEGA outputs up to 50 lowest energy conformers for each molecule, and those conformers of the fully enumerated products were used for ROCS scoring to generate the ground truth. When we increased the conformer number to up to 500 (while keeping the same energy window of 10 kcal/mol), scored the full library against the query, and used this new ground truth to compute the recall, the AUC@1000 value improved from 0.58 to 0.77 for 4X60 (Fig 6, orange bars). Improvements of varying degrees were also seen for other queries. This represents an interesting additional utility of the method, i.e. identifying shape-similar compounds based on synthon combinations that would have been missed by the full enumeration approach due to insufficient conformer sampling of the whole products. While we saw that the increased conformer sampling of the products was beneficial, we did not see any improvement of the query performance when the conformer sampling of synthons was increased (Fig S8). This was likely due to the much smaller size and fewer conformers available to the synthons.

The performance on the 1ERE ligand was very poor (AUC@1000=0.07, Fig 6, blue dashed line), but it is a special case where the query generates no valid acyclic query fragments, only ring-fragments. By default, we only scored ring-forming synthons against ring-fragments, and, as a result, any potential top compounds in the ground truth results that arise from combination of non-ring-forming synthons would not be found (Fig S9). In addition, the ratio of ring-forming reactions/synthons to non-ring-forming is ~1:9, so scoring only the ring-forming synthons greatly limited the choice of synthons. To address this issue, we included scores between non-ring-forming synthons and ring-fragments of 1ERE query (cross-scoring) in the selection pool for the final top-*m* combinations. This led to a much improved AUC@1000 of 0.90, demonstrating that this method is robust even for molecules with only fused rings. It should be noted that including cross-scoring did not benefit other queries (Fig S10) while at least doubling the time for the synthon-scoring step, which suggests it may be best to reserve cross-scoring to queries that cannot be split by cleaving only acyclic bonds.

Testing at Scale

To assess whether the high recall persists at scale, we increased the chemical space to include up to 1000 random synthons from each two-component-reaction synthon set. This resulted in a fully enumerated library of 229M products. We ran SASS with 4X60 and 4X8G because those queries represent the low and high ends of the performance on the 4M-scale.

As shown in Fig 9, the performance of this method is maintained at this larger scale. For both queries, the AUC is higher than the 4M-scale experiment at the same fraction of the library scored (black lines). For 4X8G, SASS achieved a recall AUC of 0.94 at top-*m*=20000, which corresponds to less than 0.01% of the full 229M library. Even with top-*m*=2000 (0.001% of the full library), the recall AUC was still 0.60. For 4X60, to achieve the same recall as on the 4M-scale library (where we rescored 0.5% of the full library), we needed to rescore only 0.02% of the full library.

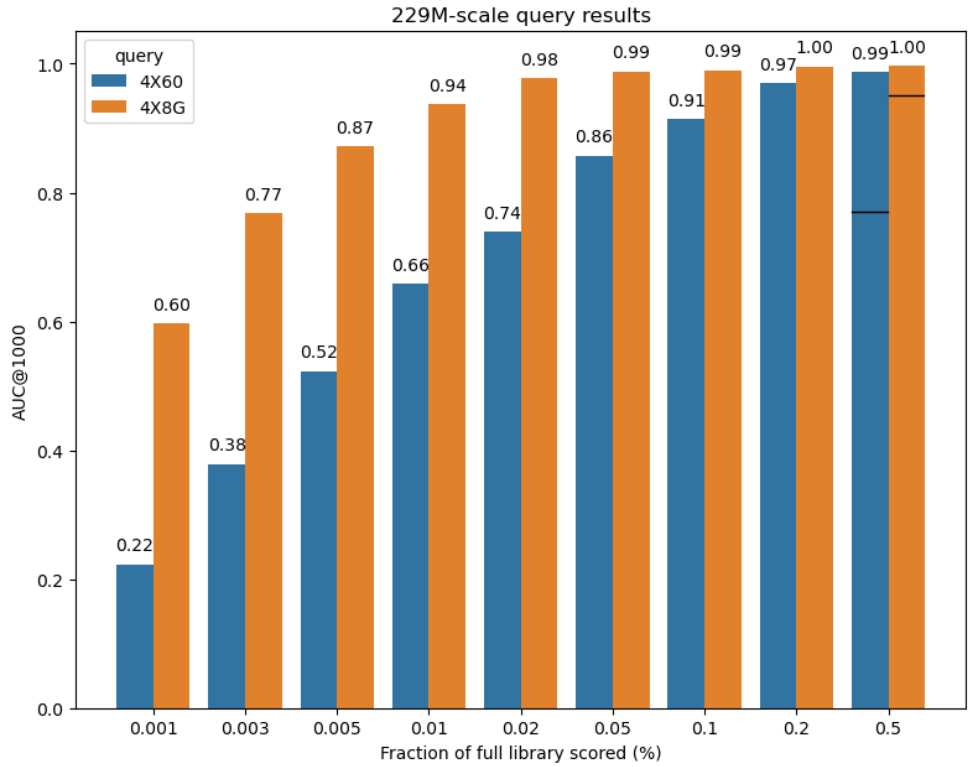


Fig 9. Recall AUC when scoring different fractions of the full library. The fully enumerated space contains 229M molecules. Black lines represent the AUC at the same fraction (0.5%) when querying on the smaller 4M library (Fig 6).

Furthermore, we also ran SASS on the entire two-component reaction space, which totaled 1.9M synthons and 2.8B molecules (estimated to grow to 5B when all stereoisomers are enumerated). Due to the scale, we did not calculate the ground truth. However, a comparison of

the top 1000 scores from the 5B-scale vs the 229M-scale experiment showed further increase in the ROCS scores of the top hits (Fig S11).

Discussion

Effect of connector atom alignment

Fundamental to the success of SASS is being able to correctly score and rank synthons. To take into consideration the alignment between the connector atoms on the synthons and those on the query fragment during synthon scoring, we applied custom color interactions between the connector atoms. To visualize the effect of the custom interaction, we examined some overlay poses between synthons and query fragments. For example, without the custom interaction between connector atoms (Fig 10A), the overlay optimization superimposed the 5,6-ring of the synthon almost perfectly with the 5,6-ring of the query-fragment, leaving the synthon connector atom and the query fragment connector atom pointing in opposite directions. On the other hand, with the custom interaction, the synthon was flipped in the best overlay pose, trading the perfect 5,6-ring overlay for having the connector atoms pointing in the same general direction (Fig 10B). The score from the overlay pose on the right is more accurate for ranking the synthons.

