Published in final edited form as:
J Chem Inf Model. 2014 January 27; 54(1): 5–15. doi:10.1021/ci400264f.

SCISSORS: Practical Considerations

Steven M. Kearnes[†], **Imran S. Haque**[‡], and **Vijay S. Pande**[†]
Department of Structural Biology, Stanford University, Stanford, CA 94305, and Counsyl, 180 Kimball Way, South San Francisco, CA 94080

Vijay S. Pande: pande@stanford.edu

Abstract

Molecular similarity has been effectively applied to many problems in cheminformatics and computational drug discovery, but modern methods can be prohibitively expensive for large-scale applications. The SCISSORS method rapidly approximates measures of pairwise molecular similarity such as ROCS and LINGO Tanimotos, acting as a filter to quickly reduce the size of a problem. We report an in-depth analysis of SCISSORS performance, including a mapping of the SCISSORS error distribution, benchmarking, and investigation of several algorithmic modifications. We show that SCISSORS can accurately predict multiconformer similarity, and suggest a method for estimating optimal SCISSORS parameters in a dataset-specific manner. These results are a useful resource for researchers seeking to incorporate SCISSORS into molecular similarity applications.

Introduction

Calculating similarity between small molecules gives insights into biological activity and provides a basis for prediction of unknown properties. For example, when one or more compounds are known to have activity against a particular target, ligand-based virtual screening (LBVS) can be performed to search a screening database for additional actives using similarity to those compounds.[1] LBVS is an attractive approach to drug discovery because it does not require structural information about the target; successful applications have been reported for diverse targets including enzymes, membrane receptors, and protein–protein interactions.[2,3]

Molecular similarity has been used in many applications besides virtual screening. Shoichet and co-workers described the similarity ensemble approach (SEA)[4] for relating proteins by the similarity of their ligands and identified several novel ligand–target interactions. Posner et al.[5] showed that similarity calculations can be used to reduce false positives in high-throughput screening. Yoon and co-workers combined similarity with docking to streamline

Correspondence to: Vijay S. Pande, pande@stanford.edu.
[†]Department of Structural Biology, Stanford University, Stanford, CA 94305
[‡]Counsyl, 180 Kimball Way, South San Francisco, CA 94080

Supporting Information Available
Supporting Figures and standard deviations for reported averages. This material is available free of charge via the Internet at http://pubs.acs.org/.

multiple-receptor docking campaigns.[6] Similarity also plays a role in methods for consensus structural alignment,[7] screening library construction,[8] and database clustering.[9]

Some applications (*e.g.*, SEA or database clustering) require calculation of a full pairwise similarity matrix. Many similarity methods analyze molecules once and generate a vector representation of each molecule (a "fingerprint") that can be used for repeated pairwise comparisons, and comparisons can often be performed rapidly with simple matrix operations. However, methods such as ROCS[10,11] that require alignments between each molecule pair become impracticable at large scales, even when using GPU-based implementations such as PAPER[12] or FastROCS.[13]

Otherwise intractable similarity calculations can be facilitated by methods that approximate similarity. The SCISSORS method described by Haque and Pande[14,15] uses similarity relative to a fixed basis set to generate vector representations for molecules that can be used for rapid pairwise comparisons. Typical basis sets are on the order of hundreds of molecules, such that speedups relative to explicit similarity calculations become increasingly significant as the size of the problem increases. In LBVS applications, SCISSORS can serve as a preliminary filter to limit the number of molecule pairs that need to be compared by more expensive explicit methods.

This paper presents a detailed analysis of SCISSORS errors and suggests several practical considerations for various applications, with focus on predictions of ROCS shape, color, and combo Tanimotos. We calculate empirical error distributions for SCISSORS predictions and show that SCISSORS prediction quality varies with true similarity. We present benchmarking results that show significant speedups compared to FastROCS in all *vs.* all and one *vs.* all scenarios, and demonstrate that SCISSORS can be used to predict multiconformer ROCS and LINGO Tanimotos. We address several algorithmic modifications and their consequences on SCISSORS performance and conclude with suggestions for practical applications.

Methods

Validation Datasets

We created 100 validation sets by sampling from PubChem3D,[16] which contains three-dimensional conformers for many of the compounds in PubChem.[17] Each subset contained 5000 molecules chosen at random (replacement was allowed between but not within subsets). In cases where downloaded molecules had more than one conformer, only the first conformer was used. Each dataset was subdivided into an ordered "basis molecule pool" (1000 molecules) and a "library" (4000 molecules). SCISSORS basis sets were chosen from the basis molecule pool and predictions were made for all unique non-self pairs in the library (~8 million pairs per dataset).

ROCS

Rapid Overlay of Chemical Structures (ROCS)[10,11] is a 3D similarity method that performs pairwise comparisons of molecular shape and chemical features. Molecular structures are represented as collections of atom-centered Gaussian functions,[18,19] allowing gradient-based

optimization of the overlap between rigid conformers of the “query” (or “reference”) and “fit” molecules. The optimized overlap volume is used for comparison of molecular shape.

The ROCS color force field measures approximate electrostatic similarity by placing “color atoms” at positions that match specific chemical groups and functionalities, including hydrogen bond donors and acceptors, charged atoms, rings, and hydrophobic regions. By default, color atoms have small effective radii (1 Å) and must overlap with another color atom of the same type to contribute to the optimized color overlap volume calculated for a molecule pair. When either molecule in a pair has no color atoms, or when they do not have any color atoms of the same type, the color Tanimoto for that pair will be zero (in contrast to shape Tanimotos, which are never zero).

ROCS shape and color Tanimotos are defined in terms of self overlap and optimized pairwise overlap volumes (note that the self overlap volume is equivalent to the molecular volume):

$$T(a,b) \equiv \frac{O_{ab}}{O_{aa}+O_{bb}-O_{ab}} \quad (1)$$

The ROCS combo Tanimoto used in this report is the average of shape and color Tanimotos (not the sum, as is the convention for ROCS).

