JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

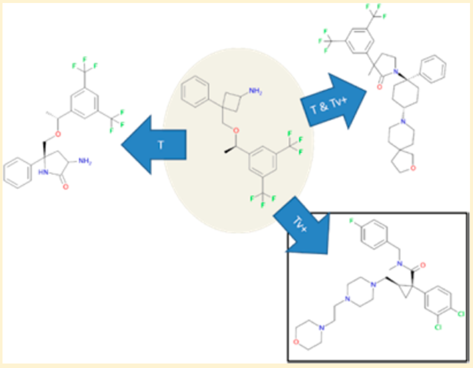
pubs.acs.org/jcim

Do Not Hesitate to Use Tversky—and Other Hints for Successful Active Analogue Searches with Feature Count Descriptors

Dragos Horvath,* Gilles Marcou, and Alexandre Varnek

Laboratoire d'Infochimie, UMR 7140 CNRS-University Strasbourg, 1 rue Blaise Pascal, Strasbourg 67000, France

S Supporting Information



ABSTRACT: This study is an exhaustive analysis of the neighborhood behavior over a large coherent data set (ChEMBL target/ligand pairs of known K_i, for 165 targets with >50 associated ligands each). It focuses on similarity-based virtual screening (SVS) success defined by the ascertained optimality index. This is a weighted compromise between purity and retrieval rate of active hits in the neighborhood of an active query. One key issue addressed here is the impact of Tversky asymmetric weighing of query vs candidate features (represented as integer-value ISIDA colored fragment/pharmacophore triplet count descriptor vectors). The nearly a 3/4 million independent SVS runs showed that Tversky scores with a strong bias in favor of query-specific features are, by far, the most successful and the least failure-prone out of a set of nine other dissimilarity scores. These include classical Tanimoto, which failed to defend its privileged status in practical SVS applications. Tversky performance is not significantly conditioned by tuning of its bias parameter α. Both initial "guesses" of $\alpha = 0.9$ and 0.7 were more successful than Tanimoto (at its turn, better than Euclid). Tversky was eventually tested in exhaustive similarity searching within the library of 1.6 M commercial + bioactive molecules at http://infochim.u-strasbg.fr/webserv/VSEngine.html, comparing favorably to Tanimoto in terms of "scaffold hopping" propensity. Therefore, it should be used at least as often as, perhaps in parallel to Tanimoto in SVS. Analysis with respect to query subclasses highlighted relationships of query complexity (simply expressed in terms of pharmacophore pattern counts) and/or target nature vs SVS success likelihood. SVS using more complex queries are more robust with respect to the choice of their operational premises (descriptors, metric). Yet, they are best handled by "pro-query" Tversky scores at $\alpha > 0.5$. Among simpler queries, one may distinguish between "growable" (allowing for active analogs with additional features), and a few "conservative" queries not allowing any growth. These (typically bioactive amine transporter ligands) form the specific application domain of "pro-candidate" biased Tversky scores at $\alpha < 0.5$.

1. INTRODUCTION

Similarity-based virtual screening, a.k.a. active analogue search, is based on the similarity principle, "similar molecules have similar properties". This old workhorse of medicinal chemists was recently[1] adapted to virtual, "in silico" screening, as "molecules encoded by similar descriptor values are, statistically speaking, bound to have similar properties". Molecules M are viewed as points defined by their descriptor values in a descriptor space, or "chemical space" (CS) where each axis $D_i(M)$ corresponds to a component of the molecular descriptor vector $\vec{D}(M)$. Formally, a CS is defined by a set of descriptors associated to a dissimilarity score $\sum(m,M)$, returning a dissimilarity[2]—distance—estimate between two points associated to molecules m and M. This may, but need not represent a "metric" in the formal sense—this shorter term will be preferred here.

Molecules placed in a CS are considered similar if the distance between corresponding points is small. If so, then the difference between their property values—which can be understood as a distance in an alternative "property space" (PS)—should also be, statistically speaking, small. The nature of this expected "neighborhood behavior" (NB) of chemical spaces has long been debated.[3−8] Authors have evidenced that the correlation between distances in CS and PS is not simple. CS ("structural") similarity should imply property similarity. Molecules of similar properties, however, do not need to be structurally similar. NB is a one way inference: near neighbors in CS should—unless they happen to form an activity cliff[9−11]—be close in PS as well.

The above-mentioned aspect is very important when trying to numerically estimate the degree of NB compliance of a CS. However, in all-day Similarity-driven virtual screening (SVS), when a single active compound M is compared to candidate molecules m from a database, the foremost issue is the choice of the CS in which to operate. In particular, the question of the metric $\sum(m,M)$ is paramount—how should $\vec{D}(M)$ and $\vec{D}(m)$ be compared in order to maximize NB compliance? Equally important—how similar is similar? Up to what threshold of $\sum(m,M)$ should one reasonably expect active analogues to

Received: February 14, 2013

dx.doi.org/10.1021/ci400106g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

occur? In practice, $\sum(m,M)$ allows a prioritization of candidates, starting from the nearest neighbor of M. Chemical common sense may sometimes decide how far down this list one should go. Things are however not that simple, especially if the CS and metric were chosen to be complementary to the innate scaffold-based sense of similarity of chemists, in order to support scaffold hopping.[12−16] In such cases, it is often difficult to even decide whether the nearest neighbor is near enough to justify synthesis/purchase and testing, and reading the absolute value of $\sum(m,M)$ would not really help. Even though various studies concerning to-date most widely used chemical spaces did suggest reasonable cutoff values of $\sum(m,M)$, these mostly concern specific binary fingerprints in conjunction with the Tanimoto score. The "Tanimoto > 0.85" suggestion for Daylight fingerprint-based similarity searches quickly became a dogma of chemoinformatics, eventually shown not to be the universal answer to similarity screening.[17] Worse, while the finite nature of the range [0,1] of Tanimoto (dis)similarity allows the user to make a sometimes lucky guess on the cutoff problem, things are less obvious with open-ended (Euclidean), or tunable scores such as the asymmetric Tversky formula.

Retrospective analysis of correct setup choices may, however, return some insights improving the chances of predictive SVS. Neighborhood analysis of so-far known actives and analogues, involving maximal chemical and target diversity, might display statistical trends showing some CS being more often successful than others. If so, adopting these choices by default may globally enhance success chances—even though not every individual SVS experiment may directly benefit. Furthermore, specific analysis per target classes may highlight whether some descriptor/metric combinations are preferred by certain receptor families.

The above is the strategy followed in the present work. It first requires an objective criterion to quantitatively define the success of SVS. In this respect, the NB optimality criterion defined,[7] employed,[4,18] and eventually refined[3] in previous works will be used. It is an alternative to classical SVS success scores, like ROC AUC measures. Its advantage becomes apparent in scenarios involving active candidate molecules being genuinely dissimilar with respect to the query, when ROC AUC[19] scores are unduly penalized by the failure to retrieve these latter.

Previous NB studies, however, fell short of providing practically useful hints on how to maximize SVS success. Partly, these were concerned with the rather academic problem of global[6,18] NB compliance, by contrast to the specific monitoring of SVS success introduced as "local" NB later on.[3] The latter work furthermore focused on the specific problem of combinatorial libraries, in which the rather important similarity of forcibly related compounds did not allow any of the competing descriptor/metric combinations to emerge as being significantly more successful than others. However, further diversification of the set of considered CS, due to the later introduction of property-labeled ISIDA fragment counts,[20] revealed the existence of some descriptor spaces clearly outperforming the originally considered options. Nothing, though, could be learned about choosing the metric. All the state-of-the-art options (Euclid, Tanimoto, etc.) performed comparably in that universe of combinatorial compounds with no dramatic size fluctuations. Also, biological activity was confined to a narrow set of five proteases, rendering any analysis implicitly class-specific. Studies[6,7,21] operating on a set of marketed drugs and reference compounds (BioPrint[22,23]) addressed an activity profile of more than 150 targets, yet did not live up to expectations of providing practical guidelines for SVS. The problem stemmed mainly from the extreme bias in the choice of molecules—many singletons (individual drugs, for which the series of analogues leading to them was not included in the study) and clusters of highly similar "me too" compounds which however mostly concerned tricyclic antidepressants.

This work therefore aims to correct the above-mentioned sources of bias, eventually coming up with some general recommendations on maximizing SVS success. It will address SVS on an as large as possible basis, using distributed computing in order to perform a maximum of active analogue retrieval simulations, exploiting the wealth of structure—activity data in the ChEMBL database.[24] ChEMBL covers both an as large as possible (and yet publicly accessible) diversity spectrum in terms of both targets and compounds, all while ensuring a dense coverage of the structure—activity space of each target, including analogue series and their associated biological activities. In order to avoid ambiguities with respect to bioassay value interpretations, only compound—target pairs with reported thermodynamic affinity (pK_i) were picked (selection and data curation courtesy of Prof. J. Bajorath, University of Bonn). Only 165 targets covered by series of 50 or more ligands of known pK_i were considered. Selected actives of each target were used as query compounds. Parallel distributed calculations assessed, for each of considered chemical spaces the local NB criteria of SVS experiments. They opposed, for each of 165 targets, each of the therewith associated query compounds to a database of 10 000 (presumed) inactive, random molecules spiked with the ChEMBL compounds of known pK_i (less the current query compound).

Concerning the explored chemical spaces, the herein work will focus on integer-value feature count fingerprints—ISIDA property-labeled fragment counts and fuzzy pharmacophore triplets, since

1. These were known to display excellent NB, as far as previous studies could show.
2. They are definitely much more information-rich and chemically relevant (pH-sensitive) than binary fingerprints.
3. They are nevertheless conceptually related to either of the industry standard, state-of-the-art fingerprints such as MACCS keys,[25] Daylight fingerprints,[26] binary pharmacophore triplets,[27] and circular Pipeline Pilot fingerprints.[28] They represent generalized, information-rich adaptions of the same basic principles (fragment/substructure/local chemical context monitoring).
4. They are homogeneous and chemically interpretable, unlike vectors of whole-molecule indices—which is of paramount importance when designing adapted metrics.

Being integer rather than binary is an important common characteristic of this class relative to the selection of the metric/dissimilarity score. Albeit any metric can be formulated for both binary and integer/real value vectors, their behavior may heavily depend on the nature of the fingerprint, as will be shown later on. The herein used NB formalism offers the opportunity to explore the proficiency of any potentially meaningful score, no matter whether bound or unbound.

Classical Euclidean and Tanimoto metrics are naturally included in this benchmark, next to the intellectually very appealing asymmetric Tversky score. It provides the possibility to give different weight to the features contained in the active query vs the ones seen in the candidate compounds. Lately, this score was subject of several in-depth studies[29−32] highlighting that a positive bias (increased importance) given to the features present in query compounds tends to improve active analogue searches (in binary fingerprint spaces). A detailed discussion of this behavior, related to differential fingerprint occupancies and to molecular size is provided.[31] It highlights that actives are, in general, more complex (feature-rich) molecules than inactives i.e. feature loss likely signals activity loss. Thus, scores specifically penalizing feature loss fare better. Furthermore, sophisticated scoring schemes taking complexity into account have been designed.[33] Nevertheless, the Tversky scheme is only rarely used in real-life applications. Actually, a query of "Tversky" and "molecular similarity" as "Topics" on the Web of Knowledge[34] returned (Oct 2102) only the above-cited methodological papers—no practical applications. This is in sheer contrast to the plethora of reported uses of the Tanimoto score. A possible reason: Tversky is a tunable formula with one (or two) fittable parameter(s), which impact on the output range and implicitly make it impossible to guess a priori "how similar is similar" in terms of Tversky scores—a problem which the herein reported strategy may help to solve. There was no attempt to fit these parameters here: three a priori interesting setups were considered, like three "independent" dissimilarity scores. The goal here was to check whether Tversky at $\alpha > 0.5$ is a robust choice in SVS, no matter the exact α value—is it as straightforward to use as any other metric, or is tedious tuning needed? In addition, some novel dissimilarity scores will be introduced and tested in this context.

Next, the analysis focuses on target families, in order to evidence whether different target families show any marked preference for specific CS. If so, the user could make a knowledgeable choice of descriptors and metric in function of the targeted bioactivity class. This tentative study focused on three target classes: monoamine (rhodopsin-like) GPCRs, other GPCRs, and kinases.

Eventually, the focus shifts on the query complexity issue: does the complexity of the query compound impact on the optimal CS choice? Are certain descriptors/similarity scores more suited than others to search for analogues of complex molecules?

Eventually, it is important to mention that this work served for the development of a (first, to our knowledge) publicly accessible web portal http://infochim.u-strasbg.fr/webserv/VSEngine.html providing a Tversky-based similarity search tool within a significant chemical space of both commercially available and activity-annotated compounds. By means of this web tool, the Tversky score was used, in parallel to Tanimoto, in exhaustive SVS over an entire publicly accessible database of 1.6 M compounds, in order to provide a concrete illustration of results that can be expected when using these scores.

2. METHODS

The following will first mention the compound and target selection, the chemical spaces (used descriptors, classical, and novel metrics), then briefly revisit the NB scoring scheme and explain how it was used to monitor relative NB propensity of considered CS.

2.1. Compounds, Targets, and Query Molecules.

The source of herein used compounds and activity data is a subset, chosen, and curated by Prof. J. Bajorath, University of Bonn, and his team, from the ChEMBL database. It focuses on compounds with reported thermodynamic pK_i values for various biological targets. Only targets with 50 or more reported ligand pK_i values were kept for this study—the list of the 165 concerned proteins and their ChEMBL reference codes is available as Supporting Information. For each such target, ligands with reported pK_i values were standardized using the ChemAxon[35−37] toolkit (basic aromatization, split-charge representation of nitrogen oxides, conversion to the most probable tautomer according to ChemAxon, and submission to the ChemAxon pK_a plugin to determine the relative populations of microspecies and according assignment of pharmacophore flags). Additionally, ten thousand randomly picked commercial compounds serving as decoys in SVS were equally standardized. This setup was considered in order to mimic—as far as possible—a typical corporate database environment, consisting of many congeneric series of analogues from past and current drug discovery projects, plus commercially available molecules. Such collections are intrinsically biased, but typical in chemoinformatics, and often used[38] for similarity-based studies.

