Count your bits: more subtle similarity measures using larger radius count vectors

Florian Huber[1], Julian Pollmann[2]
Centre for Digitalisation and Digitality, Düsseldorf University of Applied Sciences, Düsseldorf, Germany
[1]ORCID: https://orcid.org/0000-0002-3535-9406.
[2]ORCID: https://orcid.org/0009-0004-3784-0019.

Abstract

Quantifying molecular similarity is a cornerstone of cheminformatics, underpinning applications from virtual screening to chemical space visualization. A wide range of molecular fingerprints and similarity metrics, most notably Tanimoto scores, are employed, but their effectiveness is highly context-dependent. In this study, we systematically evaluate several 2D fingerprint types, including circular, path-based, and distance-encoded variants, using both binary and count representations. We highlight the consequences of fingerprint choice, vector folding, and similarity metric selection, revealing critical issues such as fingerprint duplication, mass-dependent score biases, and high bit collision rates. Sparse and count-based fingerprints consistently outperform fixed-size binary vectors in preserving structural distinctions. Furthermore, we introduce percentile-based normalization, propose inverse-document-frequency (IDF) weighting, and benchmark all methods against graph-based MCES similarities. Our results offer practical guidance for selecting molecular similarity measures, emphasizing the need for conscious, task-aware fingerprinting choices in large-scale chemical analyses.

Introduction

Quantitative measures of molecular similarity (or distance) are essential for a wide range of applications in cheminformatics. They are used to identify structurally related molecules in large databases, to rank candidate compounds in virtual screening workflows [1,2] and to perform high-level data analysis tasks such as dimensionality reduction and chemical space visualization [3–6]. Molecular similarity scores are further increasingly used as a target for training and evaluating deep-learning models for the prediction of molecular structures from mass spectrometry or NMR data [7–9].
At the heart of molecular similarity is the question: *What does it mean for two molecules to be similar?* Unfortunately, there is no single, universally valid definition of molecular similarity, but rather a multitude of molecular similarity measures, each tailored to specific data analysis goals

or property predictions[10,11]. Two molecules may be considered similar if they share synthetic pathways, exhibit closely related biological or physicochemical properties, or differ only slightly in their overall molecular structure. Determining similarity thus heavily depends on the context and intended application.
For example, researchers working on lipidomics may focus more on broad structural classes while paying less attention to minor variations in fatty acid chains. In contrast, others may be interested in the exact length and saturation levels of side chains, adopting a viewpoint in which even small structural alterations can be significant.

Given these complexities, it is not surprising that a variety of computational methods, fingerprint algorithms, and similarity metrics have been proposed [12–16]. The prevailing approach generally involves first computing a molecular fingerprint as an abstract representation of a compound before applying a similarity or distance metric to these fingerprints. Among the numerous available metrics, the Tanimoto coefficient (closely related to the Jaccard index for binary data), Dice, and Cosine similarities are commonly employed [2,17]. Tanimoto scores have become a de facto standard, frequently cited as an effective choice for fingerprint-based similarity comparisons [1,2,13,17].

Commonly applied fingerprint algorithms, such as those based on circular or path-based methods, come with several favorable aspects. They can typically be computed very efficiently. Once generated, comparing large numbers of compounds by Tanimoto similarity is also highly scalable. Moreover, unlike data-driven embeddings derived from neural networks or other machine learning models, these classical fingerprints are model-independent. The bit patterns they produce retain a level of interpretability since certain bits can often be traced back to defined structural fragments in the molecule.

Nonetheless, Tanimoto-based similarities also have well-known limitations [6,18]. Even subtle structural changes can sometimes lead to unexpectedly large shifts in similarity values. Conversely, changes can be scored as irrelevant even though a chemist would classify them as relevant. Furthermore, most fingerprint algorithms may collapse distinct molecules into identical or near-identical bit patterns, resulting in no effective discrimination.
These shortcomings are not inherent to the Tanimoto metric alone, but rather they stem from the interaction between the chosen metric and the underlying fingerprint design. The common practice of broadly referring to all such measures as “Tanimoto similarity” can mask the important variability introduced by different fingerprint representations and parameter choices [12]. Moreover, the choice of the fingerprint algorithm as well as the similarity metric is often perceived as arbitrary [19].

Different molecular fingerprint classes

In this work, the focus is placed on 2D fingerprints, where structural features are encoded as sets of bits or count vectors. There is a rich space of such fingerprint types [19,20]. Two main classes predominate: circular fingerprints, which encode local neighborhoods of atoms within a

specified radius [2,21], and path-based fingerprints, which capture fragments defined by linear paths up to a certain length [2].

Both approaches have their strengths but also inherent restrictions. For example, binary fingerprints with a fixed path length or radius lose quantitative information about how many times a certain feature occurs. Count vectors can partially restore this information but still ignore broader structural arrangements beyond local neighborhoods.
More recent methods, such as the MAP4 fingerprint [14], attempt to incorporate a larger fraction of the molecular context, yet no single fingerprint method currently dominates in all scenarios. The search for a universally superior fingerprint remains ongoing (and will, sadly, also not be solved in this work).

An alternative to fragment-based representations is graph-based similarity measures, such as those derived from the Maximum Common Subgraph (MCS) or Maximum Common Edge Subgraph (MCES), which can offer a more complete structural comparison [22]. Such measures are conceptually very appealing and capable of avoiding several pitfalls of common molecular fingerprints, which is why we will use them in this work as a point of reference. Yet these measures might also fail to reflect the nuanced notion of similarity that a chemist might apply. By focusing primarily on the largest common subgraph, these methods emphasize shared structure over subtle but potentially crucial differences. For a pair of molecules sharing an extensive scaffold, for instance, replacing a methyl group with a strongly electron-withdrawing nitro substituent can dramatically change properties such as reactivity or binding affinity, yet the MCS-based similarity score would remain high.
In practice, however, the much more relevant limitation is the very high computational cost of such graph-based metrics. Although it is possible to improve overall computation times using threshold or approximated approaches [6,22], these remain many orders of magnitude slower than fingerprint-based similarity computations. The need to solve an NP-hard task results in poor scaling of such algorithms for larger and more complex molecules [23], thereby rendering them unsuitable for large-scale similarity calculations for comparing larger molecules.

Recently, a different class of data-driven approaches has emerged, leveraging deep learning and other machine learning techniques to learn abstract representations of molecules, e.g., embeddings [24–26]. Such embeddings can capture both global and local structural features in a data-driven manner. These approaches typically require large, well-annotated datasets and may lack the interpretability found in classical fingerprints. But they offer a novel and promising route to compare, analyse and visualize large chemical datasets [27]. In this work, however, we will solely focus on model-independent similarity computations.

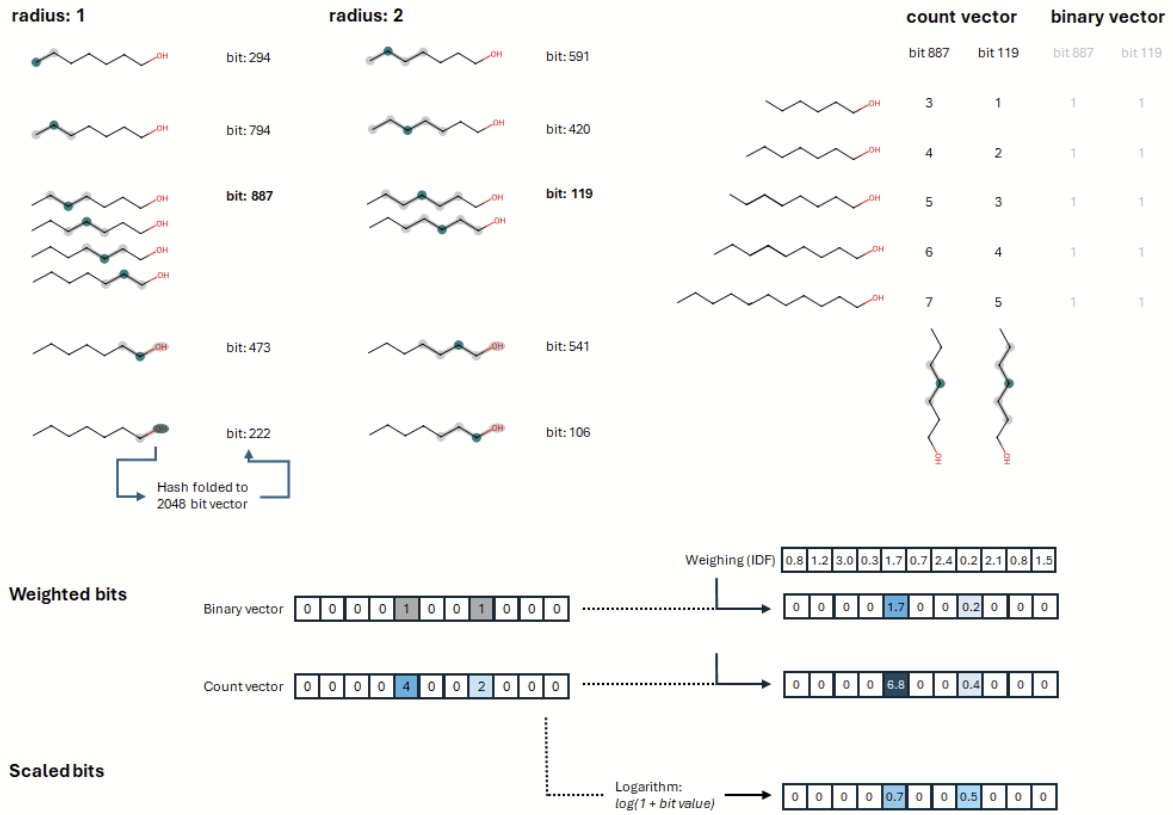

Figure 1 *Illustration of how circular fingerprints (e.g. Morgan-2 or ECFP-4) work. Per atom in the molecule, substructures constituting all atoms and bonds within a certain radius are collected. The hashes of those substructures are collected (sparse binary fingerprint), as well as their respective counts (sparse count fingerprint). All substructures can be mapped onto a fixed number of bins to create either binary or count fingerprint vectors. Here, two example substructures that would occur for Morgan-2 fingerprints are shown, one with radius=1 and one with radius=2. Fingerprint bit values can further be transformed by weighing (for instance, by multiplication with the inverse document frequency (IDF) of each bit) or by scaling (for instance, using a logarithmic scaling).*

Evaluation of molecular similarity measures

Evaluating different molecular similarity measures in order to guide the selection of appropriate methods and parameters is non-trivial [20]. Traditionally, this process often relies on tasks that approximate specific real-world scenarios, such as virtual screening benchmarks or enrichment factor analyses, that assess how effectively a similarity measure retrieves known actives from large compound sets [1]. This approach is valuable and has been applied and adapted in numerous studies [1,14,15,24].