Overlays were performed with FastROCS[13] (version 1.3.1) on an NVIDIA GeForce GT 545 GPU. Due to the grid-based nature of the FastROCS algorithm, some Tanimotos > 1 were observed, but these were constrained to unity. Molecule self overlap volumes were calculated using the OpenEye Shape TK.[20]

SCISSORS

The SCISSORS method[14,15] generates vector representations for molecules by least-squares embedding in the feature space of a molecular similarity kernel. These vector representations are then used to calculate pairwise similarity using a vector formulation of the Tanimoto coefficient:

$$T(\mathbf{a},\mathbf{b}) \equiv \frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{a}\cdot\mathbf{a}+\mathbf{b}\cdot\mathbf{b}-\mathbf{a}\cdot\mathbf{b}} \quad (2)$$

By comparison of equations (1) and (2), ROCS overlap volumes can be interpreted as inner products between vector representations of molecules: $O_{ab} = \mathbf{a}\cdot\mathbf{b}$. This representation allows molecule vectors to be approximated using kernel PCA. We define a “molecular” kernel function in terms of these inner products:

$$\kappa(a,b) \equiv \mathbf{a}\cdot\mathbf{b} = \frac{T(a,b)\,(\mathbf{a}\cdot\mathbf{a}+\mathbf{b}\cdot\mathbf{b})}{1+T(a,b)} \quad (3)$$

Next, we construct a kernel matrix containing kernel values between molecules in a preselected basis set $\{b_1, \ldots, b_n\}$:

$$K_{ij}=\kappa(b_i, b_j)$$

Eigen decomposition of *K* gives a feature space basis as rows of a matrix *B*:

$$K=BB^T=VD^{\frac{1}{2}}D^{\frac{1}{2}}V^T$$
$$\therefore B=VD^{\frac{1}{2}}$$

Finally, the feature space embedding ("SCISSORS vector") for a molecule *m* is calculated using a two-step process:

1. Construct a vector **m** of kernel values between *m* and the basis set:

$$\mathbf{m}=[\kappa(m, b_1), \ldots, \kappa(m, b_n)]^T$$

2. Calculate the SCISSORS vector **x** for *m* by least squares:

$$B\mathbf{x}=\mathbf{m} \Rightarrow \mathbf{x}=B^{-1}\mathbf{m}$$
$$=\left(VD^{\frac{1}{2}}\right)^{-1}\mathbf{m}$$
$$=D^{-\frac{1}{2}}V^T\mathbf{m}$$

The matrix $P=D^{-\frac{1}{2}}V^T$ used in step 2 can be calculated once and used for any number of embeddings. The dimensionality of the resulting vectors can be reduced by selecting a subset of eigenvectors to use when calculating *P*. In cases where the choice of kernel results in an indefinite kernel matrix, the maximum SCISSORS vector dimensionality is limited to the number of positive eigenvalues in *D*.

The original SCISSORS method[14] calculated "inner products" between molecules ($\mathbf{a} \cdot \mathbf{b}$) directly from Tanimotos under the assumption of unity self overlap values ($\mathbf{a} \cdot \mathbf{a} = \mathbf{b} \cdot \mathbf{b} = 1$). In this situation, the kernel function given in equation (3) can be rewritten as:

$$\kappa'(a, b) \equiv \frac{2T(a, b)}{1+T(a, b)} \quad (4)$$

This modified kernel is used throughout this report, except where indicated.

For indirect prediction of combo Tanimotos, shape and color Tanimotos predicted using the same basis set and dimensionality were averaged. No indirect combo Tanimotos were predicted using dimensionality values where only one of shape or color Tanimoto predictions were possible.

LINGO

LINGO[21] is a fingerprint-like similarity measure that compares molecules by matching substrings of textual molecular representations such as SMILES. Canonical isomeric SMILES representations of validation set molecules were generated with the OpenEye OE Chem TK,[22] and LINGO Tanimotos for all molecule pairs were calculated with the CPU implementation of SIML[23] (pySIML 1.5).

LINGO self overlap values were defined as the number of LINGO substrings counted for each molecule, calculated by summing rows of the "count matrix" returned by the SIML preprocessing method cSMILES to Matrices.

Validation Protocol

SCISSORS predictions were validated using all available dimensions for 18 basis set sizes ranging from 1–1000 molecules. For each dataset, a basis set of size n was constructed from the first n molecules in the ordered basis pool. Average root-mean-square error (RMSE) or mean error (ME) of predictions of self and non-self overlap values and Tanimotos across all validation sets were calculated at each basis set size and dimensionality. The maximum number of dimensions for each basis set size varied between datasets, so some averages were calculated using fewer than 100 values. Standard deviations for averages are given in the Supporting Information.

When analyzing SCISSORS performance as a function of similarity, predictions were made for an entire dataset and performance metrics were calculated for predictions of pairs whose FastROCS or SCISSORS Tanimoto similarity was within a given window. The windows are left-open, right-closed intervals, i.e. $a < T_{\text{true}} \leq b$. Twenty windows between $T_{\text{true}} = 0$ and $T_{\text{true}} = 1$ were used for this report, each having width 0.05; color Tanimoto plots have one additional window for $T_{\text{true}} = 0$.

Benchmarking

Benchmarking was conducted on a Dell Power Edge R720 running Cent OS 6.3 with the following hardware configuration: 2×Intel E5-2650 CPU, NVIDIA Tesla M2070-Q GPU, 8 × 8 GB 1600 MHz DIMM. SCISSORS calculations were performed with the Enthought Python Distribution (EPD) version 7.3-2, using functions in the numpy Python package (version 1.7.1). Numpy was linked against the Intel Math Kernel Library (MKL); results may vary when using packages without MKL support. Our benchmarking script used FastROCS to calculate shape Tanimotos, and assumed single-conformer molecules, *i.e.*, scores were saved without comparison to the current (possibly zero) value in the result container array.

Miscellaneous

Coefficients of determination (R^2) were calculated using the r2_score function in the scikit-learn Python package (version 0.14) at dimensions that were represented in both basis–library (BL) and library–library (LL) results. Figures were made using matplotlib 1.2.0. The table of contents figure also used Inkscape 0.48. Main text Figures were converted from PNG to TIFF using the convert tool in ImageMagick 6.6.9-7.

Results and Discussion

We randomly sampled from PubChem3D[16] to create 100 validation sets, each containing 4000 molecules (~8 million non-trivial molecule pairs; see Methods). The distribution of ROCS and LINGO Tanimotos in the validation sets is shown in Figure 1. Shape Tanimotos have a broader distribution than color Tanimotos, which are concentrated at small values. The LINGO Tanimoto distribution is less smooth than any of the ROCS Tanimoto distributions, and closely matches the ROCS color Tanimoto distribution.

To evaluate SCISSORS performance, we calculated average root-mean-square (RMS) errors for Tanimoto predictions globally (across all molecule pairs) and in two local contexts, with molecule pairs binned either by FastROCS or SCISSORS similarity. Binning by FastROCS ("true") similarity shows the dependence of prediction accuracy on true similarity, while binning by SCISSORS prediction provides confidence estimates for predictions in a given range. Standard deviations for averages are given in the Supporting Information.

Prediction accuracy is sensitive to SCISSORS parameters

We calculated average RMS errors for ROCS shape and color Tanimoto predictions using all available choices of dimensionality across a series of basis set sizes (Figure 2). Shape predictions had a lower maximum dimensionality than color predictions, and were most accurate near the center of the available dimensionality range, while color predictions tended to improve with increasing dimensionality. Predictions that used all available dimensions were generally poor. Haque[24] noted that small values are more difficult to approximate at low rank due to vector space considerations, which suggests an explanation for the observation that optimal color Tanimoto predictions tended to require more dimensions than optimal shape Tanimoto predictions. Increasing the basis set size appeared to "stretch out" the error distribution for both shape and color Tanimotos, widening the range of acceptable dimensionality.