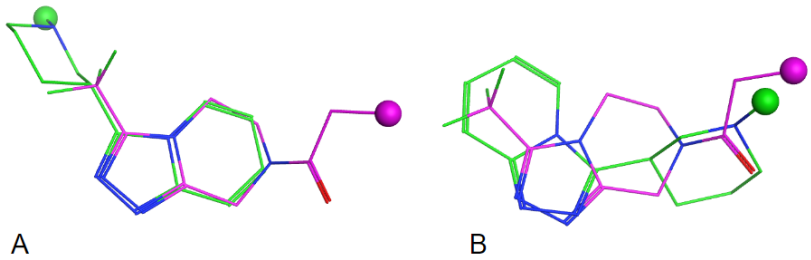


Fig 10: Best overlay pose of a synthon (green) against a query fragment (purple, connector atoms shown as spheres). A: Overlay without custom interaction between connector atoms (synthon shifted by 0.2 Å for visual clarity). B: Overlay with custom interaction between connector atoms.

Time and Space considerations

The compute times for the full-enumeration approach and SASS can be calculated as the sum of each step outlined in Fig 1. First, conformers (products or synthons) need to be generated, a process that can be particularly time consuming for the products of fully enumerated large libraries. For the full-enumeration approach, the ranked product list is yielded directly from the output of the product scoring step. For SASS, top-scoring synthon combinations are instantiated to create a subset of the full library. Conformers for this subset of products are generated and scored to yield the final product list.

Exact compute times can be affected by the hardware, the programming language, and the efficiency of parallelization. As a result, runtimes provided here are an estimate used for comparison of the full-enumeration approach to SASS.

For a library of 229M products, the full-enumeration approach (product instantiation and conformer generation) took 34,000 cpu-hr, and the ROCS query step took 9000 cpu-hr, totaling 43,000 cpu-hr. In contrast, for SASS, synthon conformer generation of the 350k synthons that make this library took 30 cpu-hr. Subsequent synthon scoring, synthon selection, product instantiation (20k), and product scoring took 20 cpu-hr, totaling 50 cpu-hr for SASS. This represents a time saving factor of 800. Because this method scales roughly linearly with respect to the square root of the full library size, we expect an even larger acceleration factor for larger libraries.

In addition to the time consideration, we also examined the space requirements for full-enumeration vs SASS. Storing billions of conformers is not trivial: in our hands, the conformers of 229M products required 6.9 TB disk space when stored in the oez format,[37] and this number scales linearly with respect to the size of the library. In contrast, the total file size of the conformers of 350k synthons is 1.3 GB, representing a space saving factor of 5000. Again, we expect larger efficiency gains for larger libraries.

Other considerations

While the recall AUC of this method is high on most queries, the performance on some queries (e.g. 4X60) is moderate. In addition, it is necessary to score and rank the top-ranked synthon combinations (as opposed to ranking products by the aggregated synthon scores directly), because while the aggregated synthon scores allow enrichment of the top ground truth products in the top-*m* selection, the exact rankings of aggregated synthon scores do not correlate well with the rankings of top ground truth products. These observations are likely due to the fact that the shape similarity of two partial molecules is not always additive (partial solutions for 3D searching are not independent[3]), i.e. the conformations of synthons can change when they are combined into products, due to sterics and/or electrostatic repulsion between the two synthons. Such interactions cannot be predicted when synthons are examined individually. Thus, two synthons that match well to two query fragments may not generate a product that matches well to the shape of the whole query molecule (see SI for examples).

In addition, for this study we limited the chemical space to products generated from two-component reactions. Future work will expand this method to 3- or 4-component reactions, which will cover an even larger chemical space.

Finally, while we used ROCS as the scoring function, other 3D shape-based scoring methods should also be suitable.

Conclusion

We presented Shape-Aware Synthon Search (SASS), a shape similarity-based method for exploring synthon-based chemical spaces by fragmenting queries and scoring synthons against query fragments. This method recalled a large proportion of the top-scoring molecules while consuming significantly less compute and storage resources than an exhaustive search. The efficiencies introduced by SASS provide a new 3D ligand-based method to efficiently search

very large synthon-based chemical spaces that are not amenable to a full enumeration approach.

Acknowledgement

We thank Enamine for access to the reaction and synthon data, Professor Stephen Cox for editorial assistance, Alberto Gobbi, Jeff Blaney, Kriszti Boda, and Hans Purkey for helpful discussions, and OpenEye support staff for help with OpenEye Toolkit usage.

References

[1] Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discov. Today*, **2019**, *24* (5), 1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013.

[2] Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.*, **2022**. https://doi.org/10.1021/acs.jcim.2c00224.

[3] Korn, M.; Ehrt, C.; Ruggiu, F.; Gastreich, M.; Rarey, M. Navigating Large Chemical Spaces in Early-Phase Drug Discovery. *Curr. Opin. Struct. Biol.*, **2023**, *80*, 102578. https://doi.org/10.1016/j.sbi.2023.102578.

[4] Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; et al. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature*, **2020**, *580* (7805), 663–668. https://doi.org/10.1038/s41586-020-2117-z.

[5] Stein, R. M.; Kang, H. J.; McCorvy, J. D.; Glatfelter, G. C.; Jones, A. J.; Che, T.; Slocum, S.; Huang, X.-P.; Savych, O.; Moroz, Y. S.; et al. Virtual Discovery of Melatonin Receptor Ligands to Modulate Circadian Rhythms. *Nature*, **2020**, *579* (7800), 609–614. https://doi.org/10.1038/s41586-020-2027-0.

[6] Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; et al. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature*, **2019**, *566* (7743), 224–229. https://doi.org/10.1038/s41586-019-0917-9.

[7] Gorgulla, C.; Padmanabha Das, K. M.; Leigh, K. E.; Cespugli, M.; Fischer, P. D.; Wang, Z.-F.; Tesseyre, G.; Pandita, S.; Shnapir, A.; Calderaio, A.; et al. A Multi-Pronged Approach Targeting SARS-CoV-2 Proteins Using Ultra-Large Virtual Screening. *iScience*, **2021**, *24* (2), 102021. https://doi.org/10.1016/j.isci.2020.102021.

[8] Grebner, C.; Malmerberg, E.; Shewmaker, A.; Batista, J.; Nicholls, A.; Sadowski,

J. Virtual Screening in the Cloud: How Big Is Big Enough? *J. Chem. Inf. Model.*, **2020**, *60* (9), 4274–4282. https://doi.org/10.1021/acs.jcim.9b00779.