"Actives" of each target T, are sorted by decreasing pK_i (larger = more active). The associated set of queries M_1^T, M_2^T, ..., M_i^T, $i = 1 \ldots Q^T$ is composed of (a) the top 1/5 actives of this list—truncated to 100 if more than 500 ligands are reported, plus (b) the 1/5 of binders of medium potency, truncated again to 100 and starting from the median compound of the list. This ensures that there would be no bias with respect to the often encountered use of top nanomolar inhibitors in SVS, whereas in practice analogue search of moderately actives is important in drug design.

2.2. Molecular Descriptors.

Descriptors were generated for each involved molecule, then dispatched into separate folders associated to targets. Each such folder stores the descriptors of the "actives" (in the above-mentioned sense) of every target in separate files. The SVS featuring query i of target T, based on descriptor set D is then easily initialized by (1) copying line i of the descriptor file T/D into the current query descriptor file of T, (2) copying all other lines of T/D as the current "database" file of T, and (3) adding descriptors D of the inactive decoys to the latter.

All the considered descriptors belong to the family of integer-value fragment/feature counts, for reasons already mentioned in the Introduction:

Fuzzy Pharmacophore Triplets (FPT)[21,39] represent fuzzy counts of monitored triplets of pharmacophore features (hydrophobe, aromatic, H-bond donor and acceptor, cation, and anion) carried by atoms, at given topological interfeature distances, i.e. "edge lengths" of the considered triangles. Out of the considered FPT setups discussed in the original publication,[21] the default **FPT1** was employed here.

Property-Labeled ISIDA Fragments were introduced[20] as a generalization of "classical" (atom symbol-labeled) molecular fragment counts.[40−42] The specific fragment types "*ffPPlu*" were named according to the convention in ref 20, namely "fragment type *ff*" + "coloring property *PP*" + "lower and upper fragment size *lu*", where

- Fragment type may be one of
  - **seq**—sequence
  - **aa**—augmented atom (i.e., atom-centered, circular fragments)

    - **tree**—trees, i.e. augmented atoms in which the nature of all atoms except the center and the terminal "leafs", is ignored.
- The presence of a "**b**" label following the fragment types means that bond order information is taken into account (ignored by default).
- Considered coloring properties are
    - **SY**—atomic symbols
    - **PH**—pH-sensitive pharmacophore type at pH = 7.4
- Eventually, the following two digits refer to the lowest and the highest (sequence length or, respectively, circular fragment radius): **seqSY37** are counts of symbol-colored sequences of three to seven atoms, like "**CCNCC**"—matching anything from diethylamine from ethylaniline (because this scheme ignores the orders of involved bonds). However, "HHPHH" in pharmacophore-labeled **seqPH37** (H = "hydrophobic", P = "positive charge") is populated, at the used physiological pH, only in aliphatic amines—in anilines, the nitrogen is not protonated and counts as an acceptor. A pharmacophore tree descriptor **treePH03** will count, for example, a circular fragment of radius two around the α carbon of alanine zwitterion [N+]C(C)C(=O)[O—] like "a hydrophobic atom connected, through one branch, to a cation at one bond apart (—NH_3^+), through a second branch to two hydrogen bond acceptors at two bonds apart (the carboxylate oxygens), and to a hydrophobe at one bond apart (β carbon)". Atoms one bond apart from the center (N and β-carbon) are included in this tree fragment of radius two only because they (prematurely) terminate the respective branch. However, the nature of the carboxylate C, which in our typing strategy is the carrier of the anionic property flag, is ignored. The same tree fragment is therefore present in 1-aminonitroethane, for example. This does not mean that information about the anionic nature of the carboxylate is lost: that feature will be captured by other tree fragments, since all possible radii between 0 (plain feature counts) and three are considered. Please refer to the original publication for more details.

Note that all pH-sensitive descriptor sets are, as shown in cited works, not simple colored subgraph or triplet counts, but fuzzy average estimators of the overall occupancy of corresponding subgraphs/triplets in a population of the various microspecies of the molecule at proteolytic equilibrium in a neutral buffer of pH = 7.4. Triplets are also fuzzily mapped on the basis triangles enumerated by the fingerprint. Formally, however, these fuzzy occupancy levels are nevertheless rendered as integers. One subgraph present once in all the molecules, irrespective of their protonation status, is given a population level of 100, rather than 1. A colored subgraph specific to, say, the conjugated acid form, will be assigned an occupancy [A] < 100, where A is the percentage of the acidic form at equilibrium, and "[]" means integer truncation.

The six descriptor sets effectively used in this work are **FPT1**, **seqPH37**, **seqSY37**, **aabPH02**, **treePH03**, and **treeSY03**. Please note, moreover, that the herein used descriptors are restricted to counts of fragments/features present in the bioactive and commercial compound reference set used to calibrate the Kohonen nets used to speed up similarity searching.[43] Monitoring features other than participating in the Kohonen maps is practically useless in our setup. Yet, as this initial set was quite diverse, the risk of losing valuable chemical information because of discarding additional fragment counts is not important. Therefore, the descriptors behave like de facto fixed-size vectors, albeit this size may be quite large (**aa** and **tree** terms include up to 20 000 fragments, while sequences reach few thousands. The 4418 most often encountered basis triplets of the 4494 monitored in the full **FPT1** vector are exploited here).

2.3. Dissimilarity Scores.

This work being confined to integer-value feature count spaces, it may be assumed that molecules M and m to be represented by vectors $\vec{D}$ and $\vec{d}$, respectively, where D_i and d_i represent occupancy levels associated to some subgraph/triplet ad hoc associated to position i of the vector. Whereas FPT vectors are fixed-size, each monitored triangle being beforehand associated to a position in the vector, fragment/subgraph count vectors are not upper bound in terms of number of elements. Therefore, vector dimension will vary from pair to pair of molecules, since $\vec{D}$ and $\vec{d}$ must be constructed as follows: (1) copy the occupancies of the $N(M)$ features occurring in M as first elements of $\vec{D}$; (2) fill the first $i = 1 \dots N(M)$ elements of $\vec{d}$ with zeroes if element i is absent from m or with the corresponding occurrence level from m otherwise; (3) add counts of subgraphs seen in m only into $\vec{d}$, starting from position $N(M) + 1$, and set the corresponding elements in $\vec{D}$ to zero. Thus, the final dimension of $\vec{D}$ and $\vec{d}$ equals the number of features seen in either of M and m, let this be called $N_{OR}(M, m)$. Similarly, let $N_{AND}(M, m)$ denote the number of features populated in both M and m, and $N_{EXC}(M, m)$ be one of the features seen exclusively in M and absent from m. Obviously, $N_{EXC}(m, M)$ is distinct from $N_{EXC}(M, m)$. With binary fingerprints, classical metrics could be formulated in terms of N_{AND} and N_{EXC} only. With integer/real vectors, the above counts must be replaced by cross product terms, as below:

$$
\begin{aligned}
N(M) &\rightarrow \mathrm{NORM}(M) = \sum_{i=1}^{N(M)} D_i^2 \\
N_{AND}(m, M) &\rightarrow \mathrm{AND}(M, m) = \sum_{N_{OR}(m,M)}^{i=1} D_i \times d_i \\
N_{EXC}(M, m) &\rightarrow \mathrm{EXC}(M, m) = \sum_{i|d_i=0} D_i^2 \\
N_{EXC}(m, M) &\rightarrow \mathrm{EXC}(m, M) = \sum_{i|D_i=0} d_i^2
\end{aligned}
\tag{1}
$$

In terms of above-defined entities, the classical dissimilarity scores (Euclid E, Tanimoto T, Tversky Tv), generically denoted as $\Sigma(M, m)$ can be written as

$$
\begin{aligned}
E(m, M) &= \sqrt{\sum_{N_{OR}(m,M)}^{i=1} (D_i - d_i)^2} \\
T(M, m) &= 1 - \frac{\mathrm{AND}(M, m)}{\mathrm{NORM}(M) + \mathrm{NORM}(m) - \mathrm{AND}(M, m)} \\
\mathrm{Tv}(M, m, \alpha) &= 1 - \frac{\mathrm{AND}(M, m)}{\alpha \mathrm{EXC}(M, m) + (1 - \alpha)\mathrm{EXC}(m, M) + \mathrm{AND}(M, m)}
\end{aligned}
\tag{2}
$$

Originally,[44] the Tversky score features two parameters α and β. However, the working hypothesis $\alpha = 1 - \beta$ has been adopted upfront in this work. The goal of the present work is not to fine-tune this score, but rather to show that Tversky is likely to be a competitive metric over a rather broad range of parameter values. If its use is indeed not preconditioned by any sophisticated calibration procedure, the ad hoc choice of replacing β by $1 - \alpha$ in the previous equation should not be penalizing. The larger α, the quicker dissimilarity will increase with respect to the number (and/or population level) of features specific to query molecule M. By contrast, its increase with respect to candidate m-specific features slows down: losing the key features which rendered the query M active is more penalizing than accepting novel features in a candidate (which, if lucky, will not interfere with−or even enhance−binding). Three parametrization schemes have been therefore employed as three "independent" metrics competing against all others in the benchmark tests: one heavily ($\alpha = 0.9$) and one lightly ($\alpha = 0.7$) biased in favor of query-specific features, plus one lightly biased in favor of candidate features ($\alpha = 0.3$). These will further on be denoted Tv+, Tv, and Tv−, respectively.

Note that, unlike with binary fingerprints, $\mathrm{EXC}(M, m) + \mathrm{EXC}(m, M) + 2\mathrm{AND}(M, m) \neq \mathrm{NORM}(M) + \mathrm{NORM}(m)$, as common features need not to be equally populated. This represents a major issue with Euclidean distances. A pair of complex molecules with a lot of overlap in terms of populated features may score very high Euclidean distances, due to the sheer accumulation of $D_i - d_i$ differences, even though both D_i and d_i are positive. By contrast, pairs of small molecules with no common features (in this case, Euclid = $[N(M) + N(m)]^{1/2})$ will come up at low distances, if both $N(m)$ and $N(M)$ are small. This makes little chemical sense, as common features−even if not identically populated−should rather signal structural likeness. Having as common point the *absence* of a feature should not render molecules similar. In order to investigate how to best respond to this problem, some straightforward adaptations of the Euclidean (and, alternatively, of the related Hamming distance, the sum absolute differences $|D_i - d_i|$) have been introduced. First, the root-mean-square R and, respectively, the absolute-mean deviations A of populated features scale the total sums with respect to the total number of features in the pair:

$$R(m, M) = \sqrt{\frac{\sum_{i=1}^{N_{\mathrm{OR}}(m,M)} (D_i - d_i)^2}{N_{\mathrm{OR}}(m, M)}}$$

$$A(m, M) = \frac{\sum_{i=1}^{N_{\mathrm{OR}}(m,M)} |D_i - d_i|}{N_{\mathrm{OR}}(m, M)} \quad (3)$$

Terms above still do not differentiate between contributions from features that are specific to either M or m, which are obvious markers of dissimilarity and contributions stemming from common, but differently populated, features−a priori less obvious markers of dissimilarity. A simple scheme to down-weigh the latter is to rescale the scores above by the ratio of specific vs total features in the pair:

$$\mathrm{RW}(m, M) = R(m, M)\frac{N_{\mathrm{XOR}}(m, M) + N_{\mathrm{XOR}}(M, m)}{N_{\mathrm{OR}}(m, M)}$$

$$\mathrm{AW}(m, M) = A(m, M)\frac{N_{\mathrm{XOR}}(m, M) + N_{\mathrm{XOR}}(M, m)}{N_{\mathrm{OR}}(m, M)} \quad (4)$$

Note that in above equations query- and candidate-specific features are equally contributing, but they could easily be rendered asymmetric.

The nine dissimilarity scores effectively employed in the following were Tv+, Tv, Tv−,T, E, R, A, RW, and AW. In conjunction with the six descriptor spaces, this yields to a total of 6 × 9 distinct CS, defined as combinations of each descriptor set with every considered metric.

2.4. SVS Proficiency: The Ascertained Local NB Optimality Score.

Local NB, as defined previously,[3] captures the quality of a SVS experiment in a single number, the ascertained local optimality criterion, all while returning the optimal dissimilarity cutoff, yielding the best compromise between active analog purity and retrieval rates. The following will briefly revisit the methodology and adapt it to the current context of pK_i activity values. Consider M to be an active compound against target T, used as an SVS query against a set of N_P molecules—here, the other ChEMBL structures of known pK_i with respect to T, plus the set of 10 000 decoys. This amounts to calculating dissimilarity scores $\Sigma(M, m)$ for all the N_P query-candidate pairs in the current CS. In classification models, these N_P pairs could be split into subsets of $N_=$ "active−active" and $N_{\neq}$ "active−inactive" pairs, as M is, by definition, active. In a property space (PS) defined by a continuous (and accurate) pK_i parameter, this classification will be replaced by a continuous activity difference score $\Lambda(M, m)$:

$$\Lambda(M, m) = \begin{cases} 0 & \text{if } \mathrm{p}K_i(M) - \mathrm{p}K_i(m) < 0.5 \\ 1 & \text{if } \mathrm{p}K_i(M) - \mathrm{p}K_i(m) > 3.0 \\ \dfrac{\mathrm{p}K_i(M) - \mathrm{p}K_i(m) - 0.5}{2.5} & \text{otherwise} \end{cases} \quad (5)$$

In other words, the activity difference is zero if the candidate is more, or roughly as active as the query (within half a log). Loss of three logs of activity or more clearly qualifies m as inactive, whereas a gradual "grey zone" covers the intermediate situation. Also, note that by definition $\Lambda(M, \mathrm{m}) = 1$ for all m of the decoy compound set, for which pK_i values are not available.