Local error analysis with binning by FastROCS Tanimoto (Figure 3) showed that SCISSORS performance depends significantly on true similarity, such that predictions of ROCS Tanimotos made with a given dimensionality were likely to have variable accuracy across the true Tanimoto range. Shape Tanimoto predictions showed a band of acceptable dimensionality within each bin, but color predictions deteriorated as similarity increased, such that predictions for similarities above ~0.4 were afflicted with significant error at most choices of dimensionality. Part of this effect could be due to the concentrated color Tanimoto distribution observed in Figure 1, but there does not appear to be any corresponding effect for shape Tanimoto errors. We observed that very few dimensions were required for accurate prediction of ROCS shape and color Tanimotos for very high similarity pairs, and that errors for these predictions tended to increase with increasing dimensionality. This is surprising, since it would be expected that very similar molecules would have very similar SCISSORS vectors, and that good accuracy would be observed for all choices of dimensionality. However, the approximate and indirect nature of the SCISSORS algorithm could lead to high-dimensional differences in SCISSORS vectors for highly similar molecules that at least partially explain the observed error distributions.

We also performed a local error analysis with binning by SCISSORS shape or color Tanimotos (Figure 4). The shape Tanimoto error distribution revealed a band of dimensionality containing high-confidence predictions across the entire predicted similarity range. In contrast, predicted color Tanimotos appeared robust only at low similarity; at color Tanimoto values greater than about 0.4, virtually all Tanimoto predictions contained substantial error. Interestingly, high similarity predictions are made with good confidence at most choices of dimensionality.

Our two local analyses suggest different optimal dimensionalities for color Tanimoto predictions in the most populated similarity range: FastROCS partitioning indicated that about 200 dimensions was best, but predictions made with ~350 dimensions appeared to have the highest confidence. This discrepancy could result from molecule pairs in a similarity window being “contaminated” by pairs from other windows. For example, if FastROCS-partitioned local analysis shows good performance for molecule pairs with color Tanimotos in the range [0.2,0.25], that result provides no information about how many pairs were *predicted* to be in that range. SCISSORS might do very well at predicting pairs with true values in a given range, but could very well achieve that result by predicting *all* pairs to be in that range. Binning molecule pairs by SCISSORS Tanimoto measures the magnitude of this misassignment problem, or the accuracy of predictions within a given range. This is reminiscent of the trade-off between false positive and false negative rates in classification; using parameters chosen from FastROCS-based analysis will improve the recovery of molecule pairs in a given range (reducing false negatives), but choosing dimensionality based on SCISSORS predictions will improve the reliability of predictions in that range (reducing false positives). Later in this report we suggest an approach for generating approximate error distributions of both types in a dataset-specific manner.

Our local error analyses reveal important variations in SCISSORS performance that are not captured by global metrics. Thus, we emphasize that global RMS error is a poor metric for applications that focus on molecule pairs in sparsely populated similarity regimes, especially very high similarity. The choice of SCISSORS parameters should take into consideration the errors observed in local analyses at levels of similarity most important for a particular application. As an alternative metric, we calculated average mean errors (ME) for shape and color Tanimoto predictions. Global analysis showed that similarity tended to be overestimated by SCISSORS at low dimensions and underestimated at high dimensions, while local analysis revealed that Tanimoto predictions skewed toward lower values as true similarity increased (Supporting Figures 1 and 2).

Unity self overlap values are a valid assumption

The original description of SCISSORS[14] includes a parsimonious assumption of unity self overlap (molecular) volumes in the calculation of inner products between molecules, such that similarity predictions are agnostic to differences in self overlap values between molecules. This assumption could have a significant effect on SCISSORS prediction accuracy, since the molecules in our validation sets have broad self overlap distributions (Supporting Figure 3) and the maximum pairwise overlap value is limited to the minimum self overlap value in a molecule pair. We compared SCISSORS predictions made with and

without the assumption of unity self overlap values and calculated average changes in RMS error for shape and color Tanimoto predictions (Supporting Figures 4–6). Most differences were quite small, with the exception of high-similarity errors for color Tanimoto predictions. Confidence for these predictions was noticeably improved without the assumption of unity self overlap values—this region of our average error distribution is highly variable, but we expect that the distribution for a specific dataset would be more smooth (in fact, predictions in this region are more variable without the assumption of unity self overlap values; compare the standard deviations in Supporting Figures 38b, 39b and 49). However, this region of the local error distribution binned by FastROCS Tanimoto showed slightly worse performance. One possible explanation for this effect is that SCISSORS makes more high-similarity color Tanimoto predictions when assuming unity self overlap values. The false positive–false negative trade-off discussed above would become relevant in this situation, and specific applications may have different requirements. Supporting Figure 7 shows correlations plots for color Tanimoto predictions at several choices of dimensionality with and without the assumption of unity self overlap values; these plots are extremely similar. We conclude that the assumption of unity self overlap values is valid for most predictions, and use it throughout the remainder of this report. Interestingly, this result suggests that Tanimoto similarity does not generally reflect differences in molecular volume.

ROCS combo Tanimotos can be predicted directly or indirectly

Figure 5 shows RMS errors for combo Tanimoto predictions made *directly* by using combo Tanimotos as input into equation (4) or *indirectly* by averaging separate shape and color Tanimoto predictions. (It is not clear how the inner products (overlap values) in equation (3) would be calculated for the prediction of combo Tanimotos, so direct predictions can only be made under the assumption of unity self overlap values.) The error distributions for direct and indirect combo Tanimoto predictions are quite different: direct combo Tanimoto predictions had the highest maximum dimensionality of all ROCS Tanimoto predictions, and optimal direct combo Tanimoto predictions required higher dimensionality than either shape or color Tanimoto predictions.

Local analyses with binning by FastROCS Tanimoto (Figure 6) showed that direct predictions had increased error at low dimensionality for mid-range similarities. (Note that we have scaled combo Tanimotos to the range [0,1] instead of using the traditional range [0,2].) Both direct and indirect predictions were less sensitive to similarity than color Tanimoto predictions. Direct combo predictions did not have a single band of dimensionality with consistent error for all similarity windows, but indirect combo Tanimotos achieved good accuracy across all similarity values using ~125 dimensions. Local analysis with binning by predicted similarity (Figure 7) further emphasized the different dimensionality requirements of the indirect and direct approaches: direct predictions had good confidence for predicted Tanimotos up to ~0.6 when ~450 dimensions were used, and indirect combo predictions showed good confidence for Tanimotos up to ~0.7 using ~250 dimensions.