[9] Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X.-P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; et al. Synthon-Based Ligand Discovery in Virtual Libraries of over 11 Billion Compounds. *Nature*, **2022**, *601* (7893), 452–459. https://doi.org/10.1038/s41586-021-04220-9.

[10] Yang, Y.; Yao, K.; Repasky, M. P.; Leswing, K.; Abel, R.; Shoichet, B. K.; Jerome, S. V. Efficient Exploration of Chemical Space with Docking and Deep Learning. *J. Chem. Theory Comput.*, **2021**, *17* (11), 7106–7119. https://doi.org/10.1021/acs.jctc.1c00810.

[11] Kwon, Y.; Park, S.; Lee, J.; Kang, J.; Lee, H. J.; Kim, W. BEAR: A Novel Virtual Screening Method Based on Large-Scale Bioactivity Data. *J. Chem. Inf. Model.*, **2023**, *63* (5), 1429–1437. https://doi.org/10.1021/acs.jcim.2c01300.

[12] Gupta, A.; Zhou, H.-X. Machine Learning-Enabled Pipeline for Large-Scale Virtual Drug Screening. *J. Chem. Inf. Model.*, **2021**, *61* (9), 4236–4244. https://doi.org/10.1021/acs.jcim.1c00710.

[13] Petrović, D.; Scott, J. S.; Bodnarchuk, M. S.; Lorthioir, O.; Boyd, S.; Hughes, G. M.; Lane, J.; Wu, A.; Hargreaves, D.; Robinson, J.; et al. Virtual Screening in the Cloud Identifies Potent and Selective ROS1 Kinase Inhibitors. *J. Chem. Inf. Model.*, **2022**, *62* (16), 3832–3843. https://doi.org/10.1021/acs.jcim.2c00644.

[14] Beroza, P.; Crawford, J. J.; Ganichkin, O.; Gendelev, L.; Harris, S. F.; Klein, R.; Miu, A.; Steinbacher, S.; Klingler, F. M.; Lemmen, C. Chemical Space Docking Enables Large-Scale Structure-Based Virtual Screening to Discover ROCK1 Kinase Inhibitors. *Nat. Commun.*, **2022**, *13* (1). https://doi.org/10.1038/s41467-022-33981-8.

[15] REAL Database - Enamine https://enamine.net/compound-collections/real-compounds/real-database (accessed Sep 29, 2023).

[16] Virtual Screening - Wuxi Biology https://wuxibiology.com/drug-discovery-services/hit-finding-and-screening-services/virtual-screening/ (accessed Sep 29, 2023).

[17] CHEMriya - OTAVA https://www.otavachemicals.com/products/chemriya (accessed Sep 29, 2023).

[18] Freedom Space 3.0 | Chemspace https://chem-space.com/compounds/freedom-space (accessed Sep 29, 2023).

[19] Explore - eMolecules https://www.emolecules.com/explore (accessed Sep 29, 2023).

[20] Neumann, A.; Marrison, L.; Klein, R. Relevance of the Trillion-Sized Chemical Space “EXplore” as a Source for Drug Discovery. *ACS Med. Chem. Lett.*, **2023**, *14* (4), 466–472. https://doi.org/10.1021/acsmedchemlett.3c00021.

[21] Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.*, **2019**, *62* (3), 1116–1124. https://doi.org/10.1021/acs.jmedchem.8b01048.

[22] Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.*, **2020**, *6* (6), 939–949. https://doi.org/10.1021/acscentsci.0c00229.

[23] Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating High-Throughput Virtual Screening through Molecular Pool-Based Active Learning. *Chem. Sci.*, **2021**, *12* (22), 7866–7881. https://doi.org/10.1039/D0SC06805E.

[24] Pedawi, A.; Gniewek, P.; Chang, C.; Anderson, B. M.; Bedem, H. van den. An Efficient Graph Generative Model for Navigating Ultra-Large Combinatorial Synthesis Libraries. In *Advances in Neural Information Processing Systems*; Koyejo S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A. ., Ed.; Curran Associates, Inc.: New Orleans, USA, November 28 - December 9, 2022, 2022; pp 8731–8745. https://doi.org/10.48550/arXiv.2211.04468.

[25] Du, Y.; Fu, T.; Sun, J.; Liu, S. MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design. **2022**, *14* (8), 1–20.

[26] Ehrlich, H. C.; Volkamer, A.; Rarey, M. Searching for Substructures in Fragment Spaces. *J. Chem. Inf. Model.*, **2012**, *52* (12), 3181–3189. https://doi.org/10.1021/ci300283a.

[27] Ehrlich, H. C.; Henzler, A. M.; Rarey, M. Searching for Recursively Defined Generic Chemical Patterns in Nonenumerated Fragment Spaces. *J. Chem. Inf. Model.*, **2013**, *53* (7), 1676–1688. https://doi.org/10.1021/ci400107k.

[28] Schmidt, R.; Klein, R.; Rarey, M. Maximum Common Substructure Searching in Combinatorial Make-on-Demand Compound Spaces. *J. Chem. Inf. Model.*, **2021**. https://doi.org/10.1021/acs.jcim.1c00640.

[29] Bellmann, L.; Penner, P.; Rarey, M. Topological Similarity Search in Large Combinatorial Fragment Spaces. *J. Chem. Inf. Model.*, **2021**, *61* (1), 238–251. https://doi.org/10.1021/acs.jcim.0c00850.

[30] Liphardt, T.; Sander, T. Fast Substructure Search in Combinatorial Library Spaces. *J. Chem. Inf. Model.*, **2023**. https://doi.org/10.1021/acs.jcim.3c00290.

[31] Meyenburg, C.; Dolfus, U.; Briem, H.; Rarey, M. Galileo: Three-Dimensional Searching in Large Combinatorial Fragment Spaces on the Example of Pharmacophores. *J. Comput. Aided. Mol. Des.*, **2023**, *37* (1), 1–16.

https://doi.org/10.1007/s10822-022-00485-y.

[32] ROCS 3.6.0.0. OpenEye, Cadence Molecular Sciences, Santa Fe, NM http://www.eyesopen.com (accessed Sep 29, 2023).

[33] We Use “Shape Similarity” as Shorthand for Combined Shape and Chemical Feature Similarity (Shape and Color in ROCS Parlance).

[34] Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front. Chem.*, **2018**, *6* (JUL), 1–21. https://doi.org/10.3389/fchem.2018.00315.