However, such benchmarking represents only one particular task, which, in our opinion, does not reflect the full scale of current applications of molecular similarity. It is commonly applied for the task of detecting active compounds in larger datasets also containing decoys. While this might represent certain virtual screening tasks well enough, it does not necessarily reflect a universal requirement for good chemical similarity measures. Additionally, focusing on the top-ranked candidates will emphasize precision in a narrow subset of all possible molecule pairs. This leaves broader aspects of similarity assessment underexplored, as methods that excel at identifying a few key hits may not perform equally well in more general contexts.

Besides screening tasks, chemical similarity metrics are frequently used to create chemical space visualizations [4,20,28] with the specific aim of revealing how well a measure preserves meaningful structural relationships across the entire space. However, typical dimensionality techniques such as t-SNE [29] or UMAP [30] also rely on a rather small subset of best matches. Moreover, quantitative comparison of dimensionality-reduction-based visualizations is inherently difficult.

Another way of assessing the fingerprint quality is by downstream classification tasks to predict molecular properties based on molecular fingerprints. However, this often delivers somewhat inconclusive results [20,31] making it difficult to decide in favor of or against a particular type of fingerprint.

To address the mentioned limitations, we adapt, propose, and test additional means of evaluating similarity measures that go beyond screening tasks. We analyse undesired fingerprint duplications in order to highlight weaknesses of individual fingerprint types.
By considering such alternative evaluation strategies, we aim to shed light on different facets of molecular similarity and encourage a more holistic benchmarking approach.

Methods

Dataset:

Organic molecules inherently display a huge diversity with large variations in the estimates of how many molecules it contains [3]. In this study, we focus on two complementary datasets drawn from natural products and metabolomics research.
First, we assembled a curated collection of compounds measured and annotated by tandem mass spectrometry. We merged the training and evaluation data from MS2Deepscore, a deep-learning model for predicting chemical similarity from mass spectra[32] with the overlapping benchmarking set, MassSpecGym [8]. The resulting **ms2structures dataset** comprises 37,811 unique compounds.
Second, we created a larger **biostructures dataset** by starting from 718,097 biologically relevant compounds drawn from the dataset used by Kretschmer et al. [6]. After removing 30 entries that RDKit could not fingerprint, 718,067 compounds remained. We then merged these with the ms2structures set, yielding 730,464 distinct structures.

Both datasets are publicly available at https://doi.org/10.5281/zenodo.15650044.

Molecular Fingerprints

Most molecular fingerprints are computed using RDKit [33]. Specifically, we generated both binary and count-based Morgan fingerprints, as well as binary and count-based RDKit (Daylight-like) fingerprints. In addition, we calculated MAP4 fingerprints via the Python package available at https://pypi.org/project/map4/).

All fingerprinting routines (dense and sparse) are implemented in optimized Python code, which can be found on GitHub: https://github.com/florian-huber/molecular_fingerprint_comparisons

Fingerprint Similarity Metrics

We applied Tanimoto scores to compute the similarity between molecular fingerprints. For binary fingerprints X and Y this meant using the following formula which is commonly referred to as Tanimoto or Jaccard index.

$$S_T(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

For count vectors X and Y containing positive integers as well as for scaled vectors containing float numbers, we use the closest relative which is often termed generalized Tanimoto coefficient or counted Tanimoto similarity (other terms are generalized Jaccard index or Ruzicka score):

$$S_{gT}(X,Y) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$$

When used with binary fingerprints, the generalized Tanimoto coefficient leads to the exact same results as the original Tanimoto coefficient. In the following, we will therefore simply refer to “Tanimoto” without further distinction. For optimal runtime, however, we used four different Tanimoto implementations for dense binary fingerprints, dense count fingerprints, sparse binary fingerprints and sparse count fingerprints. The performance optimized Python code is available on Github: https://github.com/florian-huber/molecular_fingerprint_comparisons. In future iterations, this code will either be merged or be integrated into the matchms library [34] or its ecosystem.

Percentile scores

To facilitate comparison between different fingerprint types and datasets, we convert raw similarity values into percentile scores on a 0–100 scale. Given a symmetric similarity matrix $S \in \mathbb{R}^{N \times N}$, we extract the strictly upper-triangular entries, rank them (using average ranks for ties), and then linearly scale the ranks so that the smallest similarity corresponds to 0 and the largest to 100. This transformation preserves the relative ordering of similarity scores while placing them on a uniform, interpretable scale. This makes cross-dataset and cross-fingerprint comparisons more meaningful.

RascalMCES scores

We randomly sampled 5,557,963 compound pairs from the ms2structures dataset whose precursor masses differ by at most 100 Da. RascalMCES scores [22] were computed with RDKit [33] on an Intel Core i9-13900K (settings: similarityThreshold=0.05, maxBondMatchPairs=1000, minFragSize=3, timeout=60 s). After excluding 144,286 timed-out pairs, our final benchmarking

set contained 5,413,677 pairs. These scores are available at https://doi.org/10.5281/zenodo.15650144.

To compare how well fingerprint-based similarities align with RascalMCES, we used Kendall's Tau-B, as implemented in SciPy[35]:

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - T_1)(n_0 - T_2)}}$$

where n_c is the number of pairs for which both methods rank one compound higher than the other (i.e., they agree on which is more similar), and n_d is the number of pairs for which the methods give opposite rankings. n_0 is the total number of pairs and T_1 and T_2 are the counts of tied scores under each method.
Unlike Pearson's correlation, Kendall's Tau-B emphasizes how consistently the two methods order compound pairs, without being influenced by the actual score values or their distributions.

Bit-Weighing

Fingerprints such as Morgan/ECFP or RDKit/daylight are rooted in the detection of unique subsets of a molecular graph. Each unique subset is thereby treated equally, which has the advantage that the algorithms do not need to incorporate extensive heuristics to account for chemical knowledge. However, this will also not distinguish very common, or trivial, chemical subsets from more particular subsets. This problem resembles a common challenge in Natural Language Processing, where simply counting the occurrences of every word in a sentence is often insufficient. This is because many words, such as so-called stopwords, carry little specific meaning, while others are key for determining the topic or mood of the sentence.
We here adapt the concept of term frequency (TF) as well as the inverse document frequency (IDF) to account for the overall relevance of individual fingerprint bits in a fingerprint. The term frequency is here identical to the values in a count fingerprint, which represent the number of occurrences of a specific substructure in a molecule. The IDF is computed in a normalized way as follows:

$$IDF_i = \log_{10}\left(\frac{N}{n_i}\right) / \log_{10}\left(\frac{N}{1}\right)$$

Where N is the total number of compounds in the dataset and n_i is the number of compounds that contain the substructure corresponding to bit i. For this, sparse vectors are used to avoid bit collisions. The IDF values were computed for both the ms2structures dataset as well as the biostructures dataset using Morgan-3 fingerprints.

Results

To explore the practical impact of different fingerprints, we first compared their bit occupations across a dataset of 730,464 biomolecular structures (**biostructures dataset**). Figure 2 provides an overview of the average bit occupation for various fingerprint types. As expected, clear differences emerged. The RDKit path-based fingerprint, computed with a default sequence length of 7, tends to yield a high bit occupation rate, particularly in the smaller, 1024-bit vectors. Here, about 90 \% of all fingerprint bits are occupied for more than 40% of all compounds in the

biomolecular dataset. By contrast, Morgan fingerprints exhibit significantly lower bit occupancy. Even for small 1024-bit fingerprints, only about 2% of all bits are occupied for more than 40% of all compounds.
Switching to larger 4096-bit fingerprints strongly reduced the overall occupation rates for all fingerprint types, yet the large discrepancy between RDKit and Morgan fingerprints remains. This reflects the more granular, local-neighborhood encoding strategy of Morgan-type features.

MAP4, which encodes the distance between substructural motifs and thus captures a wide variety of fragment types, shows a bit occupation profile roughly comparable to the RDKit fingerprint.

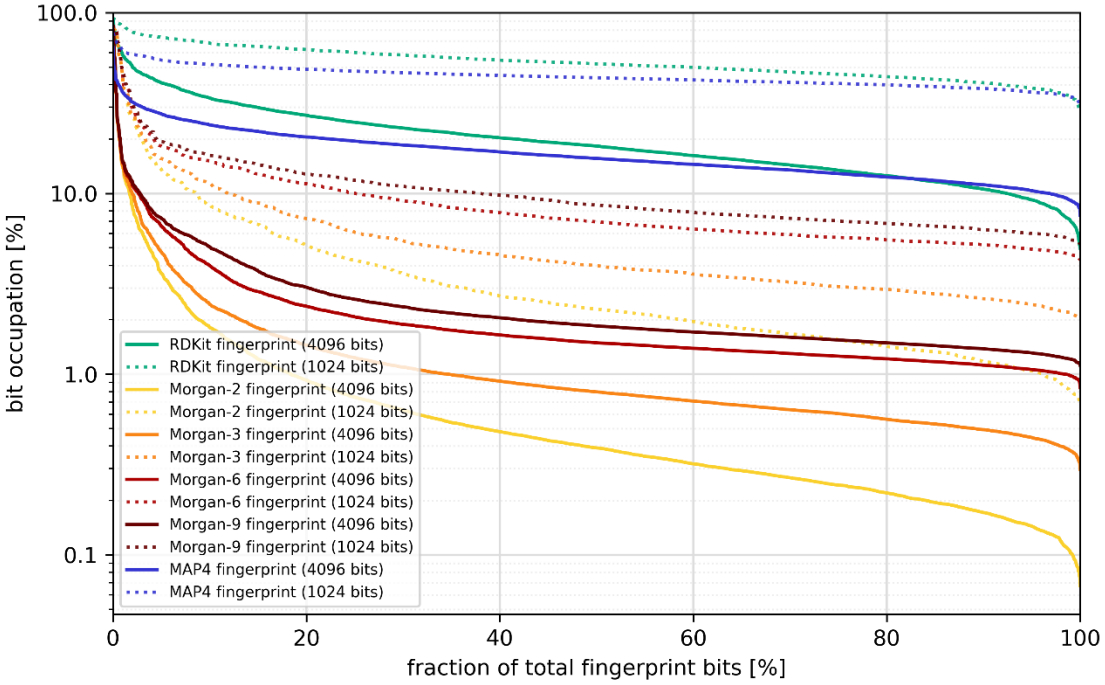

Figure 2 Bit occupation frequency for different molecular fingerprints displays a large variation for different fingerprints types and vector lengths. This ranges from nearly all bits being occupied > 30% of the time for 1024-bit RDKit and 1024-bit MAP4 fingerprints (green and blue dashed lines) to fingerprints with only a tiny fraction of bins occupied this often (4096-bit Morgan fingerprints).