In terms of $\Lambda(M, m)$, the fuzzy count of active−active pairs becomes $N_= = \Sigma_m[1 - \Lambda(M, m)]$, whereas $N_{\neq} = \Sigma_m \Lambda(M, m)$. The identity $N_= + N_{\neq} = N_P$ thus holds for the fuzzy formalism too. Note that $N_=$ and $N_{\neq}$ are constants of the entire data set and do not relate to the subset selected according to the dissimilarity radius, as will be described further on.

Now, given a dissimilarity threshold d, pairs with $\Sigma(M, m) < d$ will be selected. Let $f(d)$ denote the fraction of selected pairs.

Among the $N_P f(d)$ selected, active−inactive pairs count as "false similars" FS (activity cliffs, in other words). However, active−active pairs that were *not* selected count as "potentially false dissimilars" PFD. In terms of $\Lambda(M, m)$, the fuzzy counts of FS and PFD at cutoff d become

Table 1. Query Subsets Q^a

query set	description	size	complexity	mean success rates
all	full set of considered queries, selected among binders to one or more of 165 ChEMBL targets, of known pK_i	13514	779 ± 396	6.2% (exc.) 28% (good)
pharma-high	subset of queries with maximal numbers of populated pharmacophore triplets	3933	>950	15.3% (exc.) 43.6% (good)
pharma-low	subset of queries with minimal numbers of populated pharmacophore triplets	4453	<570	2.0% (exc.) 14.8% (good)
monoamine GPCR	queries pertaining to either of 34 monoamine GPCRs (Supporting Information)	3290	647 ± 285	1.0% (exc.) 17.1% (good)
other GPCR	queries pertaining to either of 44 other various GPCR targets (Supporting Information)	4370	943 ± 400	11.6% (exc.) 39.1% (good)
kinases	queries pertaining to either of 7 kinases (Supporting Information)	290	668 ± 204	24.5% (exc.) 60.3% (good)

[a] "Complexity" reports either average/standard deviation of populated FPT1 triplet numbers, or triplet number threshold—when this was the selection criterion. Mean success rates represent the fraction of successful SVS (at acceptance level in parentheses, "exc."—excellent) out of all the SVS experiments (54× set size) performed for each query subset.

$$N_{\mathrm{FS}}(d) = \sum_{m|\Sigma(M,m)<d} \Lambda(M, m)$$
$$N_{\mathrm{PFD}}(d) = \sum_{m|\Sigma(M,m)\geq d} 1 - \Lambda(M, m) \tag{6}$$

With respect to the above, the trade-off between increasing d—and implicitly $f(d)$—in order to minimize N_{PFD}, yet without increasing N_{FS}, can be monitored by the Ω criterion:

$$\Omega(d) = \frac{\kappa N_{\mathrm{FS}}(d) + N_{\mathrm{PFD}}(d)}{\kappa N_{\mathrm{FS}}{}^{(\mathrm{null})} + N_{\mathrm{PFD}}{}^{(\mathrm{null})}} = \frac{\kappa N_{\mathrm{FS}}(d) + N_{\mathrm{PFD}}(d)}{\kappa N_{\neq} f(d) + N_{=}[1 - f(d)]} \tag{7}$$

Here $k > 1$ ($k = 5$ in the present work) outlines the higher importance of keeping the number of false similar as low as possible. The sum to be minimized is related to its expectation value—the denominator of the fraction—corresponding to the null hypothesis that Σ does not exhibit any NB at all. In the given CS, the optimal cutoff d^* minimizing $\Omega(d)$ is used as a dissimilarity radius. The lower the corresponding $\Omega(d^*)$, the more successful this SVS. However, in order to compensate for artifacts of low $\Omega(d^*)$, due to statistical fluctuations fortuitously placing more than expected actives within the selection, an ascertained (excess) optimality criterion was defined. To this purpose, mean "⟨⟩" and standard deviation σ of $\Omega(d)$ upon repeatedly (20×) using random number series instead of calculated metrics are determined, and the Ascertained criterion $\Xi(d)$ is then defined such as to reach a *maximum* value at optimal d^*. SVS success will be thus estimate by its $\Xi(d^*)$ or, succinctly, Ξ^*:

$$\Xi(d) = \Omega(d)_{\Sigma(M,m)=\mathrm{rand}()} - \sigma[\Omega(d)]_{\Sigma(M,m)=\mathrm{rand}()} - \Omega(d) \tag{8}$$

Any positive Ξ^* signals that Σ(M, m) performs better than random, as statistical noise was already subtracted, and should ideally approach one.

2.5. Query Compound Classes. A classification of queries (Table 1) by complexity has been undertaken, using simple criteria: the number of features present in the molecules. This study focused on pharmacophore complexity, i.e. the number of populated pharmacophore triplets (**FPT1**). Out of the employed queries, one (rough) tier of lowest populated respective feature counts were classified into the "pharma-low" category. "Pharma-high" regroups, by contrast, the third tier of feature-richest queries. The intermediate tier was not considered for analysis, in order to ensure a significant gap between the complexity of the simplest "high" and most complex "low".

Also, three subsets regrouping queries associated to monoamine GPCRs, to GPCRs of other various families, and to kinases were defined. Out of the herein considered targets, 34 were identified as Rhodopsin-like monoamine GPCRs, 44 as belonging to other GPCR families, and 7 kinases (Supporting Information). They are not exhaustive, as they were gathered by means of basic text mining (regular expression search of key labels in ChEMBL target classification records, followed by rejection of false "hits"). They may thus not include targets with eclectic names not explicitly hinting to their families.

2.6. Benchmarking Procedure. For each query associated to every target (total query number being 13 514), the complete set of SVS experiments in all the 54 considered CS (6 descriptor sets × 9 metrics) took roughly 12 h to complete in parallel (one target/CPU core), on the High Performance Cluster (HPC) of the University of Strasbourg. This amounts to about 10^{10} dissimilarity score estimations, followed by ascertained optimality score determinations, for each of the 13 514 queries × 54 CS = 729 756 SVS simulations. Technical problems (file I/O errors, scheduler failures due to an important usage of the machine) randomly occurred in 1.9% of the cases—only 716 001 optimality scores were retained.

Benchmarking proceeded via a classification and counting scheme of successful vs total numbers of SVS experiments, as follows:

- According to the operational premises ***O***, SVS experiments can be naturally classified with respect to the CS in which they operated, but also with respect to metric or descriptors only.
- SVS experiments can also be subdivided with respect to the nature of their query ***Q***, following the previously visited query subsets—by pharmacophoric complexity and by target class, respectively.
- Eventually, SVS experiments classify with respect to their success rate, as "successful" if its Ξ^* exceeds some positive user-defined acceptance levels ***a***: at 0.25, 0.5, and 0.75, SVS subsets of "acceptable or better", "good or better", and "excellent" SVS experiments will be defined. For clarity reasons, the "or better" adagio will be left out—it is understood that "excellent" is a subset of "good", which is a subset of "acceptable".

This triple classification allows subsets of SVS experiments to be formally denoted as $S(O, Q, a)$. For example, $S(T$, kinases, 0.25) regroups all the SVS concerning queries associated to selected kinases, performed using the Tanimoto score (irrespective of descriptors) and having an acceptable success status. A wildcard asterisk will be used to stand for all possible choices with respect to a criterion: S(FPT1.E, *, *) denotes all the SVS experiments in the pharmacophore triplet-Euclidean metric CS, irrespective of query compound and irrespective of success status. The number of SVS experiments in such a subset S will be simply denoted as $\#S$.

2.6.1. Relative Participation of an Operational Premise to Successful SVS Events. Within all the successful SVS experiments of given query nature, are there any operational premises that are more often represented than expected? Since the operational premises are subject to a systematic scan, there are always 1/6 of SVS experiments using each descriptor set, 1/9 using each metric, and 1/54 associated to a same CS. If this ratio is significantly altered within the subset of successful SVS, it may suggest that some operational choices are more success-prone than others. This is related to economic market share analysis: out of the observed instances of *successful* SVS at a, how many are associated to a specific premise O?

$$R^{Q,a}(O) = \frac{\#S(O, Q, a)}{\#S(*, Q, a)} \times 100\% \tag{9}$$

Let the above R be denominated "relative success rate" of operational premise O over the query set Q at acceptance level a.

2.6.2. Absolute Percentages of Success. Alternatively, one may ask for the absolute percentages P—out of all SVS experiments over set Q—how many with operational setup O succeeded at acceptance a?

$$P^{Q,a}(O) = \frac{\#S(O, Q, a)}{\#S(*, Q, *)} \times 100\% \tag{10}$$

The sum over all alternative setups O of above P scores returns the average probability of SVS success with respect to queries in set Q.

2.6.3. Exclusivity Index. Eventually, the above success rates only reflect in what measure the use of a given operational premise may increase the chances to obtain a successful SVS. They do not evidence whether active analogues of a given query may be found with many, differently operated SVS (i.e., SVS is intrinsically easy), or whether it takes specific operational premises to achieve this. Therefore, an alternative exclusivity index W has been designed as follows. Consider each individual query (formally speaking, let q be the one "current" query, out of the members of Q). Then, $\#S(*, q, *) = 54$ since the total number of associated SVS experiments equals the number of scanned CS. Thereof, only $0 \le \#S(*, q, a) \le 54$ are successful. If successful instances exist and are all based on the same operational premise O, then $\#S(*, q, a) = \#S(O, q, a) > 0$. In other words, premise O is *necessary* for SVS success with q, for $\#S(O', q, a) = 0\ \forall\ O' \neq O$. If, for example, three SVS attempts with query q were successful: the CS being FPT1.T, treePH03.T, and seqPH37.T, then the necessary operational premise associated to q is the use of the Tanimoto score. There is no mandatory descriptor set, for three alternatives exist. The exclusivity index is defined as the ratio of queries within Q for which O is necessary, with respect to the total count of queries from Q for which at least some of considered setups O were successful. Below, # denotes the number of elements of the sets defined within curly brackets:

$$W^{Q,a}(O) = \frac{\#\{q \in Q | \#S(O, q, a) = \#S(*, q, a) \text{ and } \#S(*, q, a) > 0\}}{\#\{q \in Q | \#S(*, q, a) > 0\}} \times 100\% \tag{11}$$

It shows the fraction among successful SVS attempts that is strictly conditioned by using a given premise.

2.7. Exhaustive Similarity Searches in a 1.6 M Molecule Collection. Tversky and Tanimoto were used in parallel to search for analogs of seven randomly picked queries achieving an at least good SVS success level in some CS and representing targets of relevant classes (GPCRs of various types, nuclear receptors, kinases and other enzymes). Selected query compounds, together with their primary targets and the CS chosen for screening, are given in Table 3. The associated descriptors were not picked for having ensured SVS success of the query during benchmarking but chosen on the basis of the benchmarked preferences of different target types for different descriptors, like in any prospective virtual screening. The SVS experiments, piloted through the web interface at http://infochim.u-strasbg.fr/webserv/VSEngine.html, were run against the entire database featuring all the activity-annotated benchmarking sets hidden within a much larger collection of (at the time of the experiments) almost 1.6 million commercial molecules. The screening protocol uses Kohonen mapping to discard obviously remote candidates,[43] in order to speed up searching—a box of three layers of neurons around the neuron hosting the query was selected. Also, the web-based SVS protocol does not return the absolute distances to the analogue "hits" but compares them to the distribution of optimal dissimilarity cutoff values d^* associated to the Ξ^* within the currently employed CS. The actual hit-to-query dissimilarity score d is thus converted into a hit "goodness" class G, as follows: if d is within the top 10% of smallest benchmark d^* values, $G = 0$ (optimal), else if within the top 20% observed d^* then $G = 1$, etc. While the web site reports all hits of $G = 0-6$, here only top quality hits at $G < 3$ will be analyzed and discussed.