[35] Color Features https://docs.eyesopen.com/toolkits/python/shapetk/shape_theory.html#color-features (accessed Oct 21, 2023).

[36] We Randomly Selected 100 Isomers, and 96 of Those Generated Valid Conformers.

[37] OEFormat https://docs.eyesopen.com/toolkits/python/oechemtk/OEChemConstants/OEFormat.html#OEChem::OEFormat::OEZ (accessed Sep 29, 2023).

Shape-Aware Synthon Search (SASS) for virtual screening of synthon-based chemical spaces

Chen Cheng[1] and Paul Beroza[1]

[1]Discovery Chemistry, Genentech, South San Francisco, USA

Supplementary Information

Contents

Fig S1....2
Fig S2....3
Fig S3....4
Fig S4....5
Fig S5....6
Fig S6....7
Fig S7....8
Fig S8....9
Fig S9....10
Fig S10....11
Selecting products from subsets of scored synthons....11
Non-additivity of synthon scores....11

Fig S1

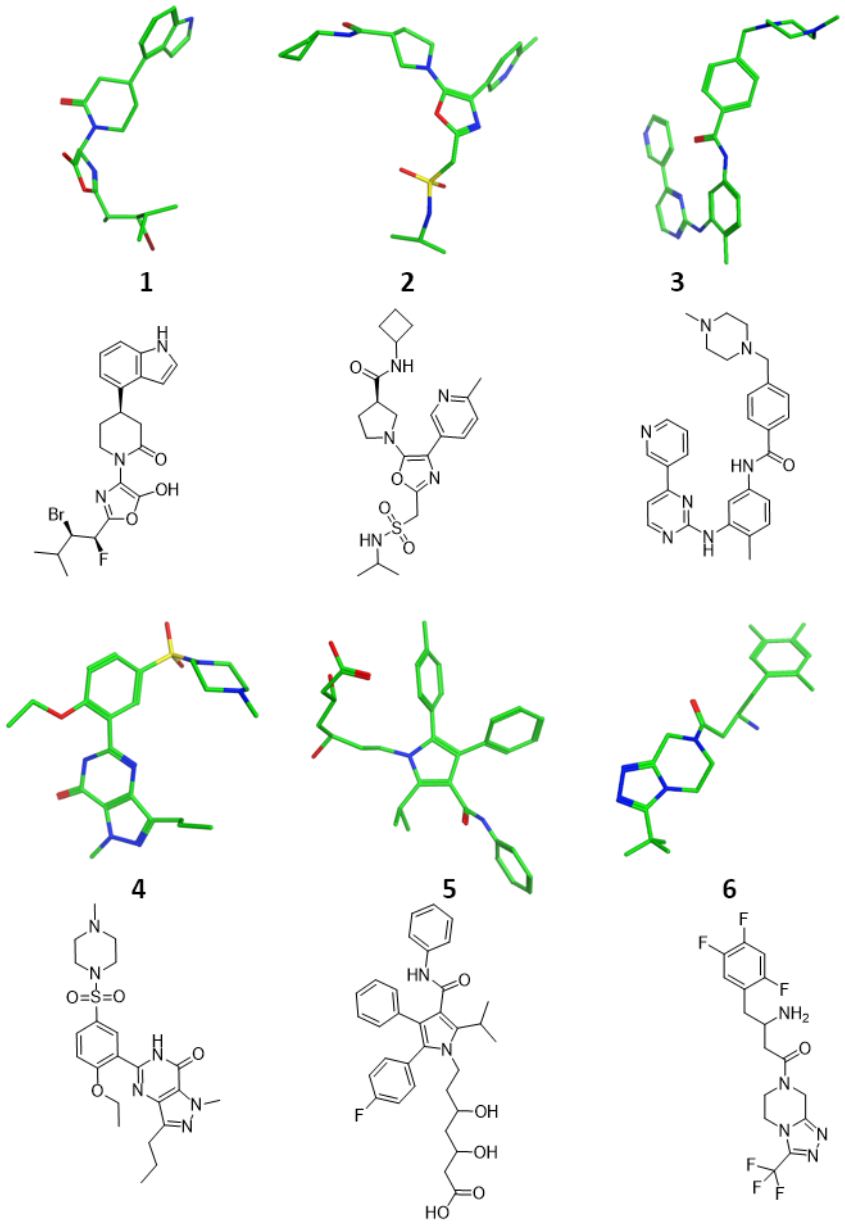


Fig S1. Query molecules 1-6 for parameter optimization shown in 2D and 3D. Molecules 1 and 2 are fictitious. Molecules 3 (1XBB), 4 (1TBF), 5 (1HWK), 6 (4FFW) are known drug molecules taken from their PDB structures.

Fig S2

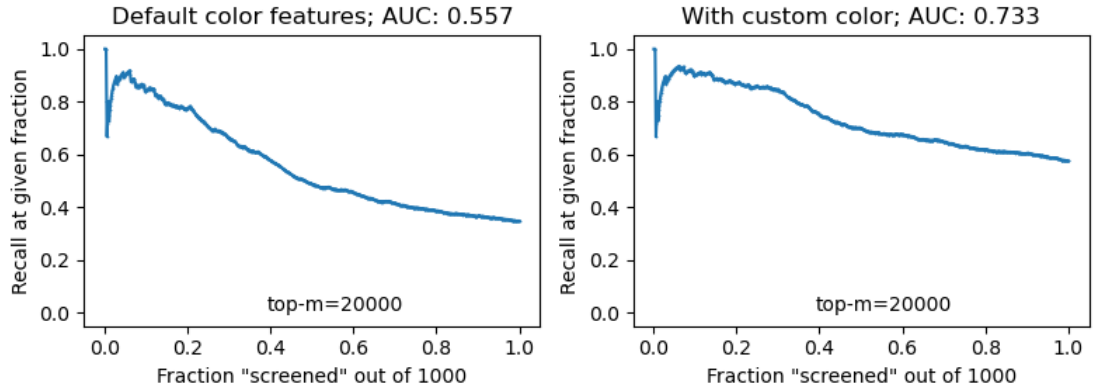


Fig S2. Recall curves for query 1. Left: with default color force field. Right: with custom interactions (w=-1, r=1) between connector atoms added to the color force field to improve connector atom alignment. The chemical space consists of all Enamine Real Space (2020-09) 2-component reactions, up to 100 synthons for each synthon set of each reaction. The fully enumerated library is ~4M compounds. The number of instantiated products scored is 20k (top-*m*=20000).