For the Morgan and RDKit fingerprints, we generated both binary and count-vector variants, although only binary vectors were evaluated for bit occupation. In the count-vector representations, each entry stores an integer that reflects the count of specific substructures rather than a simple presence/absence bit. The overall occupancy (the fraction of indices containing non-zero information) thus remains identical between binary and count-vector versions, but the ability to capture frequency may yield different similarity scores downstream.

Fingerprint duplicates

Because molecular similarity is inherently context-dependent, it is challenging to compare scoring methods in a strictly quantitative manner. Nonetheless, certain criteria can serve as clear-cut stress tests for any similarity measure. In line with the term “fingerprint”, one such property is the ability to avoid identical fingerprints for pairs of structurally distinct molecules.

We therefore identified all instances in which different molecules in the dataset shared an identical fingerprint. For each duplicated fingerprint, we computed the maximum mass difference between any two compounds bearing that same fingerprint. Figure 3 summarizes these statistics. MAP4 exhibited a particularly low number of duplicates, only 238 in total, which indicates its ability to discriminate between highly similar compounds. By contrast, the fingerprint variants derived from the Morgan family, and especially the RDKit binary fingerprint, produced far more duplicates.

The RDKit binary fingerprint with 1024 bits yielded 147,388 duplicated entries, accounting for about 20% of all compounds in the dataset. Switching to 4096-bit fingerprints only slightly reduces the number of duplicates. Moreover, these duplicates included the highest fraction of large mass differences, with many pairs differing by over 400 Da despite sharing identical fingerprints.

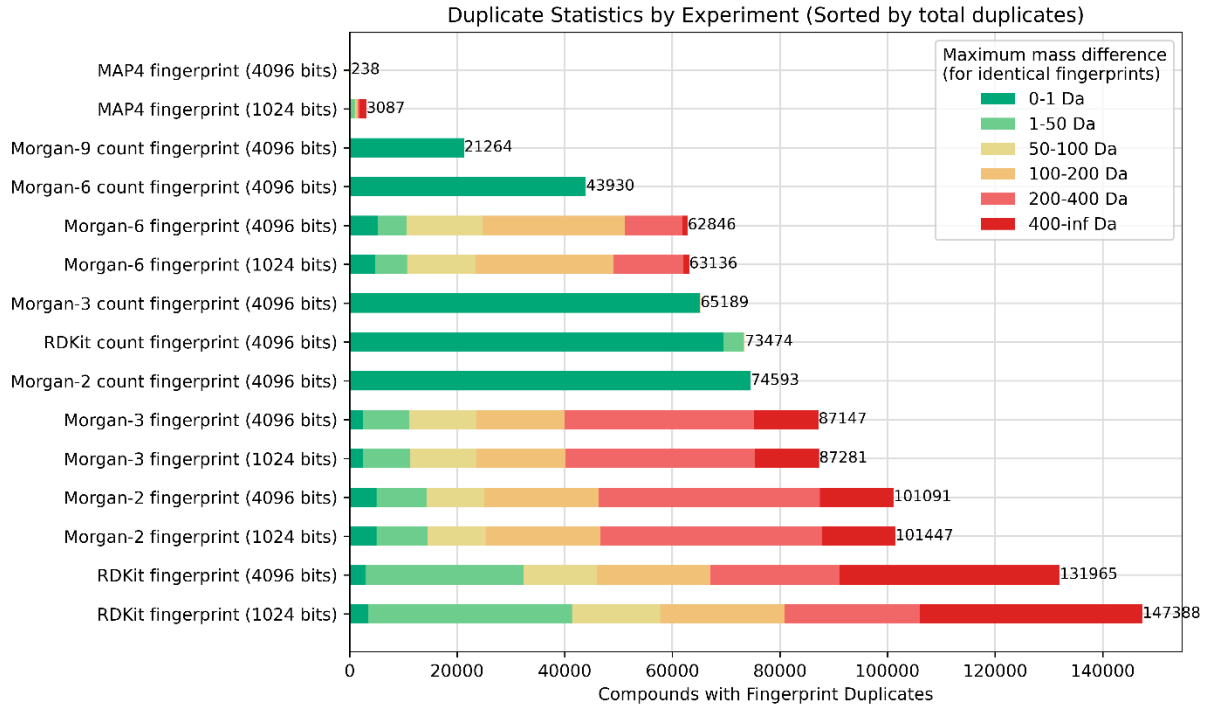

Figure 3 For different molecular fingerprints, all duplicates in a set of 730,464 unique compounds (biomolecular structures dataset) were counted. For each duplicate fingerprint, we computed the maximum mass difference among compounds sharing that fingerprint which is colored along different bins ranging from 0-1Da mass difference up to mass differences >= 400 Da.

For Morgan fingerprints the number of duplicates notably decreases with increasing radius. In addition, count fingerprints led to fewer duplicates than their binary counterparts. Even more noteworthy, however, is that for count fingerprints virtually all duplicate pairs (>99.96%) correspond to negligible mass differences of less than 1 Da.

Fingerprint Comparisons Reveal Divergent Perspectives on Molecular Similarity

Different fingerprint representations capture distinct aspects of molecular structure, and these differences are reflected in their similarity score distributions. In our analysis, we compared two widely used fingerprint types, topological fingerprints represented here by RDKit fingerprints, and circular fingerprints represented by Morgan fingerprints, across the 37,811 compounds in the **ms2structures dataset**. Although both fingerprint types are commonly employed in tasks such as compound ranking, their underlying algorithms and bit occupancy rates differ noticeably, leading to different interpretations of what it means for two molecules to be similar.

Due to its higher bit occupation rates, the default RDKit fingerprints generally produce higher Tanimoto similarity scores compared to Morgan-3 fingerprints. To enable a more meaningful comparison, the raw similarity scores were converted into percentile values ranging from 0% to 100% based on all unique compound pairs. This scaling allowed us to directly contrast how each fingerprint representation ranks molecular similarity across the entire dataset.

Comparison 1: Morgan-3 Binary vs. RDKit Binary Fingerprints
When comparing Tanimoto similarity scores computed from Morgan-3 binary (4096-bit) fingerprints with those derived from RDKit binary fingerprints of the same bit length, substantial differences in their percentile distributions were observed. Notably, a significant proportion of compound pairs, approximately 43.41%, were consistently assigned to the bottom 60% of similarity by both scoring methods.
In contrast, only a smaller fraction of pairs (0.42%) was jointly classified within the top 1$\%$ of similarity. Moreover, there were striking discrepancies for certain compound pairs. For instance, over 200,000 pairs (0.04% of all unique pairs) were placed in the top 1% by RDKit fingerprints while being relegated to the bottom 60% by Morgan-3 fingerprints.
Conversely, more than 900,000 pairs (0.13%) were ranked in the top 1% by Morgan-3 fingerprints but fell into the bottom 60% based on RDKit scores (Figure 4). These results underline that even when using the same Tanimoto metric, the two fingerprint types emphasize different molecular features, resulting in divergent similarity assessments.

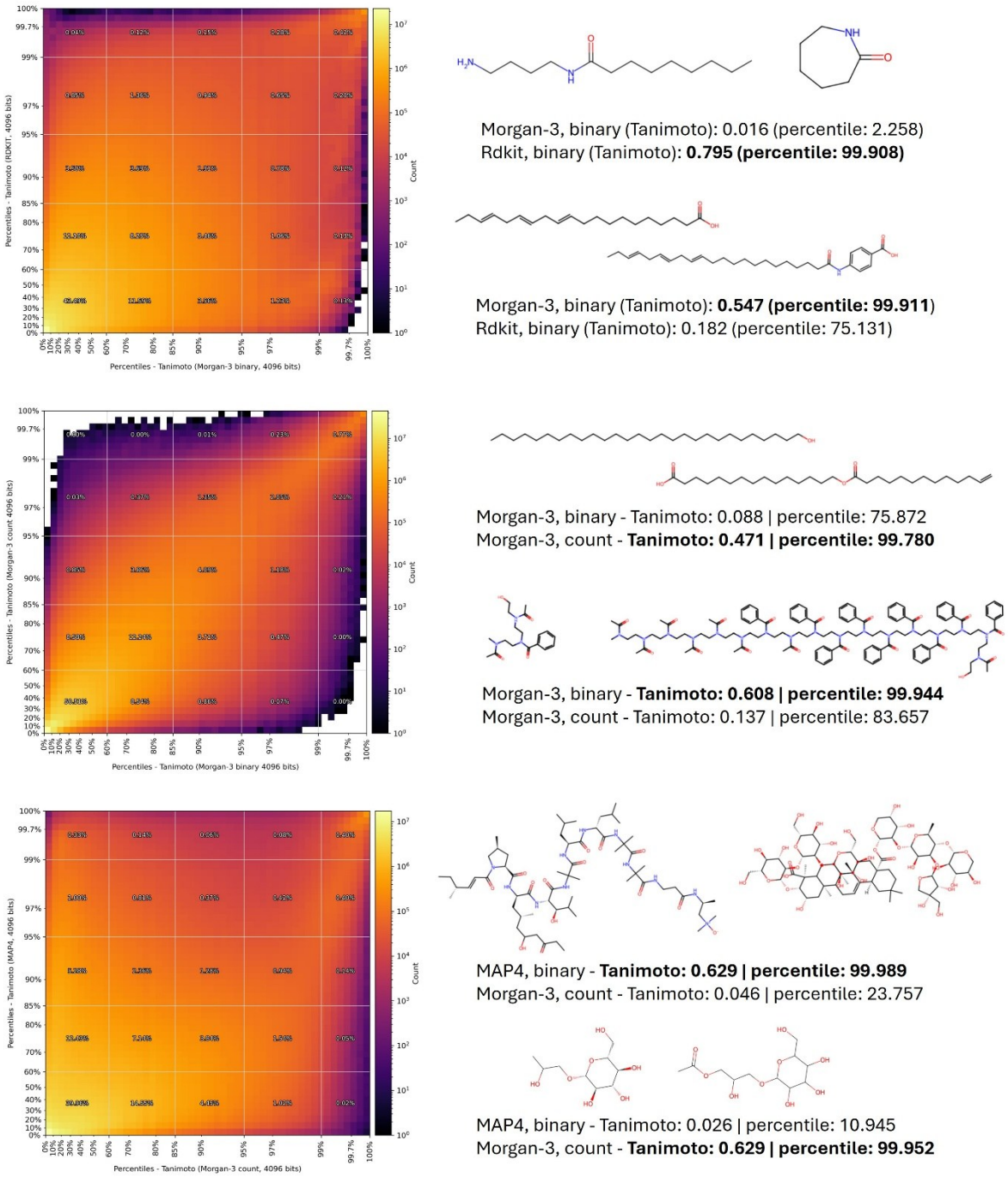

Figure 4 Comparison of all pairwise similarities for the ms2structures dataset (37,811 compounds) using various fingerprints. ***(top row)*** *Tanimoto scores computed from RDKit fingerprints and Morgan-3 binary fingerprints.* ***(center row)*** *Tanimoto scores computed from Morgan-3 binary fingerprints and from Morgan-3 count-based fingerprints.* ***(bottom row)*** *Tanimoto scores were computed from Morgan-3 count fingerprints and from MAP4 fingerprints. All fingerprints were 4096-bit vectors. Example pairs with drastic score discrepancies are highlighted on the right.*