3. RESULTS AND DISCUSSIONS

3.1. Comments Concerning the Analyzed Data Set. ChEMBL sets certainly include analogue series, within which similarity searches are deemed to be easy. This is not a problem in our study. First, frustrating activity cliffs appear well within such series of strongly related molecules—so searches within a series are not necessarily easy. Next, such a collection of diverse sets of related compounds perfectly mimics the structure of a corporate database. Most important, however, the herein monitored SVS score is not limited by an arbitrarily defined "top N nearest neighbors". In the latter case, large series of trivially close analogs might occupy these "top N" monitored positions, totally obscuring the manner in which the active analogue search handles further, less obvious active analogues. The NB-based criterion may perfectly well deal with such situations. Being self-adaptive and providing a variable-size selection as the best (and tunable) compromise between purity and recall of actives, it will favor CS able to highlight "hidden" structural similarity, retrieving both the entire subset of trivial neighbors, plus a maximum of less obvious analogues. Of course, compounds sharing a same activity may but need not be

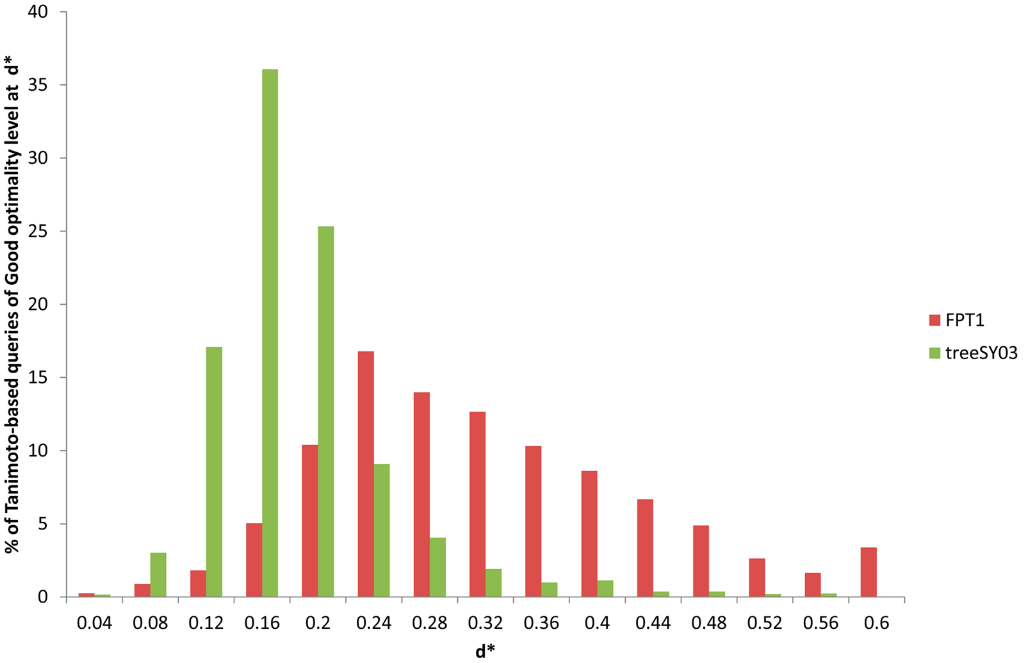


Figure 1. Distribution of the FPT1 and respectively treeSY03 descriptor-based Tanimoto-driven SVS simulations of good level with respect to their optimal dissimilarity radius d^*.

structurally similar: activity similarity should not be seen as implying some "hidden" structural similarity. The fact that, for each query, genuinely structurally unrelated actives cannot be retrieved in SVS, is not a failure of SVS−and, since these pairs are bound to fail in all CS, they do not impact on the benchmarking.

Furthermore, note that $\Lambda(M, m)$ is not a strict transposition of the similarity principle "similar molecules should have similar properties". It rather is an empirical expression of the medicinal chemist's point of view. An SVS starting with a micromolar query is a success when returning a nanomolar hit, but a failure if returning a millimolar, albeit the activity gap of 3 logs is the same. $\Lambda(M, m)$ is an asymmetric property similarity score, as required by chemical common sense. Inactive decoys m are set at $\Lambda(M, m) = 1$ with respect to active query M, like any other tested compounds m with $\mathrm{p}K_\mathrm{i}(m) < \mathrm{p}K_\mathrm{i}(M) - 3$. The fuzzy border $1 > \Lambda > 0$ between relevant actives and decoys does not coincide with the demarcation between the ChEMBL homogeneous series and the diverse ZINC decoys. Such SVS simulations are thus confronted with decoys from the same series as the actives: same size, same scaffold, and largely conserved pharmacophore pattern−a difficult scenario.

There were 4453 pharma-low queries, selected for having less than 570 populated pharmacophore triplets in the molecule. Reversely, the 3933 pharma-high queries feature >950 triplets each. Significant shifts from average complexity exist within target-specific subclasses. The 3290 monoamine GPCR binders (647 ± 285 populated FPT1 triplets) are significantly (Student[45] $t = 21.9$) simpler than average queries (779 ± 396 triplets, see Table 1). By contrast, the 4370 binders of the other GPCR classes are of above-average pharmacophore complexity (943 ± 400 triplets, $t = 23.5$). The 290 kinase inhibitors are slightly less complex (668 ± 204 triplets, $t = 8.9$) than the average. However, both monoamine GPCR and kinase inhibitors are nevertheless more complex than the pharm-low set.

3.2. Optimal Dissimilarity Radii.

As can be seen from the example in Figure 1, the answer to the key question of dissimilarity cutoff choice, "how similar is similar?", clearly seems to heavily depend on the specific query. Assuming the d^* dissimilarity cutoff maximizing the ascertained optimality criterion to be a meaningful choice for the Tanimoto dissimilarity radius, it can be seen that its value may swipe a range of [0,0.6]. In rare cases, reaching out to 60% of dissimilarity (40% of similarity) is necessary to ensure a significant active retrieval level, all while not overloading the selection with inactives. More important, the distribution of dissimilarity radii of a same metric is seen to vary significantly from one descriptor space to another (Tanimoto dissimilarity shown here, but this holds for all the others). This clearly shows, as already pointed out,[17] that shortcut assumptions like "impose Tanimoto (similarity) > 0.85, irrespective of descriptors" is not the universal solution to SVS. With FPT1, less than 10% of queries reach their optimality radius at $d^* < 0.15$ (the web server uses a cutoff of 0.163 to delimit the most stringent class of optimal quality hits $G = 0$). Ironically, Tanimoto > 0.85 is a perfect choice with atom symbol-based tree descriptors.

While various variants of binary hashed fingerprints may be more conservative in terms of distributions of their optimal dissimilarity radii, feature counts represent a quite diverse class of descriptors for which no dogmatic attempt to predefine any "default" cutoff should be undertaken. A careful choice of the dissimilarity cutoff is required to compensate for some specific artifacts (marked size dependence, sparseness, etc.) instead of adapting the similarity score.[46] For this very simple reason, we decided not to publish the cutoff sets used by the web server, for we do not wish to encourage any attempt to transfer these to other CS. They are available, upon request, to scientists interested in using the very same descriptors employed by us. We rather recommend a thorough calibration of dissimilarity radii—at least in a general context like the one illustrated here (focusing on good overall performance, all targets/queries

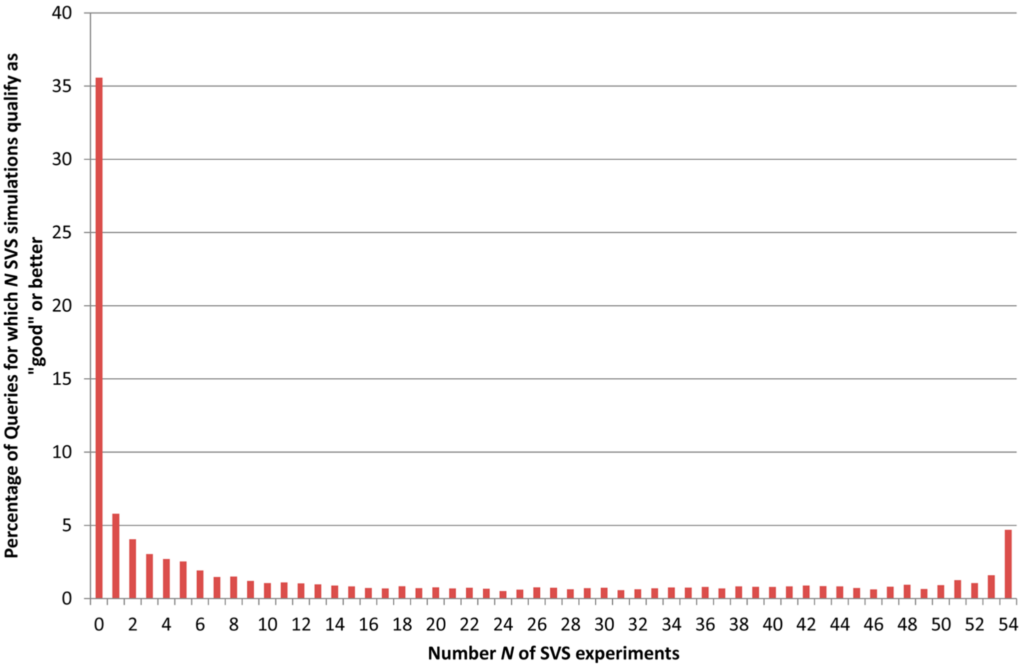


Figure 2. Distribution of queries with respect to their number of SVS simulations that qualify as good.

confounded). Even better, SVS may and should be tuned for specific contexts—targets with already known actives.

3.3. Mean SVS Success Rates, All Premises Confounded. When analyzing SVS experiments in terms of their optimality scores, it must be kept in mind that, for different SVS experiments, a same Ξ^* value may actually describe quite different SVS outcomes:

- Schematically, one extreme scenario (A) would rank candidates by similarity such as to feature only actives at the top of the list, some dissimilar actives at the bottom, and a "gap" of inactives in-between. The fewer actives at the bottom, the larger Ξ^*. However, if the target accommodates two genuinely different subfamilies of actives, then separation into top/bottom actives is expected. $\Xi^* < 1$ would, in this case, not signal a suboptimal choice of operational premises, but merely the user frustration at the plain fact that not all actives are eligible hits in active analogue searches.
- In the alternative scenario (B), an SVS may rank all the actives somewhere within the list head but mingling with inactives. Then, the dissimilarity radius will be extended to encompass this list head. Retrieval is perfect, purity is not: operational premises able to more selectively regroup actives score better. In scenario (B), $\Xi^* < 1$ is always a consequence of a suboptimal choice of CS.

However, one cannot distinguish between the two scenarios based on their Ξ^* value only, albeit Ξ^* gives more weight to purity over retrieval ($\kappa = 5$ has been kept as such, as previous studies showed that this is a reasonable choice). The actual quality of SVS[16,25,47]—including usefulness and originality of the hits—can be hardly squeezed into a number for benchmarking purposes. Ξ^* is not a perfect solution, but the best we may advocate so far. Studies cited in the Introduction typically focused one-sidedly on hit purity only,[32] or considered the overlap of dissimilarity distribution plots.[30,31,33] This latter assumes dissimilarity between two actives to be systematically lower than active−inactive scores and is therefore less well-suited to describe A-type SVS scenarios.

3.3.1. Global Mean SVS Success Rates. Out of the >700 000 SVS experiments, 66% are at least acceptable, 28% at least good, and 6.2% yield "excellent" optimality scores (see section 2.6). This is encouraging and outlines that ISIDA feature counts are chemically meaningful.

3.3.2. SVS Success Rates Per Query. The above averages include the fact that for some easy queries, the SVS may succeed several times, under various operational premises, whereas difficult queries will never be properly treated.

Alternatively, Figure 2 displays, as a histogram, the percentages of queries vs the number of SVS experiments (out of the 54 conducted per query) that qualified as good. It can be seen that, for 35% of queries, none of the 54 CS could support any good SVS campaign. On the opposite, for some 5% of queries, SVS was always good, irrespective of CS. If (results not shown) acceptable is chosen as success criterion, then only some 4% of the queries fail systematically, whereas 24% succeed indiscriminately. By contrast, barely 18% of queries manage to reach excellence status, if the CS is properly chosen, and 0.17% make excellent SVS departure points irrespectively of the CS.

3.3.3. SVS Success Rates Per Targets. At acceptance criterion "good", each target had at least one query for which at least one SVS attempt worked (a detailed histogram is available as Supporting Information). However, 65 targets out of 165 never manage to reach the excellent level. There are only 5 targets for which SVS excellence is very often encountered (between 60 and 80% of the cases—irrespective of queries and CS). Table 2 lists such easy targets, displaying at good level 80% or more of successful SVS (including the above-mentioned 5 of ubiquitous excellence, marked by asterisk). Not surprisingly, most of these are among the "exotic" targets with relatively few associated actives. They form narrow congeneric series in which actives are easily distinguished from inactives within the series—and the entire series is obvious to distinguish from the added decoys. Three of these targets are kinases (albeit only PIM1/PIM2 where included in

Table 2. Easy Targets Featuring Very High Success Rates of the Associated SVS Experiments, All Query and CS Choices Confounded[a]

target	ChEMBL ID	binder set size
gonadotropin-releasing hormone receptor	CHEMBL1855	241
inosine-5prime-monophosphate dehydrogenase 1*	CHEMBL1822	56
inosine-5prime-monophosphate dehydrogenase 2*	CHEMBL2002	56
interleukin-8 receptor B*	CHEMBL2434	102
matriptase	CHEMBL3018	63
motilin receptor*	CHEMBL2203	65
neurokinin 1 receptor	CHEMBL249	174
Neurotensin receptor 1	CHEMBL4123	57
PI3-kinase p110-alpha subunit	CHEMBL4005	68
prostanoid DP receptor	CHEMBL4427	105
purinergic receptor P2Y12*	CHEMBL2001	504
serine/threonine-protein kinase PIM1	CHEMBL2147	79
serine/threonine-protein kinase PIM2	CHEMBL4523	62
somatostatin receptor 2	CHEMBL1804	61

[a] More than 80% of their SVS simulations qualify as good. The targets labeled by "*" have between 60 and 80% of SVS reaching "excellent" levels.

the kinase subset, because the ChEMBL class of PI3 is plainly "enzyme"). The intriguing exception is the purinergic receptor P2Y12, with more than 500 associated actives. However, a quick glance at this data set revealed obvious activity-size dependence: the length of the SMILES string of these compounds is a variable that explains more than 40% of the pK_i variance throughout the series (larger being more active, as expected). It is thus clear why sophisticated molecular descriptors should be hardly necessary to discriminate stronger from weaker actives within this series. There are nine targets with significant activity−size correlation (at more than 40% of explained pK_i variance, in the above-mentioned sense). Two are easy targets of Table 2 (purinergic receptor and inosine−monophosphatase−dehydrogenase 1), but all reach high rates of at least 40% of SVS success at good level.