Haque has suggested[24] that direct combo Tanimoto prediction implicitly improves color Tanimoto predictions, but our results suggest that optimal combo Tanimoto predictions

would come from combining shape and color Tanimoto predictions made with different basis sets or and/or dimensionality. In practice, since shape and color Tanimotos are calculated simultaneously by ROCS, handling them separately should not significantly impact SCISSORS performance relative to FastROCS (see benchmarking results below).

As an alternative metric to RMS errors (see above), average mean error (ME) distributions for direct and indirect combo Tanimoto predictions are given in Supporting Figures 8 and 9, respectively.

SCISSORS is *fast*

To give quantitative description of the speedups made possible by SCISSORS, we benchmarked the direct prediction of ROCS shape Tanimotos using several different basis set sizes. We measured performance in two application contexts: all *vs.* all comparisons within a 4000 molecule dataset (Table 1), and one *vs.* all comparisons between a query molecule and a 40 000 molecule screening library (Table 2). For calculation of SCISSORS Tanimotos, we used all available SCISSORS vector dimensions. In the one *vs.* all context, vectors representing a large library are precomputed at a substantial upfront cost (Supporting Table 1), but they can then be stored and used for unlimited queries; the results in Table 2 assume that SCISSORS library vectors have been precomputed.

Our benchmarking results make it clear that ROCS overlays are the rate-limiting step in all *vs.* all SCISSORS calculations. For example, it might be surprising that using a 1000 molecule basis set with SCISSORS takes ~50% of the FastROCS time. FastROCS can take advantage of the fact that only $\binom{N}{2+N}$ comparisons need to be made (where N is the size of the library) since the all *vs.* all similarity matrix is symmetric.[25] For the 4000 molecule libraries used in our benchmarking, FastROCS requires ~ 8×10^6 overlays and SCISSORS with a 1000 molecule basis set requires 4×10^6 overlays—very close to 50%.

One *vs.* all comparisons using a precalculated SCISSORS library achieve much greater throughput than FastROCS. Interestingly, and contrary to the all *vs.* all results, FastROCS calculations were not the rate-limiting step for one *vs.* all comparisons. Instead, Tanimoto calculations using SCISSORS vectors took approximately the same amount of time as query *vs.* library FastROCS comparisons. Faster methods for calculation of Tanimotos from SCISSORS vectors would yield even better speedups relative to FastROCS.

SCISSORS can be used to predict multiconformer similarity

Three-dimensional similarity measurements are especially resource-consuming when they consider conformational ensembles of molecules instead of single conformers, but the indirect nature of SCISSORS could be problematic for predicting multiconformer similarity. To evaluate SCISSORS performance with multi-conformer molecules, we calculated generated conformational ensembles (up to 10 conformers per molecule) for each of our validation datasets using OpenEye OMEGA[26,27] (version 2.4.6). By default, OMEGA does not generate conformers for molecules with unspecified stereochemistry, and only 57 of our validation basis pools were completely expanded (*i.e.*, none of the molecules failed

conformational expansion). ROCS Tanimotos for these 57 basis pools and their corresponding libraries were calculated using FastROCS and SCISSORS, and global and local error distributions were calculated for SCISSORS predictions. The global and local error distributions for multiconformer Tanimoto predictions of all ROCS Tanimotos (Supporting Figures 10–13) were similar to those observed for single-conformer molecules, confirming the ability of SCISSORS to approximate multiconformer similarity with roughly the same fidelity as single-conformer similarity.

As an interesting point of reference, the global RMS error between single-conformer and multiconformer FastROCS combo Tanimotos was 0.1137 ± 0.0006. There are some important caveats to this number: first, we only compared results for seven libraries that retained all 4000 molecules after expansion with OMEGA; second, we are assuming that multiconformer similarity values are more “true” than single-conformer similarity. If these caveats are acceptable, this result suggests that SCISSORS predictions of similarity (at least globally) can be superior to FastROCS single-conformer comparisons.

We did not benchmark SCISSORS predictions for multiconformer molecules because not all molecules have the same number of conformers following OMEGA expansion, and this would have added additional variation to our measurements. All *vs.* all performance would not be significantly enhanced since ROCS calculations are already the rate-limiting step. However, the longer times required for multiconformer comparisons could enhance one *vs.* all performance, since FastROCS calculations would become the rate-limiting step.

Optimal SCISSORS parameters can be estimated using known similarities

The error distributions for our validation datasets are useful in evaluating SCISSORS performance in general, but our results are not guaranteed to generalize to new datasets. Accordingly, we investigated whether optimal SCISSORS parameters could be estimated by using SCISSORS to predict a set of known similarities. Since similarities between basis and library molecules must be calculated explicitly to perform the SCISSORS calculation, they are a natural choice. By calculating global and local error distributions for predictions of basis–basis and basis–library (BB+BL) Tanimotos, the optimal parameters for library–library (LL) Tanimoto predictions can be estimated, assuming that BB+BL and LL errors are somewhat correlated. We calculated error distributions for BB+BL Tanimoto predictions, as well as the average changes in RMS error when moving from BB+BL to LL predictions (Supporting Figures 14–19). We observed reasonable correspondence between BB+BL and LL errors for shape and color Tanimoto predictions, but BB+BL–LL correspondence was poor for direct combo Tanimoto predictions.

As an alternate measurement of error correspondence, we calculated the coefficient of determination (R^2) between BB+BL and LL RMS errors (as a function of dimensionality) within each validation dataset. Global R^2 values for shape and color Tanimoto predictions showed that error correlation improved with increasing basis set size. Shape Tanimoto error correlation was good for mid-range and high similarity values (Supporting Figure 20). Color Tanimoto error correlations with binning by FastROCS Tanimoto were good for low- and mid-range similarities, but binning by SCISSORS Tanimoto gave much weaker correlations for low-range similarity, and high-similarity errors were very poorly correlated (Supporting

Figure 21). Direct combo Tanimoto error correlations were generally unremarkable (Supporting Figure 22).

Although the correlation between BB+BL and LL Tanimoto prediction errors is not perfect, our results suggest that calculating the BB+BL error distribution can provide ballpark estimates of optimal SCISSORS parameters for shape and color Tanimoto predictions in a dataset-specific manner. These parameter estimates will also indicate the applicability of the error distributions reported in this paper to specific applications. As a practical note, generating the BL error distribution requires scanning across multiple dimensions for a particular basis set, but the associated computational cost can be reduced by generating SCISSORS vectors once and truncating them to the appropriate dimensionality as needed.