Comparison 2: Morgan-3 Binary vs. Morgan-3 Count

To further probe the influence of fingerprint representation, we compared Tanimoto scores computed from Morgan-3 binary fingerprints with Tanimoto similarity scores derived from

Morgan-3 count-based fingerprints. In this pairing, both methods are based on the same underlying fingerprint type, yet the representation (binary versus count) introduces noteworthy differences.
The analysis revealed that about 50.42% of compound pairs were consistently ranked in the bottom 60% by both metrics, while 0.77% of pairs were jointly assigned to the top 1%. Although the overall agreement between the two metrics is higher than in the previous comparison, notable differences persist for certain compound pairs.
The count-based fingerprints offer the advantage of recognizing repeated occurrences of motifs or substructures, as exemplified in an example pair in Figure 4, providing a more nuanced view of molecular similarity. However, this sensitivity can also be a drawback, as it may overemphasize the contribution of common substructures, such as extended carbon chains (e.g., CC chains) leading to disproportionately high similarity scores in some cases.

Comparison 3: MAP4 vs. Morgan-3 Count
The third comparison involves Tanimoto similarity scores computed using MAP4 fingerprints and from Morgan-3 count-based fingerprints, with both representations using 4096-bit vectors. As illustrated in Figure 4 (bottom row), the overall correspondence between these two metrics is less pronounced than in the previous comparisons. Only 39.97% of compound pairs fall into the bottom 60% for both metrics, and just 0.39% are classified in the top 1% by both methods. Notably, MAP4 exhibits the highest bit occupation among the fingerprints evaluated, a characteristic that appears to lead to a substantial number of bit collisions, particularly for larger molecules. This increased collision rate manifests in a sizable subset of cases: approximately 0.33% of pairs are ranked in the top 1% for MAP4 while simultaneously being assigned to the bottom 60% according to the Tanimoto score on Morgan-3 count fingerprints.
In addition, a smaller group of pairs is observed in the opposite corner of the score distribution, where the Morgan-based count fingerprints place them in the top 1% while MAP4 assigns them to the bottom 60%. We noted that those pairs are usually clear cases of unambiguous false similarity assignments by MAP4 as indicated by the two example pairs in Figure 4.

Sparse vectors to avoid bit collisions

Molecular fingerprints are usually brought to a fixed size. Common in literature, tutorials, and library default settings are vector sizes of 1024, 2048, or less frequently up to 4096 bits. Folding the found substructure bits into fixed-sized vectors is an efficient trick to simplify further use of such fingerprints, both conceptually as well as computationally. However, this approach results in information loss, and bit collisions are a well-known downside. When the number of bits is high enough, the effect of bit collisions on ranking tasks was described to be very low by some [1] while others reported that, in particular for very close analogues, results improved notably when using 16384-bit fingerprints instead of 1024-bit ones [36]. However, due to the focus on very close analogues, this might not fully reveal the extent to which bit collisions might influence the all-vs-all comparisons of a larger number of molecules.
A large part of the very pronounced differences between the various similarity scores as shown in Figure 4 can be attributed to fundamentally different fingerprint concepts. Yet, at the very least a notable fraction of very high discrepancies between MAP4 and Morgan-3 based

fingerprints pointed at more substantial fingerprint-related shortcomings (see pair examples in Figure 4). Consistent with the high overall bit occupation ratio shown in Figure 2, this suggests severe effects from bit collisions. Consequently, we also computed sparse fingerprints of all tested variants, which kept all original, unfolded bits.

We computed Tanimoto scores for all unique pairs between the 37,811 compounds in the ms2structures dataset using both sparse as well as 4096-bit Morgan-3 count fingerprints. The results confirm that the scores are generally only very mildly affected (Figure 5, left). Interestingly, in the case of Morgan-3 count fingerprints, we found that computing molecular similarities for sparse fingerprints was about five times faster than for Morgan-3 dense 4096-bit fingerprints.

In both cases, functions were used that were just-in-time compiled and parallelized using Numba[37]. This can be explained by a relatively low number of bits per molecule for Morgan-3 (see Figure 6). For fingerprints with a much higher number of occupied bits, such as the RDKit fingerprint with the default setting of paths up to the length 7, handling sparse fingerprints is still feasible but with notably slower similarity score computations.

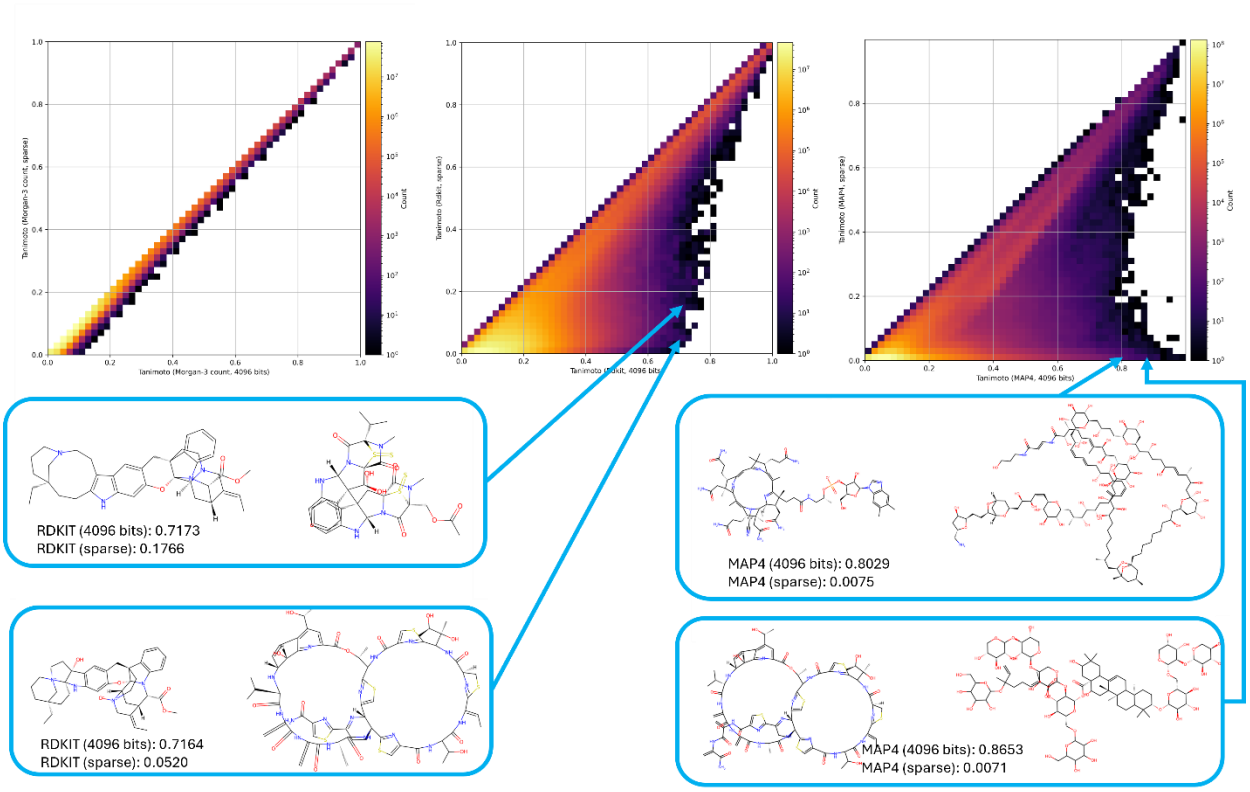

Figure 5 Comparison of sparse vs fixed-size 4096-bit RDKit fingerprints on all possible unique pairs between 37,811 compounds (ignoring pairs between identical compounds). Shown are Tanimoto scores for Morgan-3 count fingerprints (left), Tanimoto scores for RDKit fingerprints (center), and Tanimoto scores for MAP4 fingerprints (right). Due to bit collisions in the RDKit and MAP4 fingerprints, the scores based on 4096-bit vectors are generally notably higher. In some cases, pairs that should have a very low similarity according to the respective fingerprint algorithm do receive very high scores due to bit collisions (see example pairs in the blue boxes).

For the RDKit fingerprints, as well as for folded MAP4 fingerprints, sparse fingerprints lead to very pronounced shift in similarity scores, shown in Figure 5. Both for RDKit and MAP4 fingerprints, this reveals a very large fraction of pairs that have little to no overlap in actual bits (as represented in the sparse vectors), but which are shifted towards values in the 0.1 to 0.2 score range due to the folding of bits to a fixed vector size of 4096. Effectively, this raises the average similarity scores for such fingerprint types.

In addition, a smaller but notable number of pairs erroneously receive very high similarity scores of up to 0.6 or higher, even though they correctly belong in the 0 to 0.2 score range. The latter is particularly dangerous for ranking or search tasks because this can lead to arbitrary false hits. In the following steps, we aim to avoid any effects from bit collisions because those might obscure the actual logic and potential of the respective metrics. Therefore, sparse fingerprints will be used unless noted otherwise. In practice, those might come with higher computational costs, but at least for the fingerprints, metrics, and compounds tested in this work, such computation times were usually within the same order of magnitude.