Fig S3

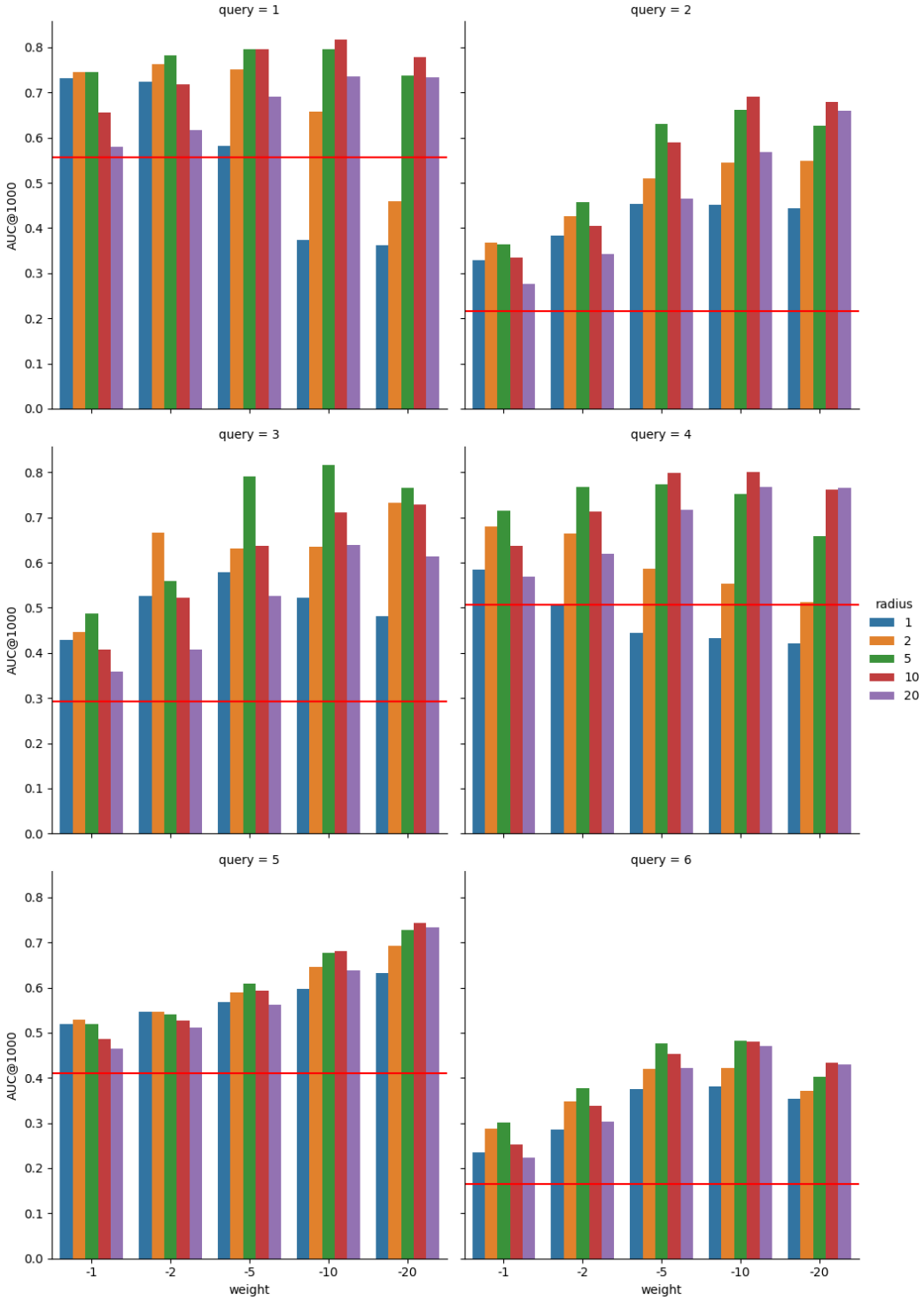


Fig S3. Performance (recall AUC@1000) on queries 1-6 with custom color interaction at various weights and radii for the connector atoms. The red horizontal lines represent the performance without custom interaction. The chemical space is the same as the one described in Fig S2, top-*m*=20000.

Fig S4

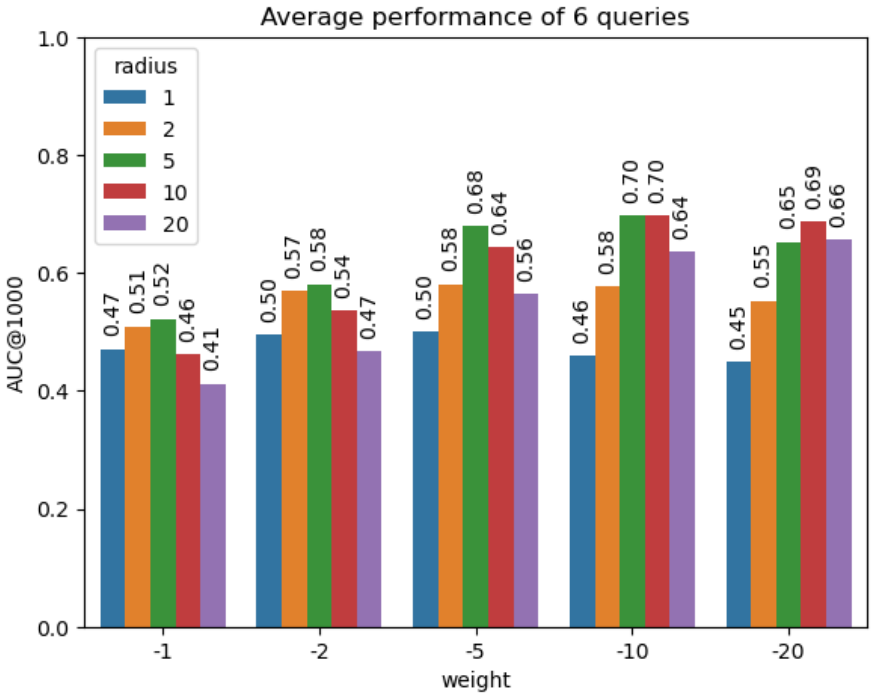


Fig S4. Average performance on queries 1-6 summarized from Fig S3.