SCISSORS accurately predicts LINGO Tanimotos

To assess the applicability of SCISSORS to similarity metrics besides ROCS Tanimotos, we calculated global and local average RMS errors for LINGO Tanimoto predictions with and without the assumption of unity self overlap values (Supporting Figures 23 and 24, respectively). (Note: overlap values for LINGO Tanimoto calculations are not interpretable as volumes; see Methods.) SCISSORS vectors generated from LINGO Tanimotos had maximum dimensionality equal to the basis set size (the kernel matrix was positive definite), and most choices of dimensionality yielded low average RMS errors when moderately-sized basis sets were used. Contrary to our observations of ROCS Tanimoto predictions, there was no increase in error at high dimensionality for LINGO Tanimoto predictions. Local analyses showed good performance at moderate to high dimensionality across the entire true similarity spectrum.

Using true self overlap values increases high-similarity prediction errors

In an effort to improve SCISSORS performance, we attempted to identify and correct major sources of error in SCISSORS Tanimoto predictions, focusing on the individual terms in equation (2). We calculated average RMS errors for predictions of self overlap (Supporting Figure 25) and pairwise overlap (Supporting Figures 26–29) values and observed strong sensitivity to basis set size and dimensionality for self overlap value predictions. To compensate for this source of error, we introduced true self overlap values into SCISSORS Tanimoto predictions, effectively reformulating equation (2) by replacing squared norms of SCISSORS vectors with the corresponding true self overlap values:

$$T(\mathbf{a}, \mathbf{b}) \equiv \frac{\mathbf{a} \cdot \mathbf{b}}{O_{aa} + O_{bb} - \mathbf{a} \cdot \mathbf{b}} \quad (5)$$

Self overlap values (molecular volumes) can be extracted from ROCS (they are not accessible from FastROCS, as far as we are aware) or calculated independently in linear time. Under the assumption of unity self overlap values, these terms are known *a priori*. The calculations reported in this section assumed that unity was the “true” self overlap value for all molecules.

When true self overlap values were used, global SCISSORS performance for all Tanimoto predictions became virtually insensitive to both basis set size and dimensionality (Figure 8 shows results for direct combo Tanimotos). However, local analyses showed that predictions for low similarity pairs were accurate at nearly all choices of dimensionality, but that prediction quality decreased at higher similarities (Figure 9). This decay in prediction quality with increasing similarity mirrors the distribution of true Tanimotos in Figure 1. Global and local error distributions for shape, color, indirect combo, and LINGO Tanimoto predictions using true self overlaps are shown in Supporting Figures 30–33, respectively. In all cases, good performance is achieved at most choices of dimensionality for low-similarity pairs, but predictions degrade at higher similarities.

The only error-prone quantity in equation (5) is the pairwise overlap value $\mathbf{a} \cdot \mathbf{b}$. It is thus not surprising that the error distributions for Tanimoto predictions made with true self overlaps are similar to the error distributions for predictions of pairwise overlap (compare, for example, Supporting Figures 26 and 30). The striking dependence of SCISSORS performance on similarity when using true self overlap values suggests that errors in self overlap predictions in the original SCISSORS method help to compensate for errors in pairwise overlap at high similarities.

Figure 10 shows correlation plots for direct combo Tanimoto predictions with and without true self overlaps using 1000 basis molecules and 300 dimensions. Including true self overlap values tightens the correlation between true and predicted ROCS combo Tanimotos less than ~0.6, but predictions begin to stray from the correlation line (in both directions) at higher similarities. Additional correlation plots for shape, color, indirect and direct combo, and LINGO Tanimoto predictions at several choices of dimensionality are shown in Supporting Figures 34–38, respectively.

Increasing prediction errors at high similarity are not ideal for LBVS and other applications where the most interesting molecules have high similarity to one or more reference compounds. Although some mis-predicted high-similarity pairs might be recovered using inexpensive explicit methods (such as topological fingerprints), most applications of SCISSORS will likely perform better without true self overlaps. Importantly, good low similarity accuracy is usually possible with the correct choice of basis set size and dimensionality, *without* using true self overlaps. One possible exception to these conclusions is the prediction of shape Tanimotos, where the degradation of performance with increasing similarity is relatively minimal. Some applications could take advantage of separate shape and color Tanimotos predicted with and without true self overlap values, respectively.

Conclusion

Similarity methods are a core component of cheminformatics and computational drug discovery, but their computational expense can be prohibitive for some applications. The SCISSORS method uses similarity between database compounds and a preselected basis set calculate approximate similarities with much higher throughput than expensive explicit methods such as ROCS. In this report, we have made an extensive mapping of the SCISSORS error distribution under many different conditions, establishing guidelines for

selection of SCISSORS parameters. We have demonstrated that SCISSORS can be applied to multiple types of similarity, including multiconformer ROCS and LINGO Tanimotos (although LINGOs are not especially expensive to calculate).

Our results suggest several “best practices” for practical applications of SCISSORS. First, combo Tanimotos should be predicted indirectly if resources permit, since separately optimized shape and color Tanimoto predictions are likely to give better results than direct prediction of combo Tanimotos. Second, because our results are not guaranteed to generalize to all datasets, optimal SCISSORS parameters for a particular dataset should be estimated by mapping the error distribution for predictions of basis–basis and basis–library similarities. Finally, it should be remembered that SCISSORS is not a replacement for explicit similarity measurements. Instead, SCISSORS should be viewed as a filter for large-scale applications that allows resources to be preferentially allocated to the most interesting molecules.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully acknowledge support from NIH and NSF, in particular NIH U54 GM072970. We also acknowledge the following award for providing computing resources: MRI-R2: Acquisition of a Hybrid CPU/GPU and Visualization Cluster for Multidisciplinary Studies in Transport Physics with Uncertainty Quantification (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0960306). This award is funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5)

Notes and References

1. Ripphausen P, Nisius B, Bajorath J. State-of-the-art in ligand-based virtual screening. Drug discovery today. 2011; 16:372–6. [PubMed: 21349346]
2. Ripphausen P, Nisius B, Peltason L, Bajorath J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. Journal of medicinal chemistry. 2010; 53:8461–7. [PubMed: 20929257]
3. Rush TS, Grant JA, Mosyak L, Nicholls A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. Journal of medicinal chemistry. 2005; 48:1489–95. [PubMed: 15743191]
4. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nature biotechnology. 2007; 25:197–206.
5. Posner, Ba; Xi, H.; Mills, JEJ. Enhanced HTS hit selection via a local hit rate analysis. Journal of chemical information and modeling. 2009; 49:2202–10. [PubMed: 19795815]
6. Lee HS, Choi J, Kufareva I, Abagyan R, Filikov A, Yang Y, Yoon S. Optimization of high throughput virtual screening by combining shape-matching and docking methods. Journal of chemical information and modeling. 2008; 48:489–97. [PubMed: 18302357]
7. Jones G, Gao Y, Sage CR. Elucidating molecular overlays from pairwise alignments using a genetic algorithm. Journal of chemical information and modeling. 2009; 49:1847–55. [PubMed: 19537722]
8. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Haigh JA, Nevins N, Jain AN, Kelley B. Molecular shape and medicinal chemistry: a perspective. Journal of medicinal chemistry. 2010; 53:3862–86. [PubMed: 20158188]
9. Butina D. Unsupervised Data Base Clustering Based on Daylight’s Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. Journal of Chemical Information and Modeling. 1999; 39:747–750.