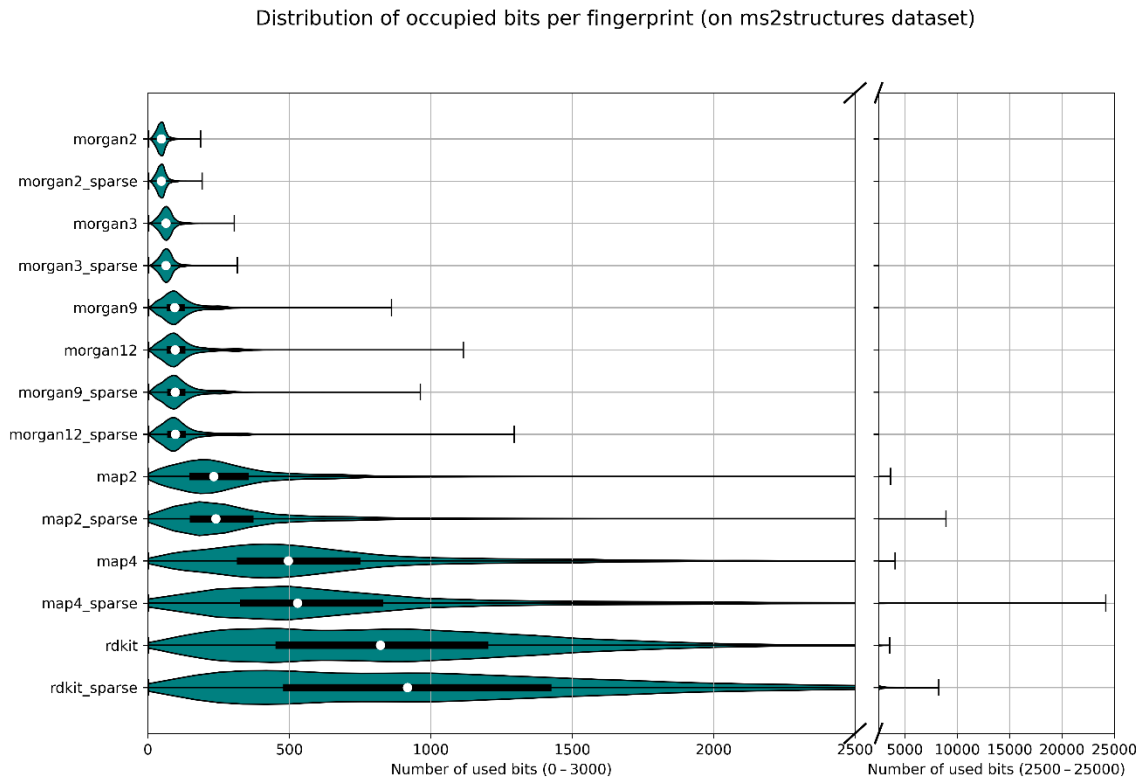

Figure 6 Distributions showing the number of non-zero bits over the entire ms2structures dataset. The subplot on the right shows the long tail of some of the distributions towards reflecting few compounds which led to very high numbers of occupied bits. The sparse MAP4 implementation, for instance, resulted in fingerprints of up to 24,183 bits.

Compound size dependence

The number of possible substructures of a compound generally increases with increasing compound size. Therefore, heavier organic compounds should, on average, have a higher number of occupied bits in a molecular fingerprint, as well as higher values for count vectors. It has been shown that this difference in bit occupation can result in different statistical properties of similarity scores [38].

Here we compare all similarity scores between the smaller molecules in the ms2-structures dataset (< 300 Da, 8,601 compounds, 36,984,300 pairs) with all scores between larger molecules (> 500 Da, 8,570 compounds 36,718,165 pairs). Both groups of compounds contain an approximately equal number of elements to avoid effects based on the sample size. Out of the scores for all possible pairs in each group, score percentiles starting at 50% were computed and are displayed in Figure 7.

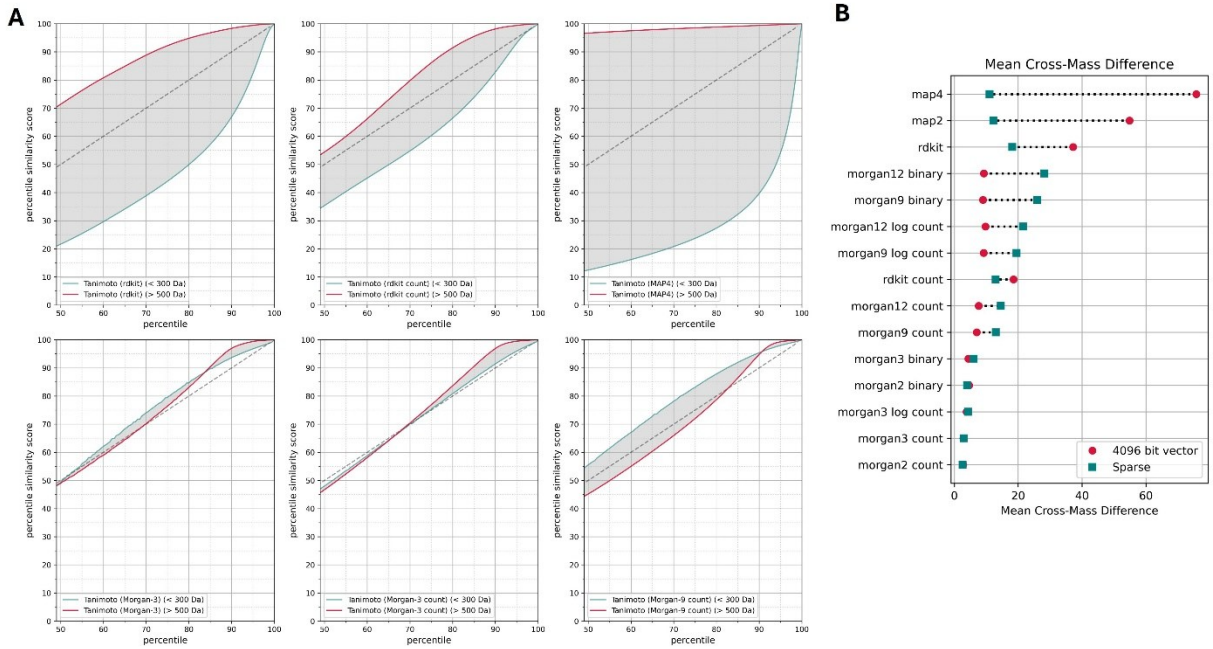

Figure 7 ***(A)*** *Fingerprint-based similarities are computed for about 36 million unique pairs between small compounds (< 300 Da) or between larger compounds (> 500 Da). The similarity scores of increasing percentiles are plotted for both the small and larger molecules, ranging from 50% up to 99.5%. This is done for Tanimoto scores of RDKit fingerprints (top left), RDKit count fingerprints (top center), Tanimoto scores of MAP4 fingerprints (top right), Tanimoto scores of Morgan-3 binary fingerprints (bottom left), Tanimoto scores of Morgan-3 count fingerprints (bottom center), as well as Tanimoto scores of Morgan-9 count vectors (bottom right). All fingerprints used had 4096 bits.* ***(B)*** *The mean difference between the high and the low mass curves was computed for multiple fingerprints, both using a sparse and a fixed-size (4096 bits) implementation.*

For the fingerprints with high bit occupation frequencies, RDKit and MAP4, the similarity score percentiles are consistently much higher for larger molecules. For the tested Morgan-3 fingerprints, this is only the case for higher percentiles starting at about 85 - 90% and the effect generally remains moderate. The median of all similarity scores, which is the percentile at 50%, is hence largely the same within small and larger molecules in the case of Morgan-3 fingerprints, both for binary fingerprints and count fingerprints.

In contrast, for RDKit and MAP4 fingerprints, the median score differs strongly depending on the molecular mass. For instance, for the RDKit binary fingerprint, about 15% of all pairs between larger molecules (> 500 Da) received a similarity score which was higher than 0.3, placing it in the top 2% of all scores. At the same time, only about 1% of pairs between smaller molecules (< 300 Da) received similarly high scores (see also Figure S1 in the supplemental material).
To better compare many different fingerprints, we computed the mean difference between the high and the low mass curves, which we will refer to as the mean cross-mass difference (Figure 7 B). This means cross-mass difference increases when we move from Morgan-2 and 3 to Morgan-9 and Morgan-12 fingerprints, especially when using binary fingerprints. Interestingly, we see higher mean cross-mass differences for similarity scores computed for sparse Morgan fingerprints when compared to fixed-size 4096-bit vectors. For RDKit and MAP4 fingerprints, we observe the opposite effect, which is particularly pronounced for MAP4 fingerprints.

Focus on the top hits: Ranking

The score comparisons over the entire dataset illustrate that the different scores and fingerprints can lead to sometimes drastically different interpretations of similarity. For certain downstream tasks the full range of scores is of high interest. For instance, for training machine-learning models to predict chemical similarities (e.g., MS2DeepScore [32,39]) or models trained to predict compounds structures which are then evaluated based on chemical similarity measures, typically Tanimoto scores of common molecular fingerprints [7–9,40].

For other tasks, such as nearest neighbor searches, however, one might argue that it is of little importance if a pair of compounds is in the lowest 20% percentile or in the lowest 50% percentile since the entire focus is set on finding the most similar compounds. The direct score comparisons in Figure 4 display that even in the highest percentiles the scores occasionally differ, but this is probably better assessed using top-n rankings. Unlike virtual screening tests such as Riniker et al [1], we here first want to inspect the ranking agreement between the different scores across the full ms2structures dataset.

For each of the 37,811 compounds, the closest 10 neighbors are computed for several different similarity scores (RDKit binary and count, Morgan-3 binary and count, Morgan-9 count, and MAP4). To account for potential issues due to bit collisions, both 4096-bit and sparse fingerprints were used. In the next step, we measure the overlap in the top-10 selections between all possible score combinations and compute the mean overlap, which is displayed in Figure 8. Not surprisingly, there is a high overlap between each score based on 4096-bit vectors versus sparse vectors.
Consistent with the very moderate difference between dense and sparse variants of Morgan-type fingerprints (e.g., Figure 5), we see a relatively high selection agreement of, on average, more than 9 out 10 candidates. This value is lower for RDKit fingerprints (8.54) and most pronounced for MAP4 fingerprints where, on average, only 7.72 of the top-10 hits based on the sparse MAP4 fingerprint are also found in the top-10 of the non-sparse implementation.

Furthermore, this ranking comparison reveals a fair agreement in the top-10 rankings between Morgan-3 count and binary fingerprints with overlaps ranging from 6.72 to 6.81. Morgan count vectors with radii 3 or 9 even show selection agreement of more than 8 out of 10. For RDKit fingerprints the agreement between count and binary fingerprint-based rankings is a bit less, with an overlap between 6.14 and 6.44 out of 10. Interestingly, the top-10 compound selections based on RDKit, Morgan, and MAP4 fingerprints all show only moderate agreement among each other, typically no more than 4 to 5 out of 10. The lowest agreement of less than 4 out of 10 is found between Morgan-9 count and RDKit binary fingerprints.