The above shows that this study, like all the others dedicated to SVS benchmarking, will inevitably be biased by the choice of molecules and targets. The above success rates cannot match real-life searches for not yet known actives, because of the seeding with congeneric series of actives. Intraseries discrimination between strong and weak actives is sometimes deceptively simple, over set sizes of hundreds. Nevertheless, this study covers the—to our knowledge—so-far largest panel of targets, and as large as possible series of certified actives with rigorously determined potencies. Most of the queries/targets have a balanced success rate in terms of SVS, which means that, for the vast majority of the 700 000 SVS experiments performed here, the operational premises **do** matter. Therefore, this study has been pursued. We see no direct interest of an in-depth statistical study of the nature of this bias, as the specifics here would not necessarily be relevant to scientists wishing to set up an optimized SVS tool, and hence bound to heed the specific sources of bias in their own collections. On a more general level, the differences between drug-like/lead-like and organic molecules were already thoroughly analyzed,[48,49] and similar trends apply to this collection.

3.3.4. Mean SVS Success Rates Per Query Subsets.

A first interesting conclusion on the basis of mean success rates over each considered subset Q of queries is that complex molecules are much more successful SVS queries than simple compounds. 43.6% of all the SVS attempts starting with a pharma-high molecule are good, and 15.3% are even excellent, compared to only 14.8% of good and 2% of excellent with pharma-low. This makes sense: similarity to a feature-rich compound is more difficult to achieve, as there are more features to be matched. The probability of a fortuitous return of a low dissimilarity score decreases. The more complex the pattern of a ligand, the more specifically it relates to its targeted binding site—for obvious economic reasons, no medicinal chemist will produce a complex ligand with its largest moiety dangling out of the binding site. Matching a complex chemotype is therefore a clear indication of belonging to the "family" of ligands of the target. The actual challenge of SVS under these circumstances remains to discriminate between strong and weak actives of the series.

In terms of target families, "other" (than Rhodopsin-like) GPCR binders also score a significantly above-average SVS success level. Six of the 44 "other" GPCRs are listed as easy (Table 2), yet many of these feature some of the largest associated compound sets, in which intraseries active/inactive discrimination plays a dominant role. With the notable exception of purinergic targets, fortuitous size-activity correlation is not an issue here. Eventually, the outstanding behavior of SVS over this set including some reputed difficult targets in drug design can be tentatively attributed to the very high complexity of the queries, in agreement with the discussion above.

Monoamine rhodopsin-like GPCR binders, also including many compound-rich targets, are by contrast, of under-average complexity and, coincidentally, significantly less successful in SVS. At this point, it cannot be argued whether low complexity is enough to explain this trend. It should be kept in mind that many of these targets have a significant binding site similarity and often share common ligands. Members of series designed for a given target have been systematically tested on related GPCRs, in order to address selectivity issues—therefore, cross-affinity data is more likely available within this class. Herein listed actives are very likely issued from various chemical series, while other less well studied targets in ChEMBL are more likely to be represented by homogeneous compound sets. This cross-affinity however extends to some receptors classified as nonrhodopsin-like (muscarinics, opioids), which also bind "bioactive amines". The detailed analysis of the structural reasons for the under-average SVS performance of monoamine GPCRs is however beyond the main scope of this work.

Intriguingly, the set of the seven considered kinase targets exhibits extremely high success rates, in spite of their low complexity. Two of the seven kinases are listed as easy targets. Kinases are "strict" targets, selective not only in terms of the chemotypes accepted as ligands but requiring well-known specific ligand−site interactions as absolute prerequisites for activity. One may expect such specific structural signatures being easily discriminated from random compounds. However, activity cliffs—small structural changes triggering dramatic activity shifts—are therefore common in the kinase activity landscape, and expectedly should make SVS difficult. Unfortunately, targets are—within the herein used, curated selection of ChEMBL—not quite compound-rich, forming the statistically most vulnerable subset of this study. Easy discrimination of the actives vs random decoy compounds is probably the paramount factor in these SVS.

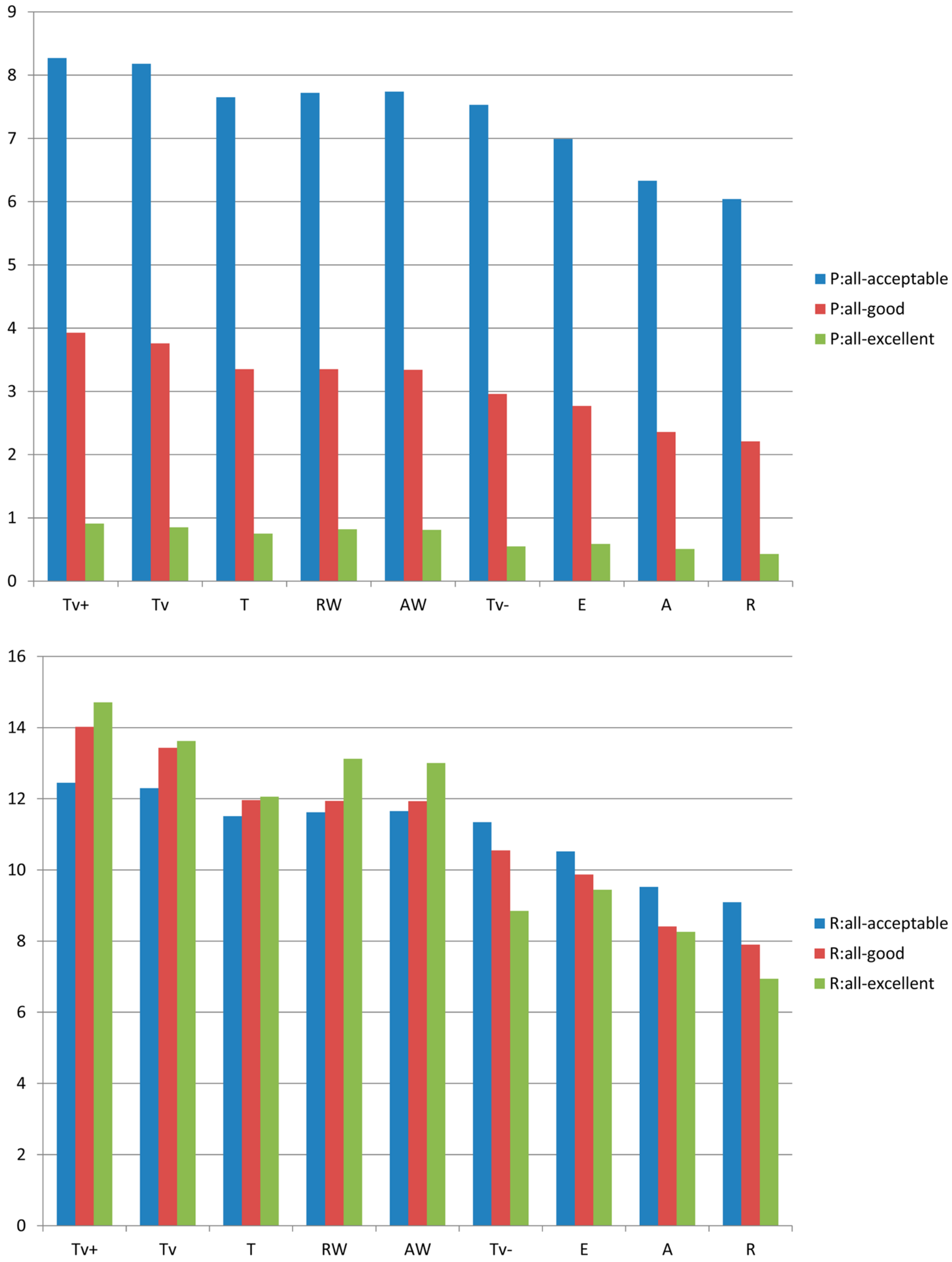


Figure 3. Absolute (above) and relative (below) success rates, at the three considered acceptance levels, of SVS experiments based on a given metric; all queries confounded.

3.4. SVS Success Rates Conditioned by the Choice of the Dissimilarity Score. The following discussions will focus on bar plots of the indices defined in section 2.6 (Absolute percentages of success *P*, relative participation to success *R*, and exclusivity index *W*) associated to premises listed on *X*. The *Y* axis units are always "%" (not shown). These average percentages are based on hundreds to thousands of items/sample. The exact number of contributing observations depends on the particular focus. For example, the average success rate for "SVS experiments using Tanimoto, all targets confounded" is based on a larger pool of members than the one of "SVS experiments using Tanimoto against kinases in FPT1 space". Larger observation pools are synonymous to better convergence toward expectation values—unless a larger observation pool mixes up divergent, local trends (putatively leading to a multimodal distribution). An in-depth statistical analysis of the behavior of reported average success rates goes beyond the purpose of this study. Instead, we have adopted a more pragmatic point of view, similar to the cross-validation approach in QSAR: if we would have disposed of only half of the available data, would we have been led to the same set of final conclusions (some fluctuation of the relative bar heights notwithstanding)? If true, then the really important question: "Will the herein conclusions still stand if more data is added to the study?" is not proven wrong and may be adopted and further challenged. An estimation of standard deviations of *R* values by monitoring their variation when calculated on randomly picked subsets of half size showed that averages

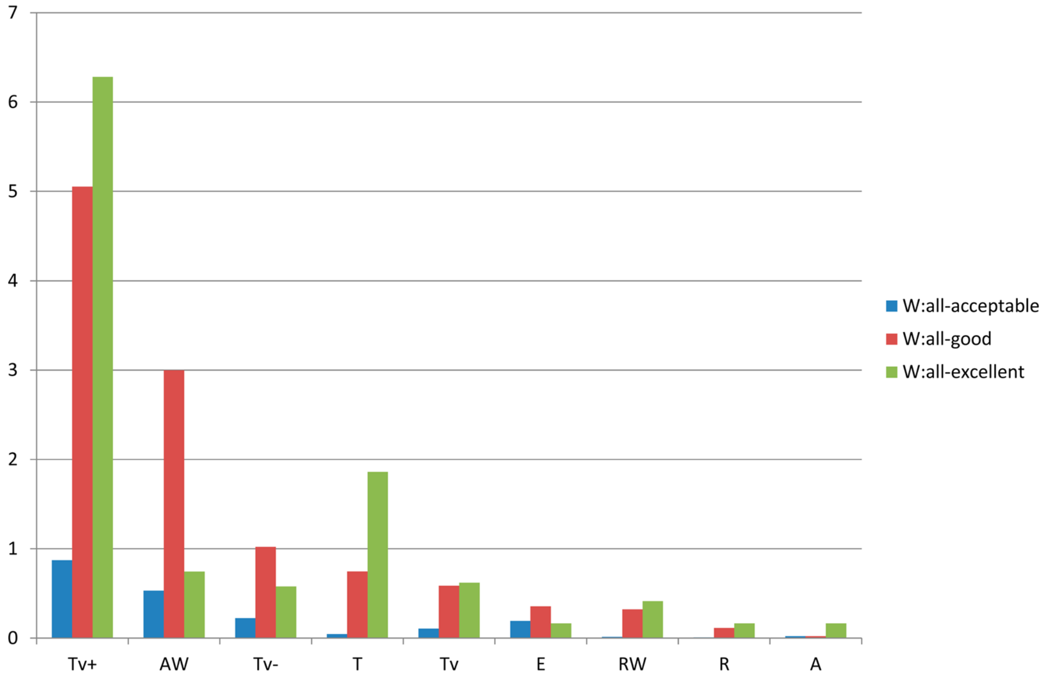


Figure 4. Exclusivity indices of metrics, at the three acceptance levels; all queries confounded.

typically remained stable within 0.1%. At such levels, there is no risk of reversing any of the important trends discussed here due to statistical noise. Therefore, the plots will not be charged with additional error bar information. Conservatively, one may consider any shift of more than a half percent as significant.

3.4.1. Overall Trends. Figure 3 displays absolute (above) and relative (below) success rates, at the three acceptance levels, over all SVS instances using a given metric, all queries confounded. The two plots are related—in the upper, the sum of absolute success rates (bars of given color) equals the global rate at corresponding acceptance rate in Table 1. In the lower, bar heights are rescaled with respect to this sum, expressing the "share" of a given metric within the "market" of successful SVS runs, after norming with respect to "market size" (bars of same color sum up to 100%).

The direct interest of the upper plot is providing an absolute probability of SVS success conditioned by the use of a metric (equaling bar height × 9—see section 2.6.2). Therefore, almost 36% of SVS runs based on the strongly asymmetric Tversky score Tv+ are good, over only 28% on average over all metrics, etc. Relative success rates are more enlightening for direct benchmarking purposes. Since each diversity score was given an equal number of opportunities to produce successful SVS, its effective share among succeeding SVS runs is an expression of its relative proficiency.

First, the R plot from Figure 3 shows that the relative proficiencies of the metrics are not strongly affected by the SVS acceptance level. They are ordered by their good success rates. Ordering by excellent would only shift the Tanimoto score two blocks to the left. It is interesting to note that, for the five best metrics (Tv+, Tv, AW, RW, T), the R scores also increase with the acceptance threshold (blue < red < green)—they are not only able to averagely increase success chances but actually have a larger involvement in the top success stories. The Tversky scores at either $\alpha > 0.5$ are dominant, and especially the strongly biased configuration at $\alpha = 0.9$. The weighing strategy of average fingerprint differences has an excellent impact on metric quality, matching or exceeding the performance of the Tanimoto "golden standard". This latter may be considered as representative of the symmetrical Tversky ($\alpha = 0.5$—strictly speaking, that would be the Dice metric). If so, a robust monotonous performance decrease with decreasing α can be observed: Tv+ > Tv > T > Tv−. Euclidean and plain average fingerprint differences are the least successful.

Analysis of exclusivity indices (Figure 4) provides some alternative insight. At low acceptance levels such exclusivity is quite low. Most queries allowing acceptable SVS typically allow more than one, based on different metrics. The higher the success threshold, the more exclusivity-prone the complying simulations are. Roughly 7% of targets yielding excellent SVS results may reach these either with Tv+ or not at all (as far as coverage by other metrics goes). Exclusivity is not correlated with relative success−in particular, the Tv score is the second most successful, but rarely exclusive. Reason−it is highly correlated, yet systematically less proficient than Tv+. Most of the queries having successful Tv-based SVS runs likely have even better Tv+-driven simulations. Yet, Tv does not completely "overshadow" Tanimoto in the above-mentioned sense−the latter seems to dispose of a specific "niche" in which it retains some exclusivity in terms of excellent SVS. Likely, AW has an applicability domain in which it dominates, at good level. More on this will be revealed in subset-specific, detailed analysis.