10. Hawkins PCD, Skillman aG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. Journal of medicinal chemistry. 2007; 50:74–82. [PubMed: 17201411]
11. http://www.eyesopen.com/rocs.
12. Haque IS, Pande VS. PAPER—Accelerating Parallel Evaluations of ROCS. Journal of Computational Chemistry. 2009
13. http://www.eyesopen.com/fastrocs.
14. Haque IS, Pande VS. SCISSORS: a linear-algebraical technique to rapidly approximate chemical similarities. Journal of chemical information and modeling. 2010; 50:1075–88. [PubMed: 20509629]
15. Haque IS, Pande VS. Error bounds on the SCISSORS approximation method. Journal of chemical information and modeling. 2011; 51:2248–53. [PubMed: 21851122]
16. Bolton EE, Chen J, Kim S, Han L, He S, Shi W, Simonyan V, Sun Y, Thiessen Pa, Wang J, Yu B, Zhang J, Bryant SH. PubChem3D: a new resource for scientists. Journal of cheminformatics. 2011; 3:32. [PubMed: 21933373]
17. Bolton, EE.; Wang, Y.; Thiessen, PA.; Bryant, SH. Annual Reports in Computational Chemistry. Vol. 4. Elsevier B.V; 2008. p. 217-241.
18. Grant JA, Pickup BT. A Gaussian Description of Molecular Shape. The Journal of Physical Chemistry. 1995; 99:3503–3510.
19. Grant JA, Gallardo MA, Pickup BT. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. Journal of Computational Chemistry. 1996; 17:1653–1666.
20. http://www.eyesopen.com/shape-tk.
21. Vidal D, Thormann M, Pons M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. Journal of chemical information and modeling. 2005; 45:386–93. [PubMed: 15807504]
22. http://www.eyesopen.com/oechem-tk.
23. Haque IS, Pande VS, Walters WP. SIML: a fast SIMD algorithm for calculating LINGO chemical similarities on GPUs and CPUs. Journal of chemical information and modeling. 2010; 50:560–4. [PubMed: 20218693]
24. Haque, IS. PhD thesis. Stanford University; 2011. Accelerating Chemical Similarity Search Using GPUs and Metric Embeddings.
25. If the diagonal elements of the matrix are assumed, only $\binom{N}{2}$ comparisons are required. We calculated diagonal elements explicitly in our experiments.
26. Hawkins PCD, Skillman aG, Warren GL, Ellingson Ba, Stahl MT. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. Journal of chemical information and modeling. 2010; 50:572–84. [PubMed: 20235588]
27. http://www.eyesopen.com/omega.

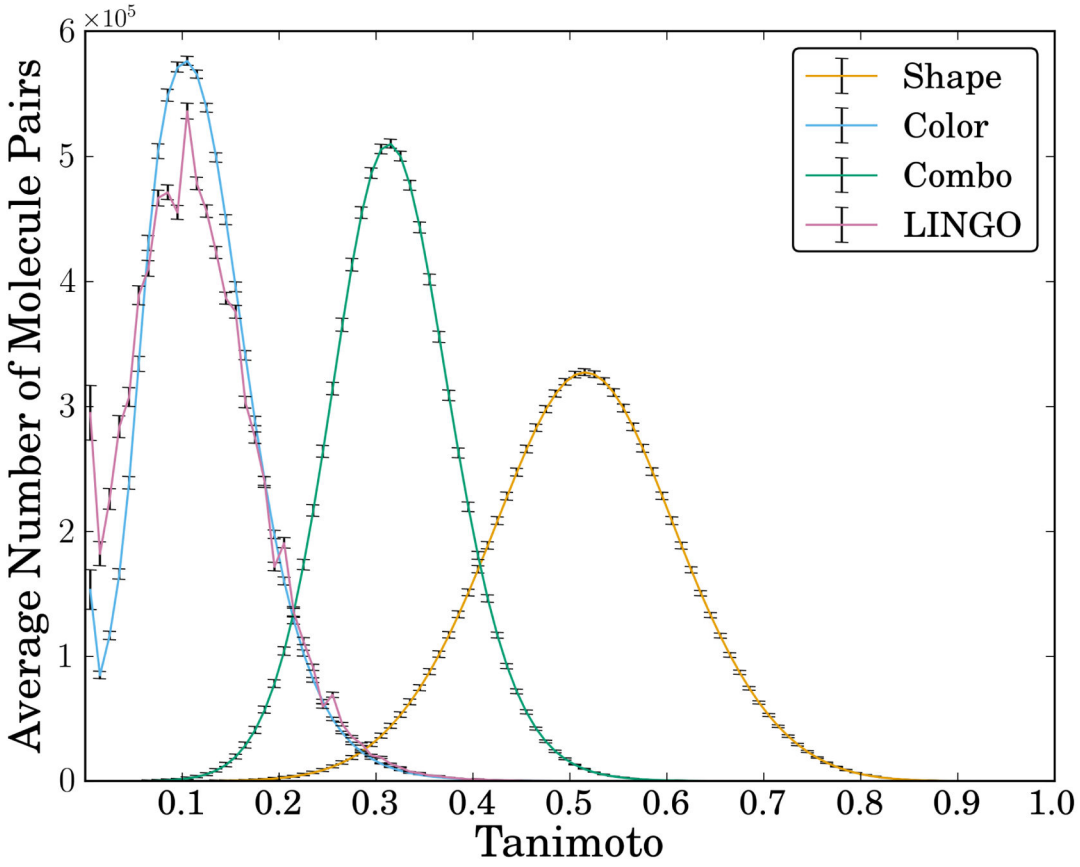
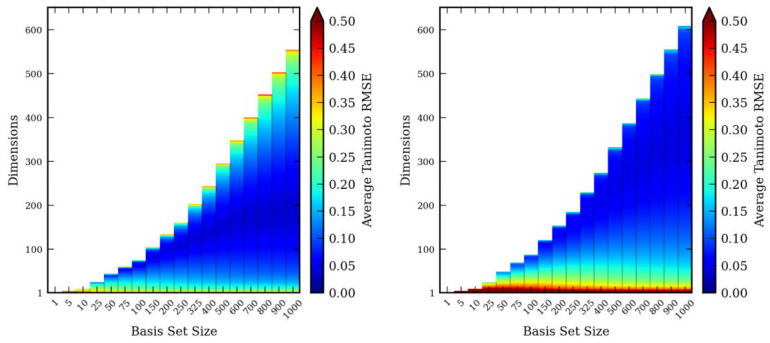


Figure 1.