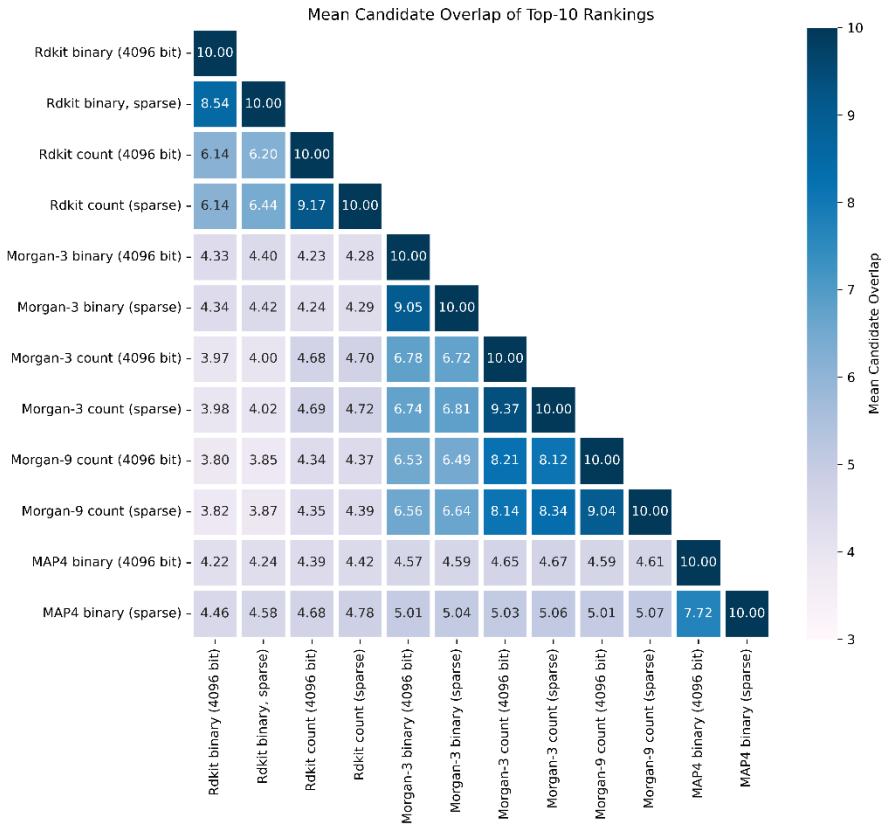

Figure 8 For any compound in the ms2structures dataset (37,811 compounds), the ten most similar compounds were selected based on the different similarity measures. For all possible combinations, the average overlap between all top-10 selections was computed.

Ranking compounds is key in many large-scale searches. Such tasks in the form of virtual screenings are probably the most common approach to benchmark molecular fingerprints and similarity measures with the open-source benchmarking platform of Riniker and Landrum [1]. This has recently been expanded by tasks involving peptides [14]. MAP4 fingerprints, and more recently chirality aware MAP4c fingerprints [15], were shown to perform particularly well on this benchmarking set. When peptides were not included though, the better average ranking of MAP4 is often not statistically significant, including when compared to Morgan-2 2024bit binary fingerprints (see supplemental material in [14]).

We contend that this ranking-based aggregation might introduce two potential issues, one with the aggregation of ranks and one with the underlying screening task. Aggregation of ranks is a

commonly applied technique, mostly because it focuses on relative performance per query, is robust to metric outliers, and avoids unjust bias towards methods optimized on high-variance subsets [41]. But converting continuous performance differences into ordinal ranks can obscure the true magnitude of those differences. Even minimal differences may result in a rank change of one, while substantial differences may be treated equivalently if they occupy consecutive rank positions. And, the averaged ranks are not independent quantities but depend on the selection of competing scores.
Secondly, the ranking tasks in the mentioned studies [1,14,15] are geared towards detecting compounds with assigned biological activity within a large pool of decoys, thereby simulating the detection of active compounds. This, however, is not directly related to the conceptual design of most fingerprints, which are rather broad descriptors of overall molecular substructures.
We ran the ranking script from Orsi and Reymond [15] with various sparse Morgan-based fingerprints. Overall ranking performance on the datasets provided for small molecules remained best for the MAP4 fingerprints (see Figure S2 in supplemental material).
Since those mean ranking plots depend on the selection of scores and hence do not allow for a flexible, general benchmarking over many different scores and future adaptations, we here also compare several other performance metrics as used in [15] (Figure S3 in supplemental material). Here, too, MAP4 frequently received the best average metrics, though not on all provided metrics.
Interestingly, we found that larger radius Morgan fingerprints, such as sparse log-count Morgan-9 fingerprints perform comparably well on the peptide task, being roughly on par with MAP4 for the mean rank (Figure S2) as well as the difference mean metrics (Figure S4).

Comparison to graph-based method

As described in the introduction, there is no generally valid ground truth chemical similarity measure. A conceptually compelling approach, however, is graph-based measures to assess the maximum common subgraph (MCS) or maximum common edge subgraph (MCES) of two compounds.

Those measures come in different flavors and implementations, but all share the disadvantage of a very high computational cost when compared to the above use of fingerprint-based methods. Fingerprint computations scale very favorably with molecular size, and once fingerprints are computed, the actual computation of molecular similarities is compound-size-independent, at least for fixed-sized vectors.
In contrast, MCS/MCES measures can sometimes require substantial computational resources for only a few such calculations. Even when making use of lower bound estimates, this disadvantage can only partly be mitigated [6,22,23]. In addition, such optimization by design makes the scores less suitable for computing chemical similarities between dissimilar compounds.

We here computed RascalMCES [22] scores implemented in rdkit [33] for a sampled subset of 5,413,677 compound pairs out of the ms2structures dataset (see Methods). To avoid too many trivial cases of entirely unrelated molecules, we only sampled from pairs with a mass difference of up to 100 Da. The rascalMCES scores were then compared to the different fingerprint-based

similarity measures by computing the Kendall correlation. This measures how often a higher fingerprint-based similarity corresponds to a higher MCES similarity.
When using fixed-sized fingerprints of 4096 bits, very low correlations can be observed for RDKit, MAP2 and MAP4 fingerprints (Figure 9). Switching to a sparse implementation can largely rescue this poor performance on this task, with sparse RDKit count fingerprints even receiving the highest overall Kendall correlation. Although not always as pronounced as in those three cases, the Kendall correlation with the RascalMCES scores increases for all scores when switching from fixed-sized vectors to a sparse implementation.

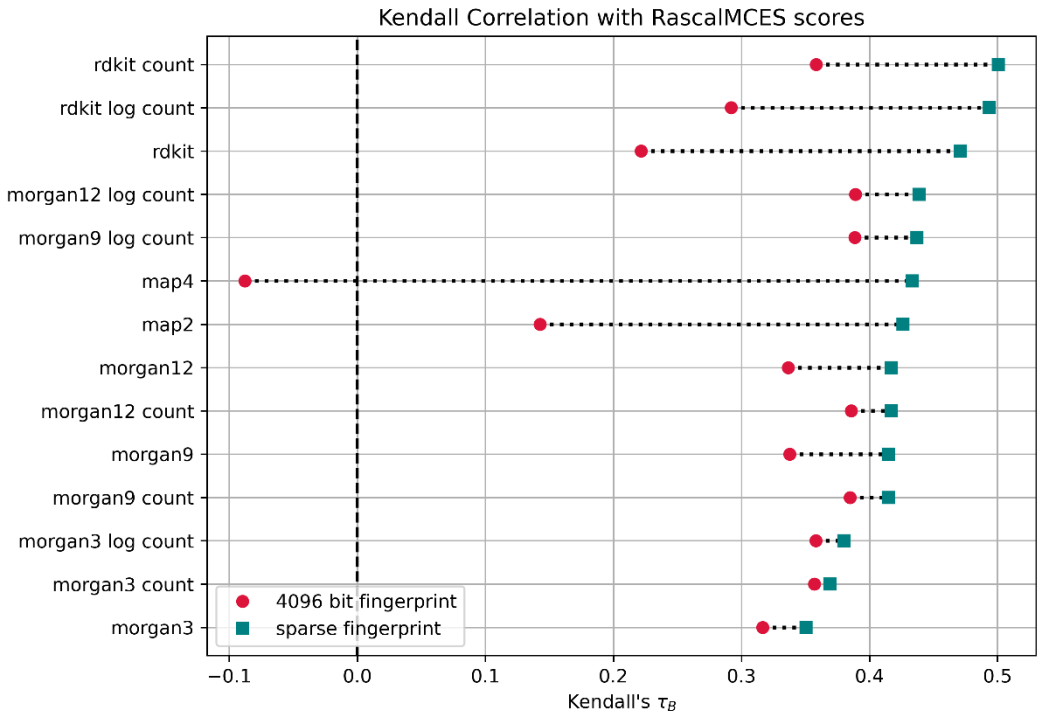

Figure 9 Kendall tau-b correlation between rascalMCES scores and various fingerprint-based similarity measures. For each fingerprint, both 4096-bit (red circles) and sparse (teal squares) implementations were used. The correlation coefficients were computed based on 5,413,677 randomly sampled compound pairs from the ms2structures dataset (see methods).

Bit frequency-based weighing

The bit occupation distributions in Figure 2 show that some fingerprint bits occur much more often than others. In the logic of both binary and count-based fingerprints, such as the RDKit and Morgan fingerprints, but also MAP4, all fingerprint bits are equally important. This means that for a similarity computation, the presence or absence of a very common substructure counts as much as the presence or absence of a rarely occurring substructure.

We weighted each fingerprint bit using the inverse document frequency (IDF, see Methods). This was done using sparse RDKit and Morgan fingerprints to avoid bit collisions. As expected, the most frequent substructures for Morgan fingerprints were of radius 0 or 1 and exclusively contained the atoms H, C, O, N (Figures S5 and S6 in supplemental material). The most frequent substructure, an oxygen atom with a double bond, was found in about 84% of all

fingerprints in the ms2structures dataset and in about 81% of all fingerprints in the biostructures dataset (see supplemental material). RDKit fingerprints contain many more substructures that occur in a large fraction of all molecules, in agreement with Figure 2. The most frequent substructures here are C-C sequences of various lengths as well as other trivial motifs like C-O or C=O (Figure S7 in supplemental material).

On the ms2structures dataset, we computed the 100 nearest neighbors for each compound using different fingerprints, including an IDF-weighted Morgan count fingerprint. Those nearest neighbors and their respective similarities were then used to compute 2D positions using UMAP [30] as shown in Figure 10.

While scatter plots from dimensionality reduction techniques are inherently hard to compare quantitatively, it can clearly be noted that the 37,811 compounds of the ms2structures dataset are more evenly distributed when using IDF-weighted fingerprints. It can also be noted that count fingerprints lead to more separation of molecule masses. This is even more pronounced when using RDKit fingerprints (see Figure S8, supplemental material).