Fig S5

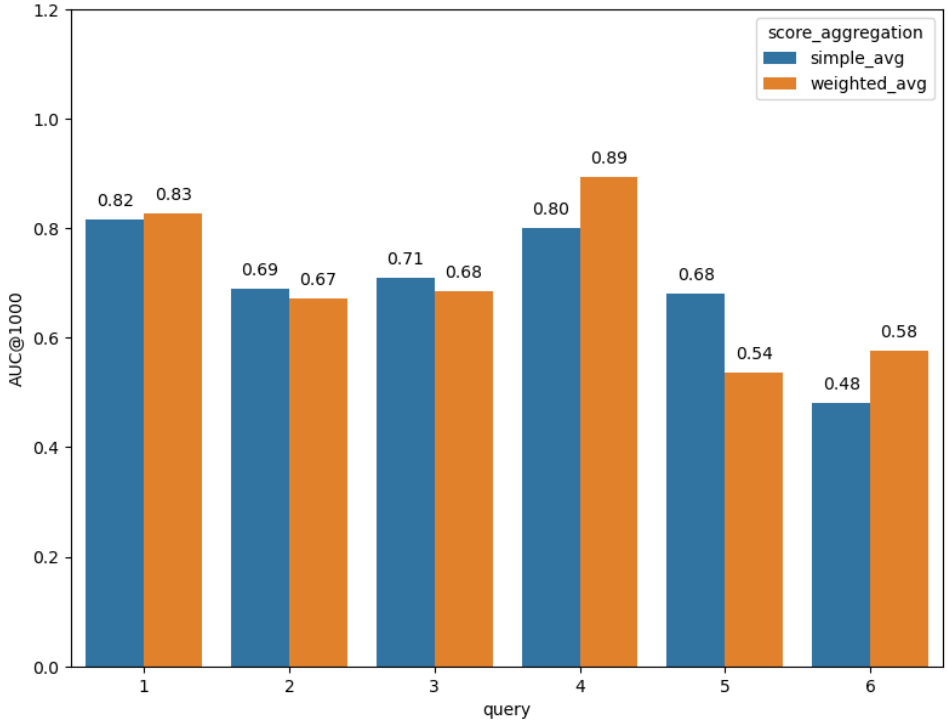


Fig S5. Comparison of performance using different synthon score aggregation method during the product subset selection step. The chemical space is the same as the one described in Fig S2, top-*m*=20000. Custom color interactions with w=-10 and r=10 were used for connector atoms.

Fig S6

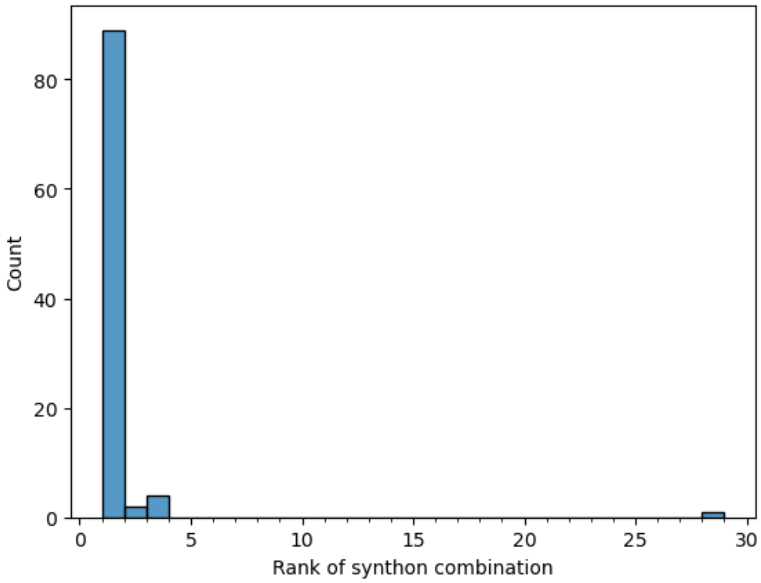


Fig S6. The distribution of the ranks of the product scores (aggregated synthon scores) for the two synthons that make up the query molecules in the “self-recall” experiment. The chemical space is the same as the one described in Fig S2, top-*m* = 20000, w = -10, r = 10.

Fig S7

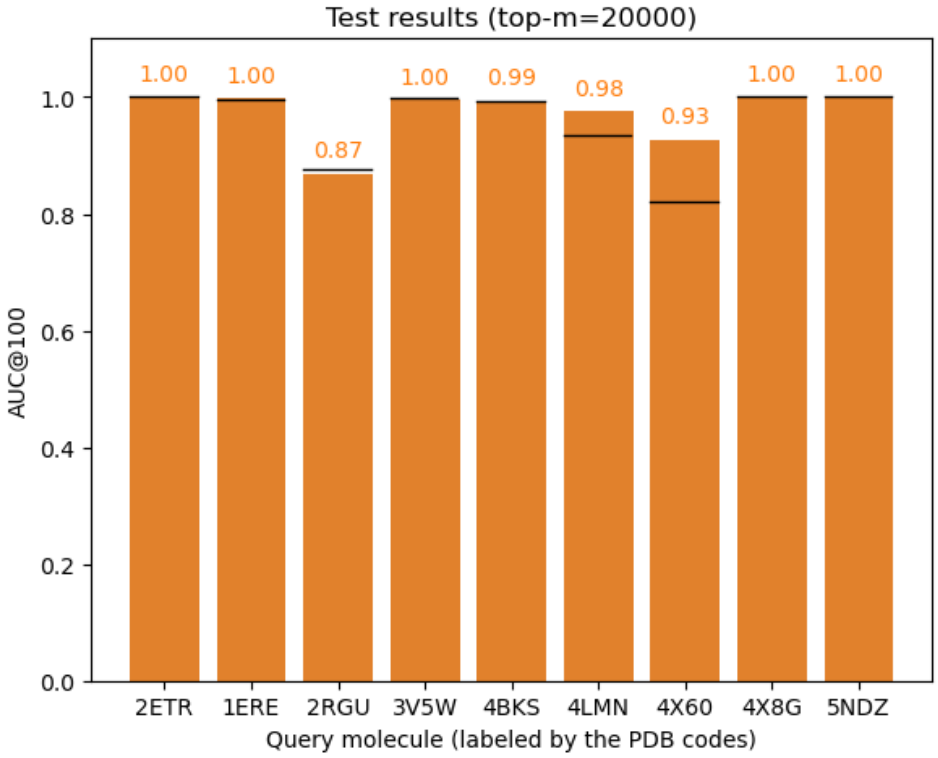


Fig S7. Search results (recall AUC@100) for various test queries.