Distribution of ROCS and LINGO Tanimotos in validation sets derived from Pub-Chem3D. The figure shows an averaged histogram for each similarity measure. Each dataset contains nearly 8 million unique non-self molecule pairs. The spike at very low Tanimotos is due to molecule pairs that have zero similarity by these metrics.



(a) Shape (b) Color

Figure 2.

Global average RMS errors for (a) shape and (b) color Tanimoto predictions as a function of basis set size and dimensionality.

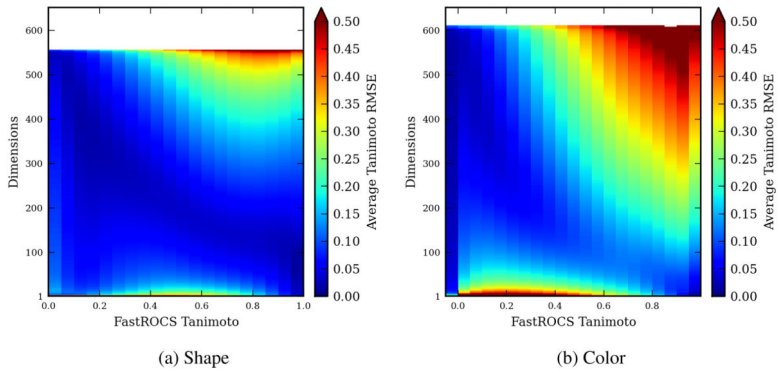


Figure 3.

Local average RMS errors partitioned by FastROCS Tanimoto for (a) shape and (b) color Tanimoto predictions. True Tanimoto similarity was partitioned into bins, each covering a Tanimoto range of 0.05, and SCISSORS prediction RMSE using 1000 basis molecules was measured for molecule pairs within each bin. The bin to the left of zero in (b) contains only molecule pairs with a color Tanimoto equal to zero, since these pairs are not included in the first positive bin.

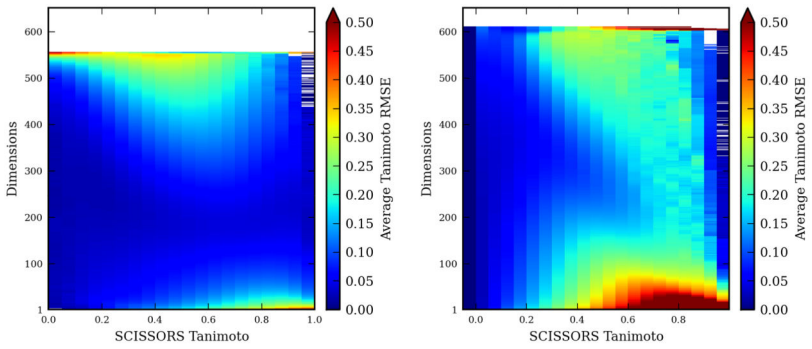


Figure 4.

Local average RMS errors partitioned by predicted Tanimoto for (a) shape and (b) color Tanimoto predictions. Predicted Tanimoto similarity was partitioned into bins, each covering a Tanimoto range of 0.05, and SCISSORS prediction RMSE using 1000 basis molecules was measured for molecule pairs within each bin. The bin to the left of zero in (b) contains only molecule pairs with a color Tanimoto equal to zero, since these pairs are not included in the first positive bin.

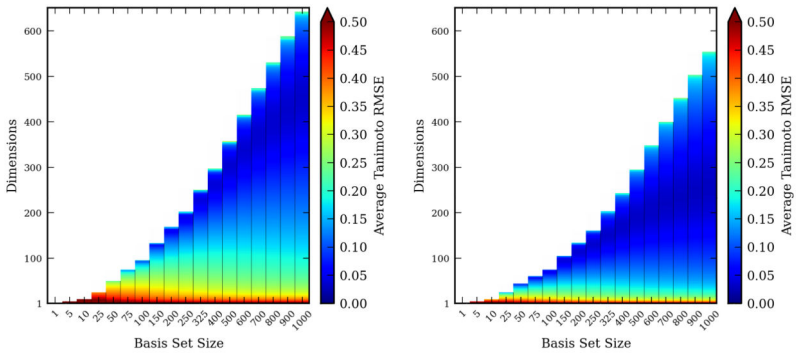


Figure 5.

Global average RMS errors for (a) direct and (b) indirect combo Tanimoto predictions as a function of basis set size and dimensionality. Indirect combo Tanimoto predictions are the average of separate shape and color Tanimoto predictions at the same basis set size and dimensionality. Direct combo Tanimotos are calculated using a kernel matrix constructed from basis set combo Tanimotos.

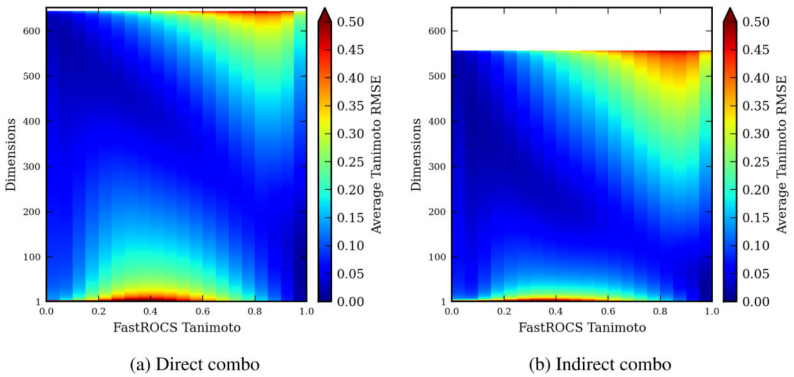


Figure 6.

Local average RMS errors partitioned by FastROCS Tanimoto for (a) direct and (b) indirect combo Tanimoto predictions.

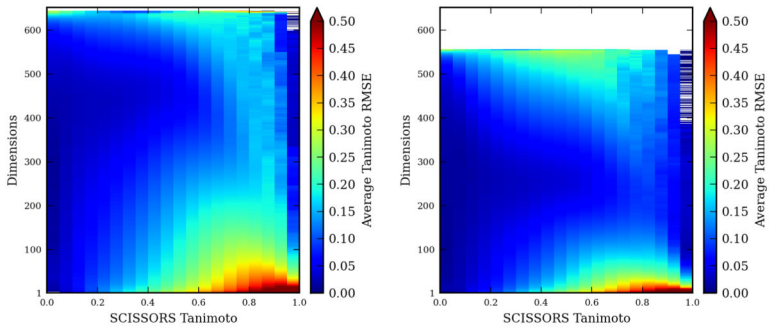


Figure 7.