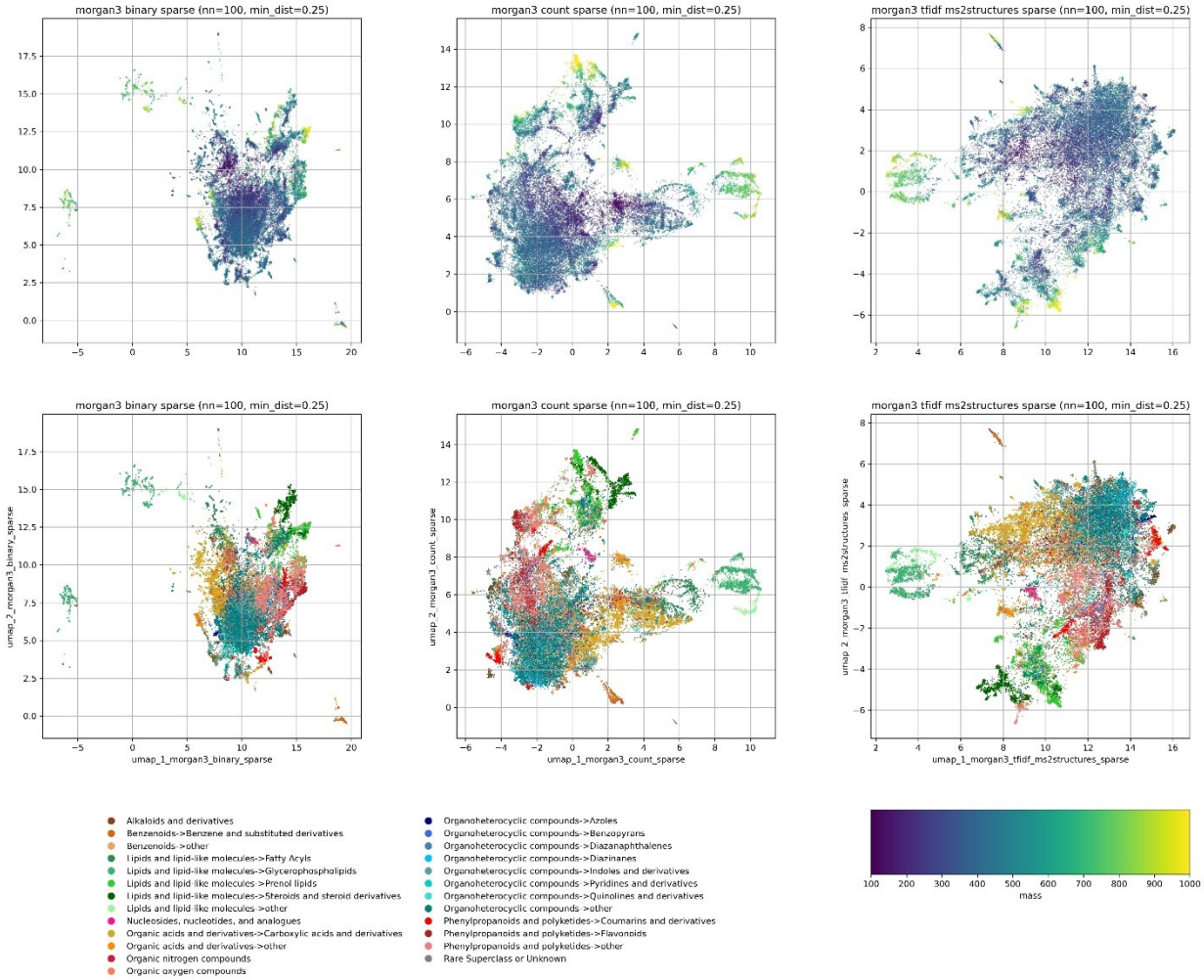

Figure 10 UMAP-based 2D coordinates of all 38,711 compounds in the ms2structures dataset were used to visualize the chemical space of this dataset according to different similarity measures. UMAP was run using the 100 nearest neighbors (and their similarity scores) computed using Tanimoto scores of Morgan-3 binary sparse fingerprints (left panel), of Morgan-3 count sparse fingerprints (center panel) as well as of IDF-weighted Morgan-3 count sparse

fingerprints (right panel). In the top row plots every molecule is colored by its molecular mass (in Da), and in the bottom row plots the molecules are colored based on their chemical class according to Classyfire [42] (less common classes were merged to obtain a limited number of categories for better visualization).

Discussion

Molecular fingerprint–based similarity measures and many of the limitations we observe are not exactly novelties. However, as dataset sizes grow and analyses become more computationally demanding, the criteria for suitable similarity measures must evolve. Our results provide guidance to avoid common pitfalls and make informed choices in future studies.

Conscious fingerprint selection

We would like to start by stating the obvious: different fingerprints will often mean entirely different concepts of what similar molecules will look like, thereby touching on the introduction question of *What does it mean for two molecules to be similar?*
Commonly used fingerprints, such as RDKit topological and circular (Morgan) fingerprints, are often selected arbitrarily [19], yet we found their similarity scores behave very differently. We neither can, nor want, to judge which fingerprint type functions best for a particular dataset or research question. But the choice might often be more fundamental than anticipated.
A substantial fraction of pairs that one fingerprint-based score might place in the top 5% may fall below the median for another score (Figure 4). This alone makes the fingerprint choice highly relevant for tasks that depend on the full range of similarity scores, notably when those are used as optimization target in the training of machine-learning models as it is common practice [8,32,39,43,44]. Similarly, search and ranking results vary substantially with fingerprint choice (Figure 8).
While our study covers only a subset of fingerprint types, future work should evaluate additional representations. We expect that in some cases, decisions may also depend on more specific need, such as compactness or explainability, as for instance the recently proposed biosynfoni fingerprint [16]. A major challenge for any broader exploration, however, will be the wide range of possibilities. Common, established libraries such as RDKit[33] and chemfp[45] , already provide a large choice of different fingerprint algorithms, often coming with multiple key parameters. Additional variations tested in this work, such as tfidf or log weighing, will increase the number of combinations even further. We hope that this work will help the cheminformatics community to find suitable benchmarking approaches for future large-scale comparisons.

Key limitations and mitigations

None of the explored algorithms will always outperform the others, but our results do clearly show some important limitations that need consideration.

Bit Collisions

Bit collisions are a well-known issue with fixed-length fingerprints. But while some studies note considerable decreases in performance [46], others find no noteworthy effects[1]. We have the

impression that bit collisions are often treated as a relatively minor, hypothetical issue that can be easily mitigated by adhering to common vector sizes.
With growing dataset sizes, and depending on the fingerprint choice, however, our experiments show that bit collisions can lead to highly detrimental effects. In general, this will mostly affect fingerprints with high occupation rates, such as RDKit or MAP4 (Figure 2 and Figure 6). Common choices for the fingerprint size are 1024 or 2048 bits, which is also the default value in RDKit. We have chosen an even more optimistic baseline of 4096 bits, and yet we observed many arbitrary, unpredictable results and failures. In particular for larger compounds, MAP4 displayed a very high bit occupation and often wrongly assigned pairs of two larger compounds a high chemical similarity (Figure 4) when fixed-sized vectors were used.
A comparison of similarities between 37,811 compounds based on 4096-bit and sparse fingerprints revealed that for a large fraction of pairs, the classical approach of using a fixed-sized bit vector led to a very considerable overestimation of similarities when using RDKit or MAP4 fingerprints (Figure 5). In contrast, Morgan fingerprints, similar to the ones used in [1], indeed suffer very little from bit collisions due to their much lower bit occupation rates.

More broadly, we observe that all fingerprints will have certain bits with much higher occupation frequencies (Figure 2). The more pronounced this effect, the more it will raise the average Tanimoto scores, which is in line with much earlier calculations [47] and explains the generally higher Tanimoto scores when using RDKit or MAP4 fingerprints compared to Morgan fingerprints. Whether a Tanimoto score of 0.4 reflects a particularly high similarity or not, therefore largely depends on the type of fingerprint that was used. In this work, we frequently moved to percentile scores to compensate for this effect in our comparisons.
More problematic is the mass dependence of similarity scores: identical Tanimoto values between small and large molecules convey different structural implications (Figure 7), biasing global searches toward larger compounds.

Mitigation: Consider Sparse Fingerprints
Sparse implementations of RDKit or MAP4, or switching to count-based or larger-radius Morgan fingerprints, largely mitigated the negative effects described above. Yet even though circular fingerprints seem much less susceptible to issues due to bit collisions, at least for the setting used in this study, it seems justified to generally question the concept of fixed-size fingerprints for large-scale chemical similarity computations.
Current computational capabilities easily allow a shift to larger vector sizes of 4096 bits and more, thereby lowering the risk of bit collisions. A more straightforward solution to this problem, however, would be a shift to sparse fingerprints if the type of fingerprint allows for efficient computation. For Morgan fingerprints, we found that this even improved similarity computation times while eradicating the risk of potential bit collisions. This performance gain will likely no longer hold when shifting to much larger compounds such as peptides. For high bit occupation fingerprints such as RDKit and MAP4, a sparse implementation can sometimes lead to fingerprints with thousands of bits. Moving to much larger compounds might then lead to computational challenges.
In all cases, we found that a shift to sparse fingerprints resulted in notably, sometimes drastically, higher correlation with graph-based molecular similarity which was here measured

using rascalMCES (Figure 9). Even though we cannot simply consider that as universal ground truth for molecular similarity, we still interpret this as a clear sign of improved consistency. Potential computational performance restrictions aside, we therefore generally consider sparse fingerprints a clear improvement over folded ones.

Missing specificity

Binary fingerprints only encode the presence or absence of a particular substructure, disregarding its multiplicity (**Figure 1**). Consequently, motif repetitions cannot be identified, which can lead to similar or identical fingerprints for (very) different compounds. For instance, using binary RDKit binary fingerprints on the biostructures dataset, we observed that approximately 20% of compounds shared fingerprint duplicates despite differing chemical structures, and this prevalence persisted, albeit at lower rates, with smaller-radius Morgan fingerprints.

Alarmingly, a significant portion of these duplicates involved compounds with substantial mass differences (Figure 3), underscoring the inability of purely binary, fixed-radius approaches to distinguish molecules that differ quantitatively rather than qualitatively.

Mitigation: larger radius counts or other, more specific bits

Transitioning to count-based fingerprints substantially alleviates this issue for both RDKit and Morgan-type fingerprints. While some fingerprint duplication remains, it occurs predominantly between molecules of near-identical mass (Figure 3).