3.4.2. Relative Proficiency of Metrics with Respect to Query Complexity. In the following, the discussion will focus on the good acceptance level as SVS success criterion−the most representative. Figure 5 compares the R scores of metrics with respect to query complexity (top), while its bottom reports the associated exclusivity indices W. Both concur to highlight that

(a) the relative ranking in terms of metric proficiency is largely the same as witnessed over the entire set, irrespective of query complexity, but

(b) high query complexity renders SVS less sensitive to the choice of metric.

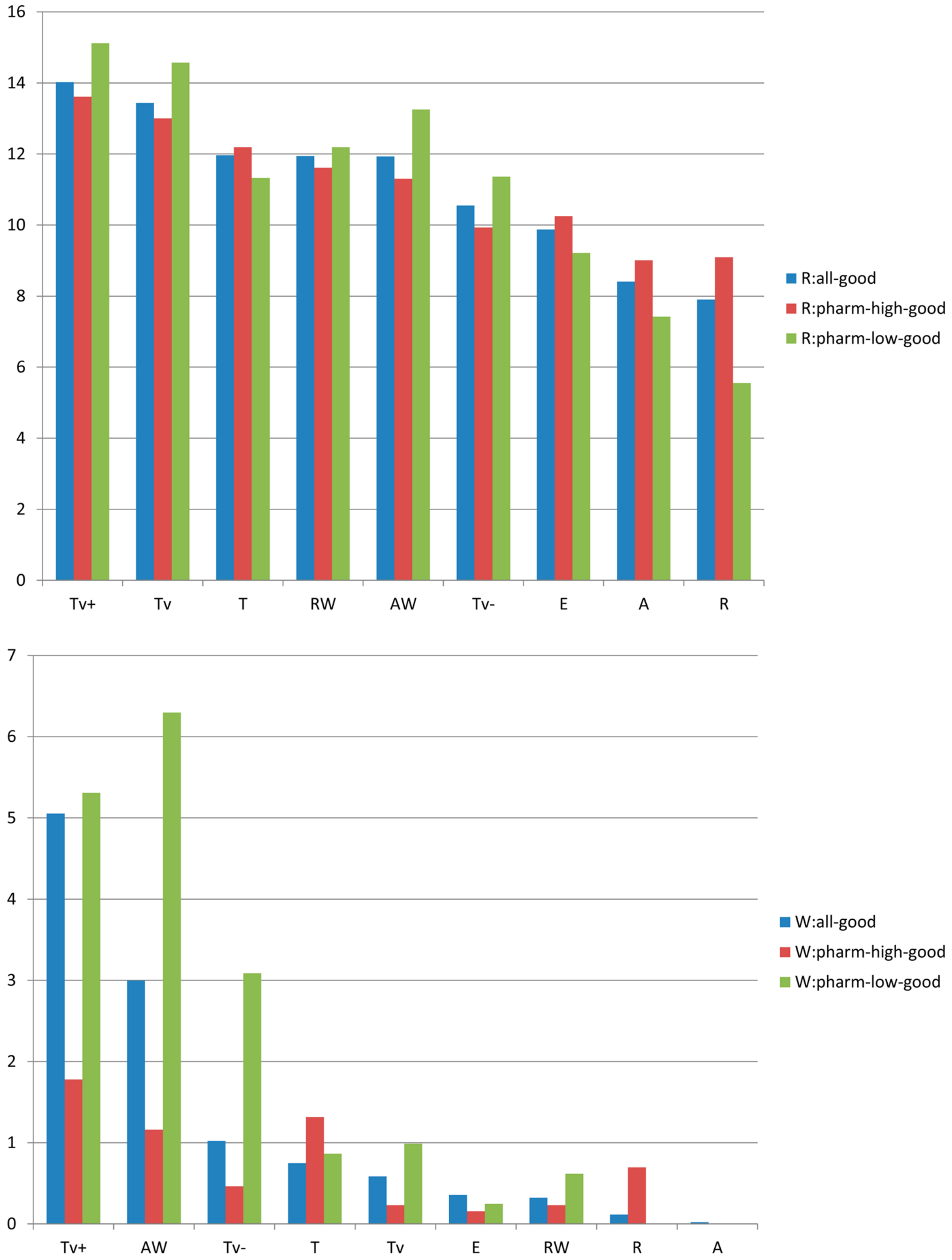


Figure 5. Relative success rates (top) and exclusivity indices (bottom) at good acceptance level, of SVS experiments based on a given metric, with respect of the complexity of their queries (overall levels shown as witness).

Albeit Tv+ and Tv are still dominant, while *E*, *A*, and *R* are weaker performers, the red pharma-high bar plot in Figure 5 (top) is significantly flatter than pharma-low. Coherently (bottom plot), there is less opportunity for a metric to be the exclusive condition of SVS success when query complexity is high. It is now possible to see that the exclusivity “niche” of the AW score highlighted by Figure 4 contains low-complexity queries.

The decrease, at high complexity levels, of the relative advantage of Tv+ and Tv over plain Tanimoto does make sense: with a feature-rich query, feature loss is, physically, less penalizing. If the query is compact (near-spherical), but feature-rich, as the number of features interacting with the receptor scales like the query−site contact surface, but their total number scales like its volume−the probability to have lost a key feature becomes intrinsically lower. If the query is of linear shape and binds in an extended cleft, then the loss of a key feature contributes less to activity loss, as the anchoring happens in many points. By contrast, losing a key feature out of few is more likely to trigger activity loss−hence, increasing query feature loss penalties by increasing Tversky’s α makes more sense in simpler queries.

Intriguingly, Tv− relative proficiency also improves with low complexity queries. Shifting the bias from extra penalty for query feature loss ($\alpha > 0.5$) to equal penalty ($\alpha = 0.5$) for either feature loss or extra feature presence in the candidate, leads the symmetric Tanimoto score to perform less well at low complexity, in line with the above discussion. However, the

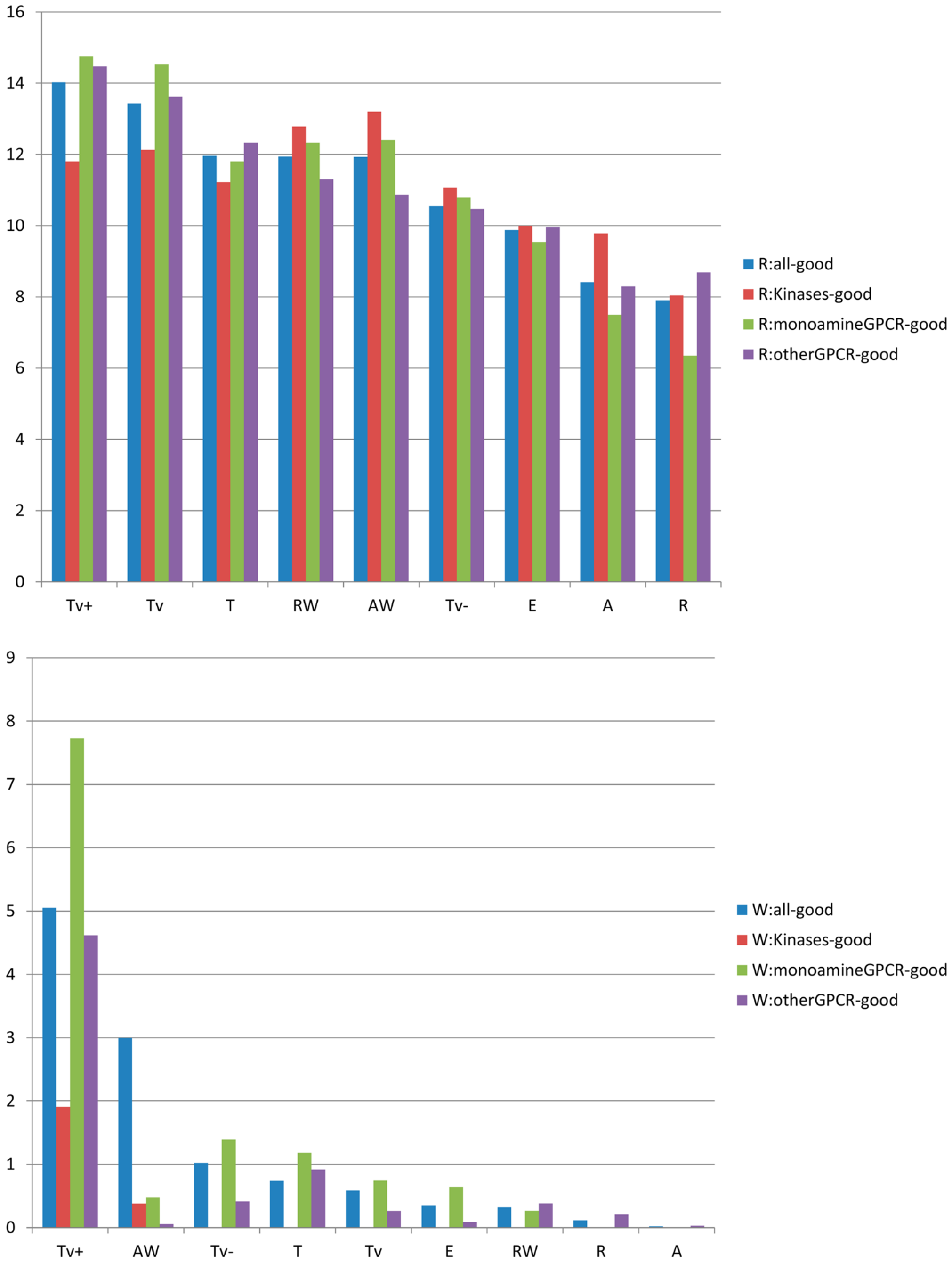


Figure 6. Relative success rates (top) and exclusivity indices (bottom) at good acceptance level, of SVS experiments based on a given metric, with respect of the target families of queries (overall levels shown as witness).

trend reverses if the bias is now pushed in the opposite direction: extra penalty for "new" features in candidates, not matched by anything in the query. Tv− is a dissimilarity score that could potentially allow to "decomplexify" a query, by prioritizing less feature-rich analogues, as hinted in the literature.[31] Surprisingly, its relative success with respect to other metrics improves when the queries are *already* of low complexity. For example, the Tanimoto score is 1.23 times more successful (a *T*-driven SVS is 1.23 times more likely to reach good level) than Tv− over high complexity queries, but only 0.99 times over pharma-low queries. Tv+ remains more successful than Tv− at low complexity, but less markedly so (1.32× vs 1.37× over pharma-high). Furthermore (Figure 5 bottom), the gain of Tv− is exclusively achieved with respect to specific pharma-low queries.

These observations could be sketchily explained by considering two categories of low-complexity queries. One class regroups those for which loss of features in the query likely implies an activity loss, but which may allow for more complex active analogues—they could be entitled "growable" queries. Growable queries are best served by the Tv+ score. At the opposite, "conservative" queries are also likely to lose activity upon feature query loss but are even less tolerant to the presence of novel features in the candidate. These may represent ligands matching perfectly complementary active sites, in which no additional functional group could be accommodated. Actually, out of the 89 queries for which

Tv− is the exclusive metric leading to good results, 21 are inhibitors of the norepinephrine transporter (CHEMBL 222). Evolved as a carrier of a well-defined bioactive amine, this could be a good example of a Tv− conservative target. Other twelve targets witness Tv− exclusivity with respect to more than one but less than six queries each. Their qualification as conservative is far less certain—nevertheless they include the related dopamine transporter (CHEMBL 238) and the 5HT-2, D1, and H3 receptors.

3.4.3. Relative Proficiency of Metrics with Respect to Target Classes. For kinases, the smallest target family of this study, *R* and *W* profiles of the various metrics from Figure 6 intriguingly resemble the high-complexity profile previously discussed. Albeit these queries are not among those of highest pharmacophore complexity, kinases seem to have a less marked preference in terms of metrics. Interestingly, the weighted average differences AW and RW are preferred by a narrow margin (AW appears only 1.12× more successful than Tv+). Furthermore, AW and RW are quite covariant, neither showing a clear advantage/exclusivity. The quite compact families of kinase inhibitors of known K_i are likely to encourage B-type SVS scenarios, as outlined in the introductory comments of section 3.3. Segregation between series members and external decoy molecules is enough to ensure a good SVS level−indeed, recognition of the series would return a set in which roughly one of two molecules is highly active (excellent purity and perfect retrieval rate). SVS operational premises are expected to compete in terms of their ability to make a finer distinction between actives and inactives within the series. However, they virtually all fail in this respect−the fine SAR of the rugged[50] kinase landscape cannot be explained in terms of plain similarity: key features must be identified and weighted accordingly. Therefore, irrespective of the employed dissimilarity score, all kinase-based runs return roughly equally successful B-scenario SVS runs. This is in agreement with the observed *R* and *W* profiles of metrics and also (vide infra) the equally flat, insensitive profile with respect to the descriptor choice.

The other considered target classes display profiles similar to the overall average. The "other" GPCR family, regrouping high complexity queries, has a signature comparable to the pharma-high subset. Tv+ and Tv are the dominant metrics, irrespective of the targets.