Local average RMS errors partitioned by predicted Tanimoto for (a) direct and (b) indirect combo.

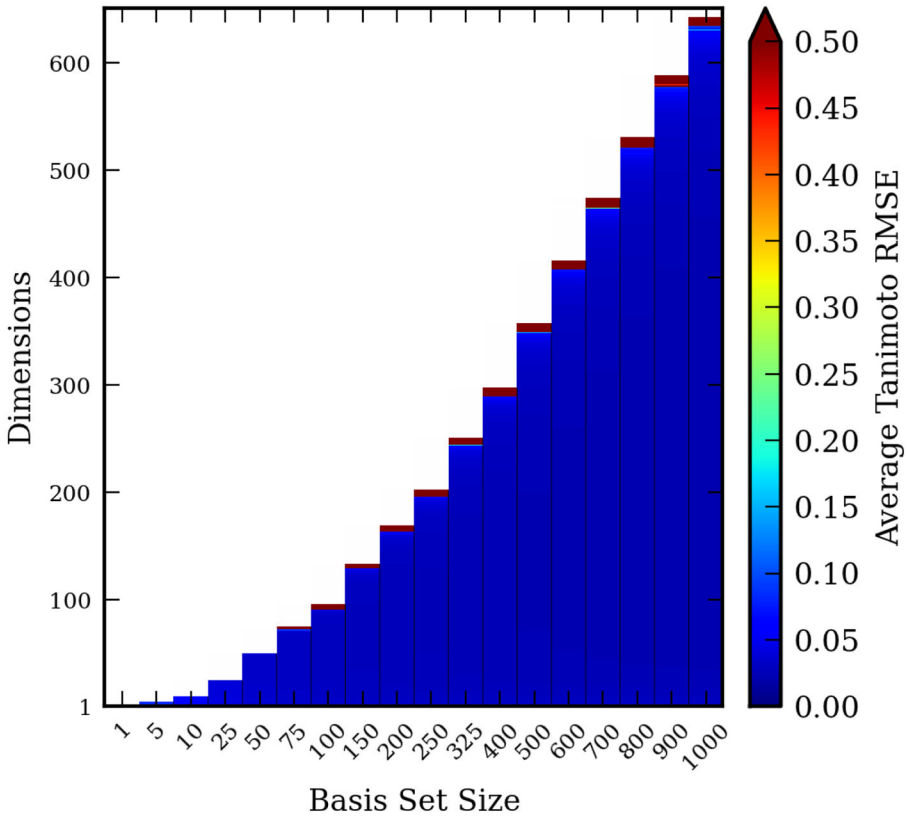
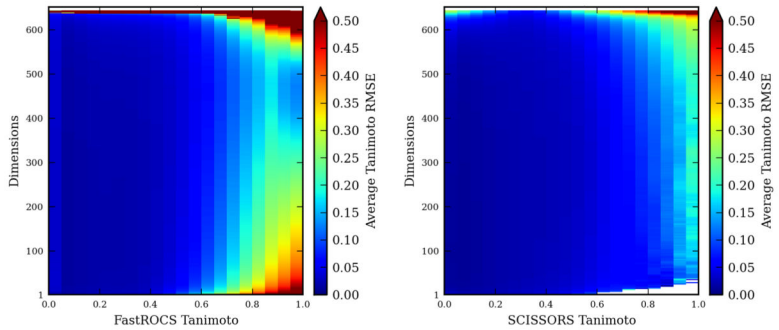


Figure 8.
Global average RMS errors for direct combo Tanimoto predictions using true self overlap values.



(a) Local — FastROCS (b) Local — SCISSORS

Figure 9.

Local average RMS errors partitioned by (a) FastROCS Tanimoto and (b) SCISSORS Tanimoto for direct combo Tanimoto predictions using true self overlap values.

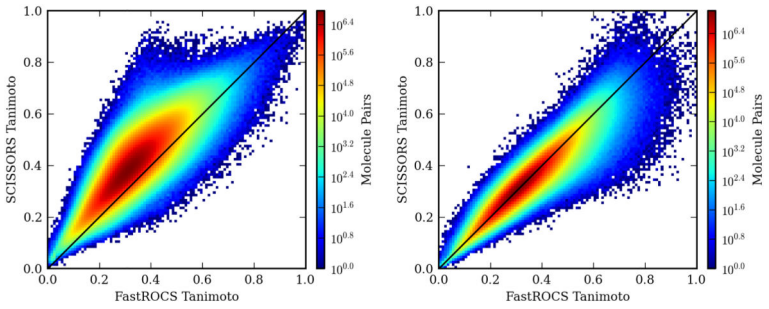


Figure 10.

Comparison of SCISSORS and ROCS direct combo Tanimotos for 100 PubChem3D subsets (~800 million molecule pairs) using 1000 basis molecules and 300 dimensions. Molecule pairs with predicted Tanimoto similarity outside the interval [0,1] are not shown (< 1% of pairs).

Table 1

Benchmarking results for calculation of all *vs*. all ROCS Tanimotos for a 4000 molecule library.

		Time (s)[a]			
	Basis Set Size	**FastROCS Similarity**	**Vector Generation**	**Tanimoto Calculation[b]**	**Percentage of FastROCS Time**
FastROCS	-	383.(3)	–	–	-
SCISSORS	125	21.2(2)	0.05(1)	1.37(2)	6
	250	43.4(5)	0.045(6)	1.37(2)	12
	500	89.(1)	0.11(1)	1.37(2)	24
	1000	188.(2)	0.32(7)	1.42(1)	50

[a] Average value for 10 validation sets. Standard deviation of the last digit is given in parentheses (rounded to the nearest single digit). For example, 2.5(3) means 2.5 ± 0.3.

[b] Tanimoto calculations used all available SCISSORS vector dimensions.

Table 2

Benchmarking results for calculation of ROCS Tanimotos between a single molecule and a 40 000 molecule library using precomputed SCISSORS library vectors.

		Time (s)[a]			
	Basis Set Size	**FastROCS Similarity**	**Vector Generation**	**Tanimoto Calculation[b]**	**Percentage of FastROCS Time**
FastROCS	-	3.22(7)	–	–	-
SCISSORS	125	0.145(5)	0.0013(6)	0.151(5)	9
	250	0.168(7)	0.00115(5)	0.17(1)	11
	500	0.198(6)	0.0012(2)	0.20(1)	12
	1000	0.250(8)	0.002(1)	0.27(1)	16

[a] Average value for 10 validation sets. Standard deviation is given in parentheses (rounded to the nearest single digit).

[b] Tanimoto calculations used all available SCISSORS vector dimensions.