Fig S8

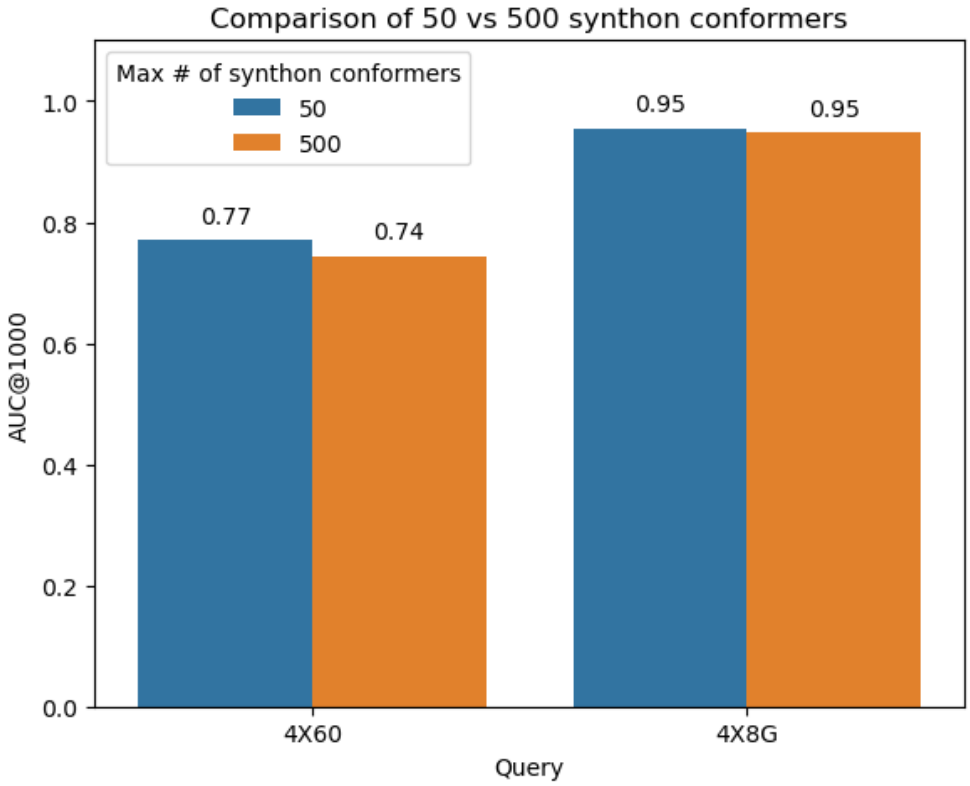


Fig S8. Comparison of using up to 50 vs 500 conformers of the synthons during synthon scoring. Increasing the number of conformers for the synthons did not improve the query performance. The chemical space is the same as the one described in Fig S2, top-*m*=20000, w=-10, r=10.

Fig S9

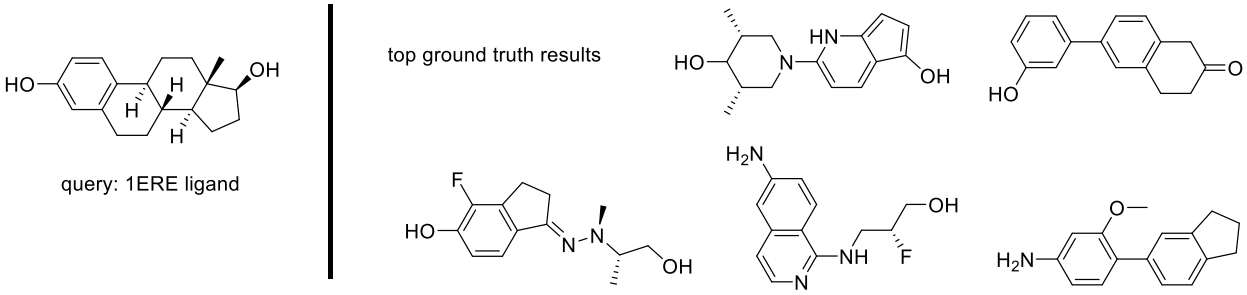


Fig S9. Top ground truth results of query 1ERE ligand. All these products are made of two non-ring-forming synthons joined with an acyclic bond.

Fig S10

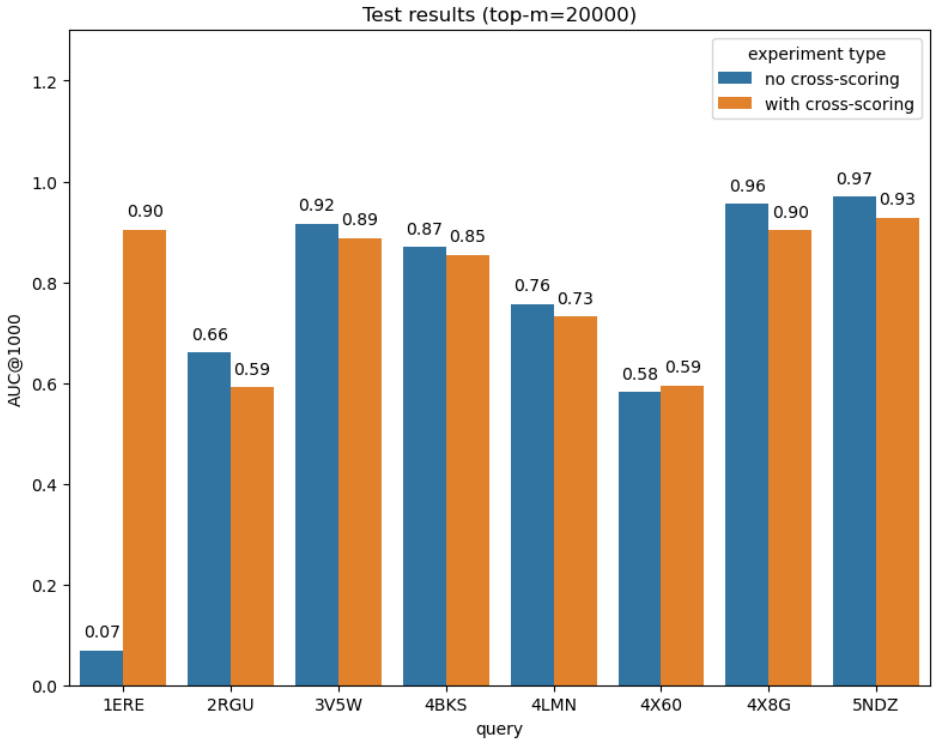


Fig S10. Comparison of query performance with or without cross-scoring. The chemical space is the same as the one described in Fig S2, top-*m*=20000, w=-10, r=10.