3.5. SVS Success Rates Conditioned by the Choice of Descriptors.

Considered descriptors were chosen to offer a diverse range of perspectives of viewing the molecular structures, both in terms of fragmentation strategy (sequences, circular fragments, triplets), fuzziness (pharmacophore point fuzziness in triplets, but also structural "fuzziness" by ignoring the exact nature of intermediate atoms between the center and the outer level in tree fragments), and coloring (element type, pharmacophore). Since descriptor benchmarking has already been addressed,[3,20] and the present study basically reinforces previous finds, the detailed discussion and histogram plots will be made available as Supporting Information. Only the main conclusions will be highlighted here:

These are a representative subset of descriptors witnessing excellent NB, and there is no overwhelming advantage of either premise over others. Again,[20] the tree descriptors emerge as winners, with pharmacophore triplets and the less sophisticated sequence counts seqSY37 being somewhat less successful.

In terms of exclusivity indices, symbol-colored tree counts treeSY03 are dominant. Trees are information-rich fragment counts—specifying the decomposition of a molecular graph in terms of symbol-colored trees provides an extensive description of the chemical context of every atom and seems sufficient to implicitly fix the pharmacophore types. Explicit pharmacophore typing with tree descriptors seems to provide some limited benefits.

High complexity queries appear less sensitive with respect to the choice of descriptors. This is coherent with the reduced sensitivity to metrics (section 3.4.2): similarity of two complex patterns is a more obvious indicator of comembership to a same activity class therefore more robust with respect to its operational premises.

If the metric-insensitive behavior of kinase-specific SVS runs is correctly explained by the "family recognition" scenario B, then—since all off the herein used descriptors are able to recognize structurally homogeneous series—there should be no significant dependence of kinase SVS success on the descriptor choice either. This is precisely the case—with a notably decreased performance of fuzzy FPT1 which, given their propensity for scaffold hopping, are more likely to allow some of the external decoy compounds enter the selection. By contrast, the strict symbol-based sequences are the optimal "guardians" of class purity.

There is no specific common trend—other than the one already associated to a complexity increase—over the diverse family "otherGPCR". A special affinity of rhodopsine-like GPCRs and the FPT1 could be evidenced. Many queries of this category only succeed with fuzzy triplets and represent a significant part of the exclusivity niche of these descriptors. Indeed, the classical aromatic ring−cationic group signature of bioactive amines can be conjugated in terms of various aromatic ring sizes and/or ring-charge spacer length, without losing GPCR affinity.

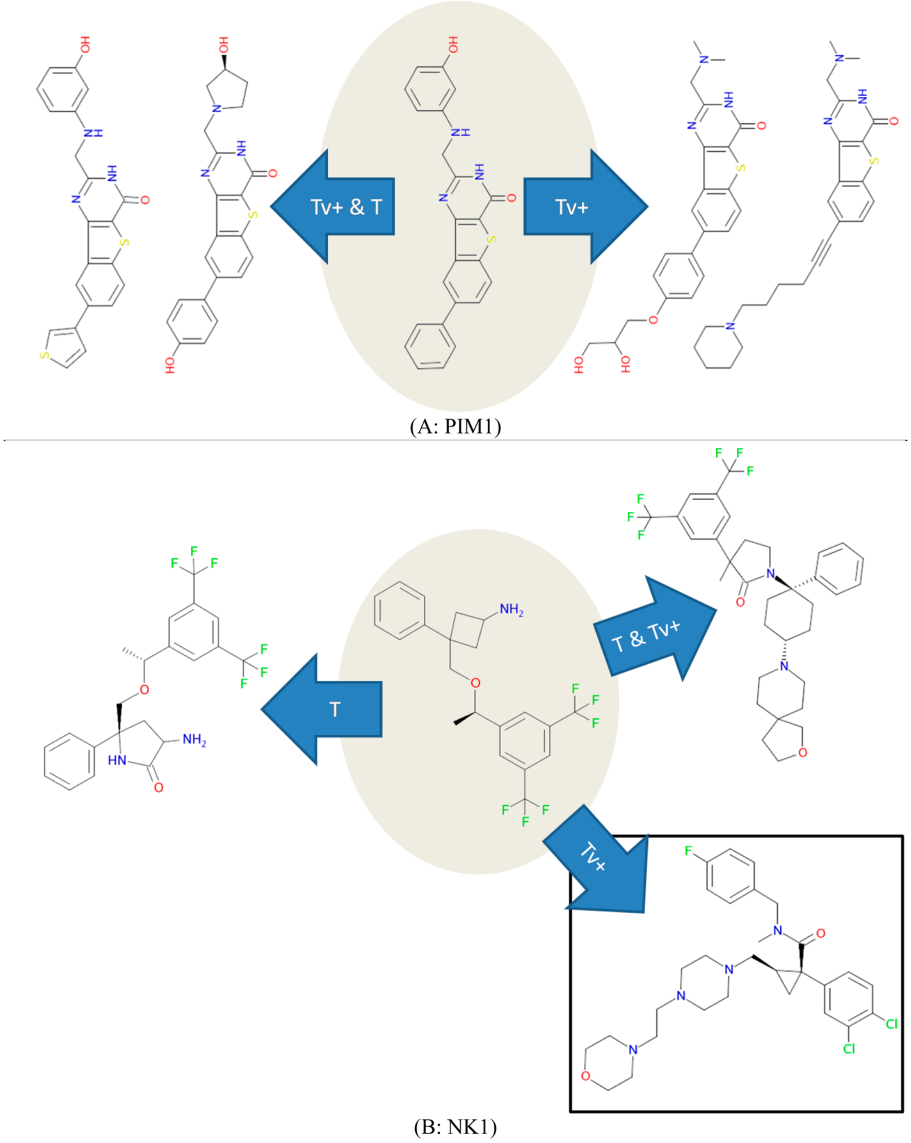
3.6. Exhaustive Similarity Searches.

The results in Table 3 are still too optimistic with respect to prospective SVS. Exhaustive searches are much more challenging than simple benchmarking, for various reasons:

Table 3. Queries Subjected to Exhaustive SVS against 1.6 M Molecules, via the Web-Driven Protocol[a]

query CHEMBL code	target	descriptors	score	hits	hits on target
1258223	serotonin 4 (5-HT4) receptor--CHEMBL1875	treePH03	*T*	26	17
			Tv+	42	41
195	muscarinic acetylcholine receptor M4--CHEMBL1821	FPT1	*T*	28	1
			Tv+	83	1
212428	neurokinin 1 receptor CHEMBL249	FPT1	*T*	24	22
			Tv+	11	3
1170849	inosine-5prime-monophosphate dehydrogenase 2 CHEMBL2002	treeSY03	*T*	9	9
			Tv+	28	21
382542	thrombin CHEMBL204	treePH03	*T*	352	11
			Tv+	84	11
573011	serine/threonine-protein kinase PIM1 CHEMBL2147	seqSY37	*T*	254	24
			Tv+	294	44
1672548	progesterone receptor CHEMBL208	treeSY03	*T*	62	7
			Tv+	20	7

[a] "Hits" refers to the total number of hits of goodness classes 0−2, out of which "hits on target" are known actives against the target.

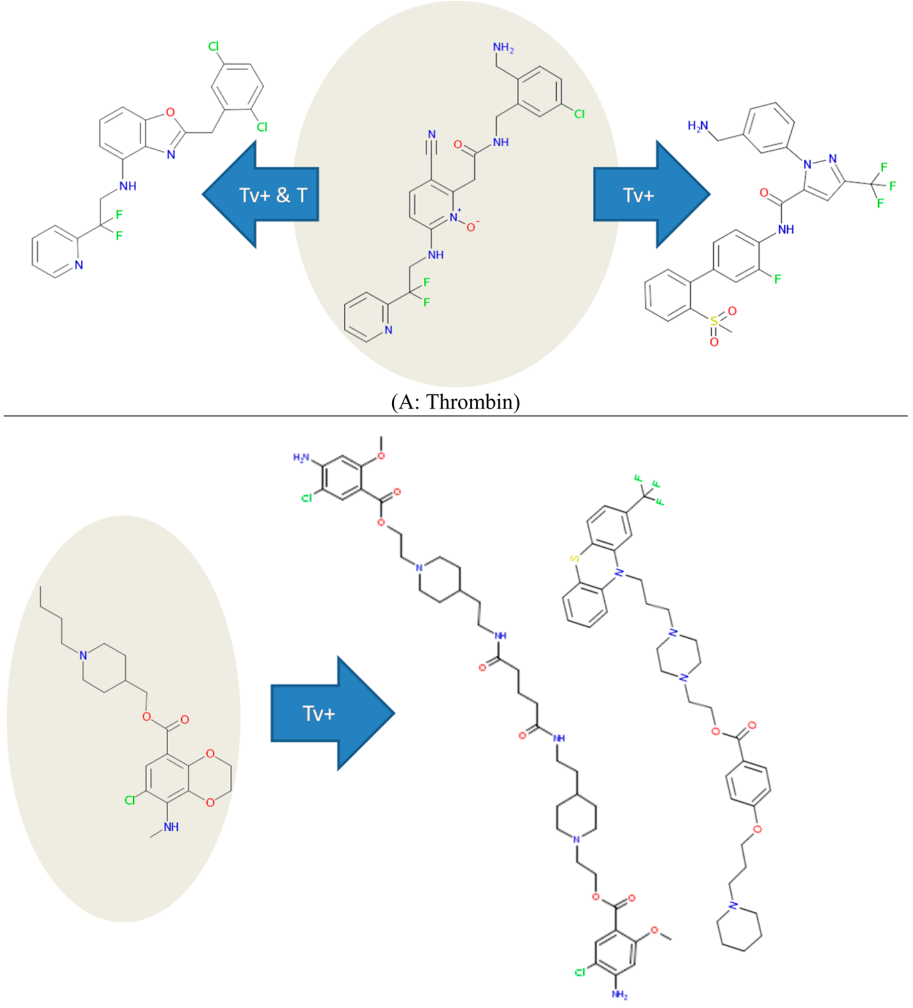


(A: PIM1)

(B: NK1)

Figure 7. (top; A): Examples of active analogues of the central PIM1 kinase ligand, found by both Tanimoto and Tversky (left) or exclusively by the Tversky score (right). (bottom B): Active analogues of the central Neurokinin 1 ligand, found by Tanimoto only and by both Tanimoto and Tversky. The boxed analogue bottom right is selected by Tversky only but is registered as a neurokinin 3 binder—no information about its potential NK1 affinity lies at hand.

- The "decoy" set, represented by >1.5 million commercial molecules is incomparably more rich and diverse than the 10 000 benchmarking decoys.
- In benchmarking, ChEMBL molecules binding to related biomolecules, but of no known K_i to the current target, did not participate in benchmarking. Presently, the union of *all* sets of actives is accessible to *any* query, and such, likely promiscuous, binders, by default, count as "inactives". Such situations have been followed up manually, and the entire ChEMBL base was interrogated in order to check whether activity data other than K_i were reported for the current target—if so, their status was changed to "hits on target" in Table 3.
- As established, using Tv+ on a complex query returns high-complexity SVS hits—and these were, indeed, "family-own" actives by contrast to less complex decoys: hence the great success of Tv+ on pharma-high. Now, having the entire lot of >30 000 annotated ChEMBL actives among candidates means that high complexity is no more synonymous to family membership. Will Tv+ preferentially retrieve any complex ligands, irrespective to their actual activities? Such a trend was indeed observed (on Neurokinin 1, notably—but also for the hits of lower goodness classes 3–6, not shown). Nevertheless, in terms of top hits, the Tv+ selections actually show higher purity rates (no. hits on target/total no. hits) in four out of the seven reported cases.
- Eventually, a specific bias against Tv+ could represent the fact that, prior to actual SVS, candidates are prescreened according to their position relative to the query, on a prebuilt Kohonen map. Or, the original study had highlighted that, albeit Kohonen mapping operates on the basis of Euclidean distances, molecules similar in terms of Tanimoto scores are consistently expected to "inhabit" neurons neighboring the residence of the query. This is a much less obvious working hypothesis with a



(A: Thrombin)

(B: 5-HT4)

Figure 8. (top A): Active analogues of the central thrombin binder, jointly found by both Tanimoto and Tversky (left) and exclusively by Tversky (right). (bottom B): Much larger, and yet active "analogs" of the given 5-HT4 binder, which may only be uncovered by a highly asymmetric dissimilarity score, biased in favor of query feature conservation.

highly asymmetric Tv+. A rather large area including the three surrounding neuron layers has been selected for screening—and this proved sufficient for the majority of reported cases.

In three outlined cases—progesterone, muscarinic receptor, and inosine dehydrogenase—there is no opportunity for scaffold hopping. Both *T* and Tv+ basically return identical hits, of same structural families as the queries. Tv+ ensures a higher purity for progesterone, and a better retrieval rate for inosine dehydrogenase. This is likely a classification artifact: with Tv+ additional hits made it into the considered $G = 2$ category, while they were relegated to $G = 3$ in the Tanimoto-driven SVS.

The muscarinic query is the apparently worst overall result, for both Tv+ and *T*. The hit list is however populated by bioactive amine-like molecules, binders to the plethora of rhodopsine-like GPCRs, displacing reported active muscarinic compounds to $G > 2$. Such are often seen to have muscarinic side effects, albeit no explicit information in this sense could be found in ChEMBL.

PIM1 kinase does not allow for scaffold hopping either—at least not in the strict sense. However, Figure 7A shows that Tv+ exclusively supports the "discovery" of ligands featuring significant topological and pharmacophore variations among the substituents of the conserved central scaffold.

In terms of hit rates only, Tv+ is less successful with Neurokinin 1 (NK1), where it oversamples more complex molecules in detriment to congeneric analogs, exclusively recognized by Tanimoto (Figure 7B). Both *T* and Tv+ succeed in scaffold hopping, identifying the upper right, alternative chemotype. Tv+ pushes scaffold hopping even further, to exclusively highlight the lower right (boxed) binder of related NK3, for which no NK1 activity was reported. This is an example where Tv+ and *T* should have been used alternatively in order to maximize the coverage of potentially interesting hits.