Fig S11

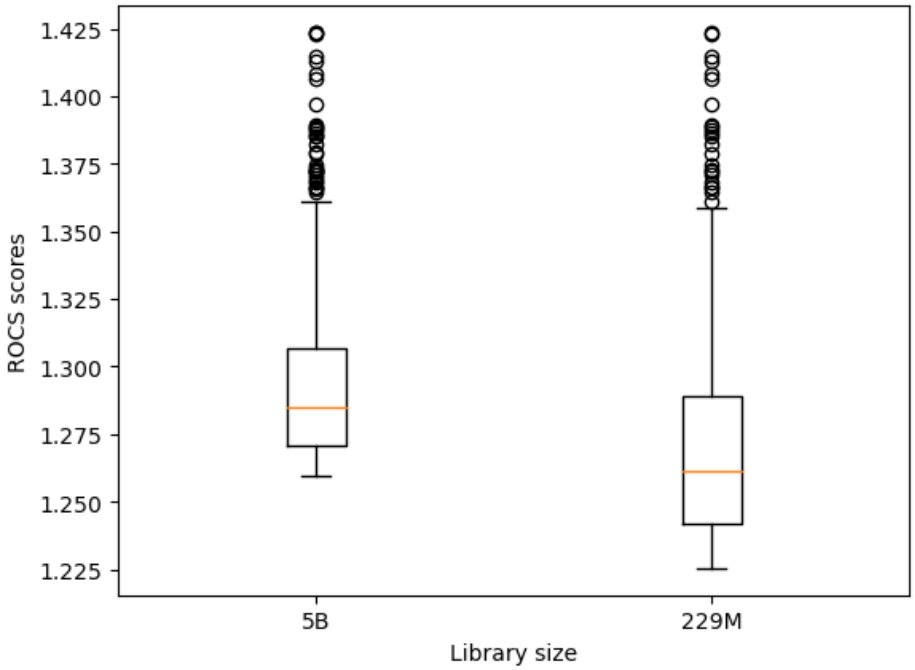


Fig S11. Boxplots showing the distribution of the top 1000 ROCS scores from the results of the experiments with 5 billion and 229 million full library sizes. The score distribution of the 5B experiment is higher than the score distribution of the 229M experiment, indicating that compounds with higher shape similarity to the query molecule were found.

Selecting products from subsets of synthon score lists

During the synthon combination and product selection step (after synthon scoring, before product subset instantiation), we do not need pairwise combination of the entire synthon score lists, but only a subset of each synthon's score list. For example, for query 1, we ran an experiment with the chemical search space limited to one reaction comprising synthon sets S1 (40k synthons) and S2 (10k synthons) with a full library size of 6 M. During product selection step, we either aggregated all synthon scores or only the top 10% of each score list. The results for both runs were very similar (AUC@1000 = 0.61 vs 0.56), suggesting that there's high degree of enrichment of "good" synthons at the top of each synthon score list. This also represents additional efficiency gain during the product selection step as only 1% of the synthon pairs were aggregated and the scores sorted.

Non-additivity of synthon scores

Shown below are some examples where each of the two synthons match well to the query fragments (their combined scores are ranked very highly among all aggregated synthon scores), the rank of the ROCS score of the instantiated product is low among all products.

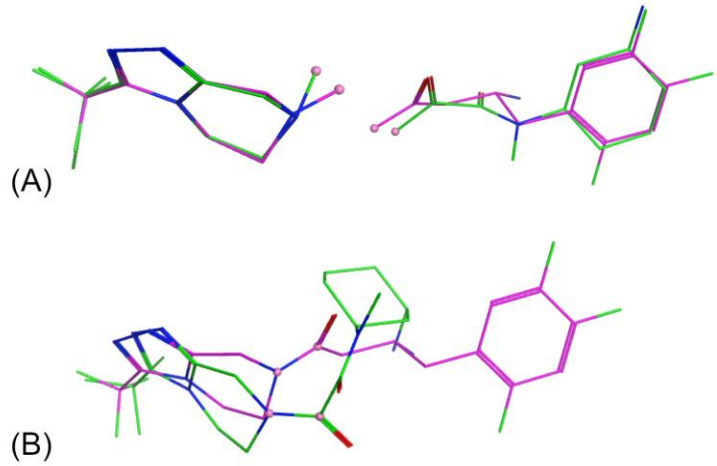


Fig S11. Comparison of the overlays between synthons and query fragments with the overlay between the product and whole query molecule (query 6). (A) Both synthons (green) aligned well with the query fragments (purple), and the aggregates synthon score of this combination ranked #5 among all possible synthon combinations. (B) The product from those two synthons (green) did not overlay well with the whole query (purple). The product ranked > #30000 among all products based on the ROCS score. The connector atoms on synthons/query fragments and the atoms along the bonds formed by the reaction are highlighted.

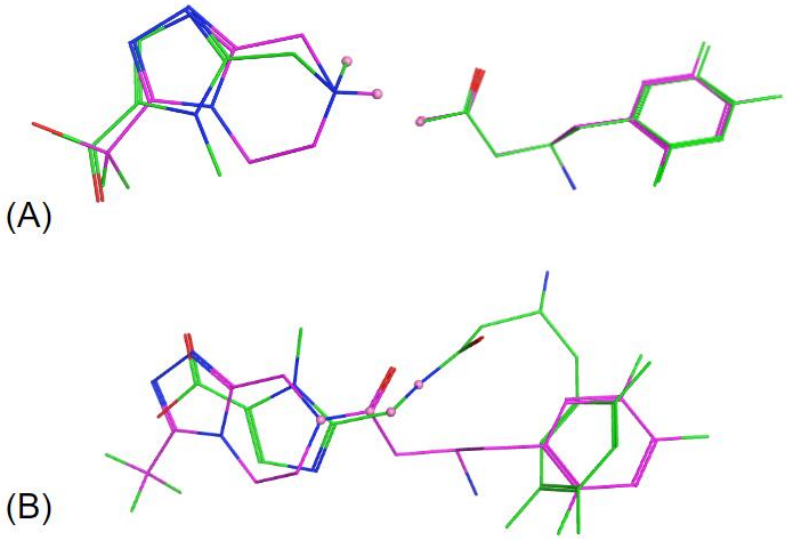


Fig S12. Another example where the overlay between synthons and query fragments are good (A), but the overlay between the instantiated product and the whole query is poor (synthon/product in green, query fragment/whole query in purple).