The same trend to more audacious scaffold hopping of Tversky vs Tanimoto, yet this time backed by significant "discovery" of genuine actives of different chemotypes, can eventually be witnessed for both the Thrombin (Figure 8A) and the 5-HT4 serotonin receptor (Figure 8B) ligands. This is in particular quite dramatic in the 5-HT4 example, where the first illustrated hit ($G = 2$) is virtually a dimer of the query. Since it contains basically all query features, the additional fragments in the "dimer" did not push the Tv+ dissimilarity outside the relevance range—correctly so, for these dimeric forms are as active. The second illustrated hit ($G = 2$ as well) is a genuine scaffold hopping example.

4. CONCLUSIONS

This study is an exhaustive analysis of the NB principle, the foundation of SVS, analyzing a data set as large as could be accessed publicly, all while ensuring that the biological activities are of high quality and obtained according to compatible protocols, thus strictly comparable. It considered SVS success as a weighted compromise between purity and retrieval rate of active "hits" in the neighborhood of an active query, all while acknowledging that retrieval rate may be objectively low−because similarity to any known active is not a prerequisite to activity. The study has highlighted various key aspects of SVS, such as behavior with respect to query complexity and target classes.

On the basis of more than 700 000 distinct SVS experiments, some robust insights emerged:

- A double perspective has been employed to understand the relative proficiency of dissimilarity score—"relative success" (R) and "exclusivity" (W). Yet, from both perspectives, the most successful score (globally, as well as within specific query subsets) is hardly used in practice, in spite of recent theoretical studies dedicated to it: Tversky at $\alpha > 0.5$. This should change.
- The uncertainty about the optimal α value should not be a deterrent against the use of the Tversky score. There is no need to fine-tune it—both default settings 0.9 and 0.7 gave top results, the former being better. Tanimoto ranks as average, being also seriously challenged by the here-introduced weighted average fingerprint deviations and fails to defend its status of today's "default" SVS score.
- Tversky-driven searches were exemplified in terms with good practical results, in realistic exhaustive SVS experiments in a million-range compound database. Tversky (at herein advocated parameter value) is a promoter of scaffold hopping in SVS runs, as could be seen from the reported SVS examples. If used in parallel to symmetrical dissimilarity scores (or if the screened database is prefiltered in terms of complexity), its tendency to oversample more complex analogues can be easily counterbalanced.
- An objective approach deciding "how similar is similar" when using highly asymmetric Tversky scores, based on comparing the actual query−hit dissimilarity to typical optimal dissimilarity cutoffs met in this NB study, should allow nonexpert users to apply Tversky-driven SVS (http://infochim.u-strasbg.fr/webserv/VSEngine.html) as easily as any Tanimoto-driven search, over a significant chemical space (1.6 M commercial and activity-annotated compounds) without having to rely on some absolute cutoff.
- Intriguingly, and in disagreement with the expectation[31] that Tversky at $\alpha < 0.5$ could serve to "decomplexify" large actives in order to discover smaller and pharmacokinetically more compliant leads, Tv− was *not* seen to emerge as the natural choice to use with top complexity queries. On the contrary, Tv− stood out with respect to above-mentioned conservative targets. As far as this study can tell, the best way to find active hits of lower complexity than a given query is to preselect the set of candidates at desired lower complexity level, then apply Tversky at $\alpha > 0.5$ to pick virtual hits minimizing the number of query features lost.
- Feature count descriptors are well NB-compliant and interpretable. As counts of actual fragments or pharmacophore group patters, which may or may not directly serve to anchor the ligand to its target, they are perfectly well-suited for use with asymmetric dissimilarity scores. They provide a straightforward illustration of molecular complexity.
- Tree descriptors are largely dominating in terms of SVS success, providing the globally best compromise between explicit monitoring of the topology and fuzziness introduced by ignoring the nature of the atoms connecting their center to the outer shell atoms. However, rhodopsin-like GPCRs binding small bioactive amines characterized by a rather fuzzy aromatic-cation pharmacophore pattern are better treated by fuzzy pharmacophore triplets. At the other end of the spectrum, kinases are best dealt with by atom symbol-colored fragment counts.
- Last but not least, this study highlighted once more that optimal dissimilarity radii in SVS should not be considered as metric-specific, universally transferable "Tanimoto >0.85" values to be used irrespectively of the descriptor space. On the contrary, they should be carefully reassessed—at minimum, for each CS (descriptor/metric combination), or, even better, tuned for specific targets and ligand classes.

This paper describes an empirical, robust approach to configure a web server providing a coherent similarity screening tool. Making absolute statements about statistically sophisticated criteria describing metric or descriptor performances is less important to us, knowing that various sources of bias are inevitable, irrespectively of considered compound collections. We encourage other groups, advocating different classes of descriptors and/or metrics, to employ similar procedures in order to set up complementary similarity search tools.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org:

- List of the 165 covered targets, associated number of compounds and considered number of queries, as well as the family based target subsets (ci400106g_si_001.xlsx)
- All the considered target-compound pairs (ci400106g_-si_003.txt), featuring the current target ID (column 1), the ChEMBL target ID, the ChEMBL compound ID, and eventually the current compound ID (column 4).
- A zip archive of a list of successful SVS instances at given acceptance level, per query subsets: succ.query-subset.acceptance-level.txt: four-column files featuring the used CS, concerned *current* target ID, *current* query compound ID (please refer to ci400106g_si_003.txt for current-ChEMBL "translation") and the rank of this SVS experiment over the set of all experiments associated to that query (i.e., "1"−the associated CS was the one returning the best Ξ*score for that query. If a target−query combination is not listed in a file at acceptance *a*, it means that its best Ξ* score, all CS confounded, was below *a*).

- An additional document, with plots and discussions left out from the main text (ci400106g_si_004.pdf).

AUTHOR INFORMATION

Corresponding Author

*E-mail: dhorvath@unistra.fr.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The Laboratory of Chemoinformatics wishes to thank the High Performance Computing center of the University of Strasbourg for assistance. Many thanks to Prof. Jürgen Bajorath, for providing the clean and coherent ChEMBL compound subset.

ABBREVIATIONS

AUC, area under curve; CS, chemical space; FS, false similars; GPCR, G-protein coupled receptor; NB, neighborhood behavior; PFD, potentially false dissimilars; PS, property space; ROC, receiver operating characteristic; SVS, similarity-based virtual screening

REFERENCES

(1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley Interscience: New York, 1990.

(2) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Model* **1998**, *38*, 983−996.

(3) Horvath, D.; Koch, C.; Schneider, G.; Marcou, G.; Varnek, A. Local neighborhood behavior in a combinatorial library context. *J. Comput.-Aided Mol. Des.* **2011**, *25* (3), 237−252.

(4) Papadatos, G.; Cooper, A. W. J.; Kadirkamanathan, V.; Macdonald, S. J. F.; McLay, I. M.; Pickett, S. D.; Pritchard, J. M.; Willett, P.; Gillet, V. J. Analysis of Neighborhood Behavior in Lead Optimization and Array Design. *J. Chem. Inf. Model* **2009**, *49* (2), 195−208.

(5) Horvath, D.; Barbosa, F. Neighborhood Behavior − the Relation Between Chemical Similarity and Property Similarity. *Curr. Trends Med. Chem.* **2004**, *4*, 589−600.

(6) Horvath, D.; Jeandenans, C. Neighborhood Behavior of In Silico Structural Spaces with respect to In Vitro Activity Spaces − A Benchmark for Neighborhood Behavior Assessment of Different In Silico Similarity Metrics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691−698.

(7) Horvath, D.; Jeandenans, C. Neighborhood Behavior of In Silico Structural Spaces with respect to In Vitro Activity Spaces − A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680−690.

(8) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39* (16), 3049−3059.

(9) Medina-Franco, J. L.; Martınez-Mayorga, K.; Bender, A.; Marın, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: Consensus Activity Cliffs. *J. Chem. Inf. Model* **2009**, *49* (2), 471−491.

(10) Guha, R.; VanDrie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model* **2008**, *48* (3), 646−658.

(11) Maggiora, G. M. On outliers and activity cliffs - Why QSAR often disappoints. *J. Chem. Inf. Model* **2006**, *46*, 1535−1535.

(12) Nair, P. C.; Sobhia, M. E. Fingerprint directed scaffold hopping for identification of CCR2 antagonists. *J. Chem. Inf. Model* **2008**, *48* (9), 1891−1902.

(13) Mauser, H.; Guba, W. Recent developments in de novo design and scaffold hopping. *Curr. Opin. Drug Discovery Dev.* **2008**, *11* (3), 365−374.

(14) Bergmann, R.; Linusson, A.; Zamora, I. SHOP: Scaffold HOPping by GRID-based similarity searches. *J. Med. Chem.* **2007**, *50*, 2708−2717.

(15) Sheridan, R. P. Chemical similarity searches: when is complexity justified? *Expert Opin. Drug Discovery* **2007**, *2* (4), 423−430.

(16) Schneider, G.; Schneider, P.; Renner, S. Scaffold-hopping: how far can you jump? *QSAR Comb. Sci.* **2006**, *25*, 1162−1171.

(17) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45* (19), 4350−4358.

(18) Bonachera, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. *J. Chem. Inf. Model.* **2006**, *46* (6), 2457−2477.

(19) de Luca, A.; Horvath, D.; Marcou, G.; Solovev, V.; Varnek, A. Mining Chemical Reactions Using Neighborhood Behavior and Condensed Graphs of Reactions Approaches. *J. Chem. Inf. Model.* **2012**, *52* (9), 2325−2338.

(20) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. Isida Property-labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29* (12), 855−868.

(21) Bonachera, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1 - Topological Fuzzy Pharmacophore Triplets and adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **2006**, *46*, 2457−2477.

(22) Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon, J. C. Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 470−480.

(23) Cerep BioPrint Database. http://www.cerep.fr/cerep/users/pages/Collaborations/Bioprint.asp (accessed 2007).

(24) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40* (D1), D1100−D1107.

(25) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008**, *48* (5), 941−948.

(26) DayLight Fingerprints and Similarity. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed 2013).

(27) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.

(28) Hu, Y.; Lounkine, E.; Bajorath, J. Improving the Search Performance of Extended Connectivity Fingerprints through Activity-Oriented Feature Filtering and Application of a Bit-Density-Dependent Similarity Function. *ChemMedChem* **2009**, *4* (4), 540−548.

(29) Senger, S. Using Tversky Similarity Searches for Core Hopping: Finding the Needles in the Haystack. *J. Chem. Inf. Model.* **2009**, *49* (6), 1514−1524.

(30) Wang, Y.; Bajorath, J. Balancing the influence of molecular complexity on fingerprint similarity searching. *J. Chem. Inf. Model.* **2008**, *48* (1), 75−84.

(31) Wang, Y. A.; Eckert, H.; Bajorath, J. Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. *ChemMedChem* **2007**, *2* (7), 1037−1042.

(32) Chen, X.; Brown, F. K. Asymmetry of Chemical Similarity. *ChemMedChem* **2007**, *2* (2), 180−182.

(33) Wang, Y.; Bajorath, J. Development of a Compound Class-Directed Similarity Coefficient That Accounts for Molecular Complexity Effects in Fingerprint Searching. *J. Chem. Inf. Model* **2009**, *49*, 1369−1376.

(34) ThomsonReuters ISI Web of Knowledge. http://apps.isiknowledge.com/UA_GeneralSearch_input.do?product=UA&SID=Q1kHJjnfj9nkMPONkHI&search_mode=GeneralSearch (accessed Nov 2012).

(35) ChemAxon *Standardizer*. http://www.chemaxon.com/jchem/doc/user/standardizer.html (accessed Feb 2013).

(36) ChemAxon *Tautomer Plugin*. http://www.chemaxon.com/marvin-archive/4.1.3/marvin/chemaxon/marvin/help/calculator-plugins.html#tautomer (accessed Feb 2013).

(37) ChemAxon *pKa Calculator Plugin*. https://www.chemaxon.com/products/calculator-plugins/property-predictors/ (accessed Feb 2013).

(38) Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(39) Bonachera, F.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 2. Application of Topological Fuzzy Pharmacophore Triplets in Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model* **2008**, *48* (2), 409−425.

(40) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4* (3), 191−198.

(41) Varnek, A.; Fourches, D.; Solov'ev, V.; Klimchuk, O.; Ouadi, A.; Billard, I. Successful "in silico" design of new efficient uranyl binders. *Solvent Extr. Ion Exch.* **2007**, *25* (4), 433−462.

(42) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693−703.

(43) Bonachera, F.; Marcou, G.; Kireeva, N.; A., V.; Horvath, D. Using self-organizing maps to accelerate similarity search. *Bioorg. Med. Chem.* **2012**, *20*, 5396−5409.

(44) Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, *84* (4), 327−352.

(45) Welch, B. L. The generalization of "Student's" problem when several different population variances are involved. *Biometrika* **1947**, *34*, 28−35.

(46) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42* (17), 3251−3264.

(47) Tanrikulu, Y.; Nietert, M.; Scheffer, U.; Proschak, E.; Grabowski, K.; Schneider, P.; Weidlich, M.; Karas, M.; Goebel, M.; Schneider, G. Scaffold hopping by "fuzzy" pharmacophores and its application to RNA targets. *ChemBioChem* **2007**, *8*, 1932−1936.

(48) Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8* (3), 255−263.

(49) Proudfoot, J. R. Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. *Bioorg. Med. Chem. Lett.* **2002**, *12* (12), 1647−1650.

(50) Sisay, M. T.; Peltason, L.; Bajorath, J. Structural Interpretation of Activity Cliffs Revealed by Systematic Analysis of Structure-Activity Relationships in Analog Series. *J. Chem. Inf. Model.* **2009**, *49* (10), 2179−2189.