Increasing the radius of Morgan fingerprints further reduces duplication rates, confirming that more extensive substructure enumeration improves the fingerprints' specificity. Alternatively, MAP4 fingerprints, by encoding combinations of substructures along with their pairwise distances, generate a far more diverse fingerprint space, resulting in a negligible number of duplicates even for complex biostructures.

Counts and beyond

Not all fingerprint types allow consistent count variants. Circular (Morgan) fingerprints natively support count variants, offering a straightforward switch from presence/absence to occurrence frequency. The classical RDKit topological fingerprint also provides an option to compute count-based fingerprints in RDKit[33]. These counts, however, do not necessarily reflect true substructure frequencies. Because the algorithm does not de-duplicate overlapping features, structures containing ring systems or internal loops, such as aromatic rings, can yield inflated counts where the same motif is registered multiple times. Exploring atom or bond disjoint variants would be an interesting future road but was beyond the scope of this work.

Even when a fingerprint type supports both binary and count representations, the decision to use counts remains nuanced. Count fingerprints generally exhibit higher discriminative power: they substantially reduce the number of fingerprint duplicates, particularly among compounds with large mass differences, and align with reports demonstrating improved performance of count-based descriptors in machine learning tasks and environmental screening models [48]. Moreover, count vectors often outperform binary variants in virtual screening benchmarks, yielding better hit rates among top-ranked candidates. Therefore, when in doubt, count fingerprints are often preferable.

On the other hand, count vectors may “overreact” to highly repetitive motifs, for instance, long aliphatic chains in lipid-like molecules. This can skew similarity metrics towards compounds that share generic repetitive features rather than chemically distinctive ones.

In our experiments, we also explored a log-count transformation, where substructure counts were scaled logarithmically. For Morgan-type fingerprints, this approach yielded superior correlation with maximum common substructure similarity (Figure 9), suggesting that moderate compression of extreme count values can mitigate the overemphasis of repetitive patterns. Going beyond simple counting, one could adopt dataset-adaptive scaling strategies akin to TF-IDF weighting in natural language processing: bits corresponding to ubiquitous substructures within a chemical library would be down-weighted relative to rarer, more specific fragments. For example, in our ms2structures dataset, over 84% of compounds contained a substructure featuring a doubly bonded oxygen atom (e.g., a carbonyl or similar group); A TF-IDF-like scheme would assign a lower weight to this trivial motif while highlighting distinctive ring systems or functional group arrangements.
We anticipate that such approaches will further enhance the suitability of fingerprint-based similarity measures for diverse cheminformatics applications, for instance, by enabling an adjustment to specific chemical reference sets or chemical spaces.

In summary, we recommend a more conscious choice of the fingerprints used. Depending on the dataset and the question in mind, some of the described limitations should be considered, but data and research question might also help determine which concept of similarity is best suited for the task. If, for instance, the actual length of long carbon chains is crucial, a count-based fingerprint is mandatory, whereas if it is more about the presence of specific chemical groups, log-count or binary fingerprints could be more adequate.
We also realized that it can be very promising to reevaluate and reiterate established concepts. For instance, while Morgan-2 and Morgan-3 are commonly used fingerprints, we found that larger radii often work better, and so does the switch to sparse implementations and counts (or scaled counts). On the tasks run in this work, such Morgan-9 or Morgan-12 fingerprints were certainly not so good as to rule them all, but they generally performed well overall, making them a very good starting point.

Acknowledgements
We want to thank Andrew Dalke, Niek de Jonge and Justin J.J. van der Hooft for their very helpful feedback on an earlier version of this manuscript.
This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 528775510.

Authors' contributions
FH conceived the study, designed the experiments, and supervised the project. **FH** and **JP** wrote the code, performed the experiments, analyzed the data, and wrote the manuscript.

References

(1) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *Journal of Cheminformatics* **2013**, *5* (1), 26. https://doi.org/10.1186/1758-2946-5-26.

(2) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.

(3) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722–730. https://doi.org/10.1021/ar500432k.

(4) Medina-Franco, J. L.; Martinez-Mayorga, K.; Giulianotti, M. A.; Houghten, R. A.; Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Current Computer - Aided Drug Design* **2008**, *4* (4), 322–333. https://doi.org/10.2174/157340908786786010.

(5) Cihan Sorkun, M.; Mullaj, D.; Koelman, J. M. V. A.; Er, S. ChemPlot, a Python Library for Chemical Space Visualization. *Chemistry–Methods* **2022**, *2* (7), e202200005. https://doi.org/10.1002/cmtd.202200005.

(6) Kretschmer, F.; Seipp, J.; Ludwig, M.; Klau, G. W.; Böcker, S. Coverage Bias in Small Molecule Machine Learning. *Nat Commun* **2025**, *16*, 554. https://doi.org/10.1038/s41467-024-55462-w.

(7) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: De Novo Structure Generation from Mass Spectra. *Nat Methods* **2022**, *19* (7), 865–870. https://doi.org/10.1038/s41592-022-01486-3.

(8) Bushuiev, R.; Bushuiev, A.; de Jonge, N. F.; Young, A.; Kretschmer, F.; Samusevich, R.; Heirman, J.; Wang, F.; Zhang, L.; Dührkop, K.; Ludwig, M.; Haupt, N. A.; Kalia, A.; Brungs, C.; Schmid, R.; Greiner, R.; Wang, B.; Wishart, D. S.; Liu, L.-P.; Rousu, J.; Bittremieux, W.; Rost, H.; Mak, T. D.; Hassoun, S.; Huber, F.; van der Hooft, J. J.; Stravs, M. A.; Böcker, S.; Sivic, J.; Pluskal, T. MassSpecGym: A Benchmark for the Discovery and Identification of Molecules. *Advances in Neural Information Processing Systems* **2024**, *37*, 110010–110027.

(9) Kim, H. W.; Zhang, C.; Reher, R.; Wang, M.; Alexander, K. L.; Nothias, L.-F.; Han, Y. K.; Shin, H.; Lee, K. Y.; Lee, K. H.; Kim, M. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. DeepSAT: Learning Molecular Structures from Nuclear Magnetic Resonance Data. *Journal of Cheminformatics* **2023**, *15* (1), 71. https://doi.org/10.1186/s13321-023-00738-4.

(10) Sheridan, R. P.; Kearsley, S. K. Why Do We Need so Many Chemical Similarity Search Methods? *Drug Discovery Today* **2002**, *7* (17), 903–911. https://doi.org/10.1016/S1359-6446(02)02411-X.

(11) López-Pérez, K.; Avellaneda-Tamayo, J. F.; Chen, L.; López-López, E.; Juárez-Mercado, K. E.; Medina-Franco, J. L.; Miranda-Quintana, R. A. Molecular Similarity: Theory, Applications, and Perspectives. *Artificial Intelligence Chemistry* **2024**, *2* (2), 100077. https://doi.org/10.1016/j.aichem.2024.100077.

(12) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49* (1), 108–119. https://doi.org/10.1021/ci800249s.

(13) Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem. Inf. Model.* **2012**, *52* (11), 2884–2901. https://doi.org/10.1021/ci300261r.

(14) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *Journal of Cheminformatics* **2020**, *12* (1), 43. https://doi.org/10.1186/s13321-020-00445-4.

(15) Orsi, M.; Reymond, J.-L. One Chiral Fingerprint to Find Them All. *Journal of Cheminformatics* **2024**, *16* (1), 53. https://doi.org/10.1186/s13321-024-00849-6.

(16) Nollen, L.-M.; Meijer, D.; Sorokina, M.; Hooft, J. van der. Biosynfoni: A Biosynthesis-Informed and Interpretable Lightweight Molecular Fingerprint. ChemRxiv February 12, 2025. https://doi.org/10.26434/chemrxiv-2025-cwq74.

(17) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *Journal of Cheminformatics* **2015**, *7* (1), 20. https://doi.org/10.1186/s13321-015-0069-3.

(18) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (3), 379–386. https://doi.org/10.1021/ci970437z.

(19) Yang, J.; Cai, Y.; Zhao, K.; Xie, H.; Chen, X. Concepts and Applications of Chemical Fingerprint for Hit and Lead Screening. *Drug Discovery Today* **2022**, *27* (11), 103356. https://doi.org/10.1016/j.drudis.2022.103356.
(20) Boldini, D.; Ballabio, D.; Consonni, V.; Todeschini, R.; Grisoni, F.; Sieber, S. A. Effectiveness of Molecular Fingerprints for Exploring the Chemical Space of Natural Products. *J Cheminform* **2024**, *16* (1), 35. https://doi.org/10.1186/s13321-024-00830-3.
(21) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.
(22) Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *The Computer Journal* **2002**, *45* (6), 631–644. https://doi.org/10.1093/comjnl/45.6.631.
(23) Ehrlich, H.-C.; Rarey, M. Maximum Common Subgraph Isomorphism Algorithms and Their Applications in Molecular Science: A Review. *WIREs Computational Molecular Science* **2011**, *1* (1), 68–79. https://doi.org/10.1002/wcms.5.
(24) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem. Sci.* **2019**, *10* (6), 1692–1701. https://doi.org/10.1039/C8SC04175J.
(25) Li, Z.; Jiang, M.; Wang, S.; Zhang, S. Deep Learning Methods for Molecular Representation and Property Prediction. *Drug Discovery Today* **2022**, *27* (12), 103373. https://doi.org/10.1016/j.drudis.2022.103373.
(26) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A Review of Molecular Representation in the Age of Machine Learning. *WIREs Computational Molecular Science* **2022**, *12* (5), e1603. https://doi.org/10.1002/wcms.1603.
(27) Yüksel, A.; Ulusoy, E.; Ünlü, A.; Doğan, T. SELFormer: Molecular Representation Learning via SELFIES Language Models. *Mach. Learn.: Sci. Technol.* **2023**, *4* (2), 025035. https://doi.org/10.1088/2632-2153/acdb30.
(28) Medina-Franco, J. L.; Chávez-Hernández, A. L.; López-López, E.; Saldívar-González, F. I. Chemical Multiverse: An Expanded View of Chemical Space. *Molecular Informatics* **2022**, *41* (11), 2200116. https://doi.org/10.1002/minf.202200116.
(29) Maaten, L. van der; Hinton, G. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* **2008**, *9* (86), 2579–2605.
(30) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv September 17, 2020. https://doi.org/10.48550/arXiv.1802.03426.
(31) Gedeck, P.; Rohde, B.; Bartels, C. QSAR − How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46* (5), 1924–1936. https://doi.org/10.1021/ci050413p.
(32) Jonge, N. F. de; Chekmeneva, E.; Schmid, R.; Joas, D.; Truong, L.-J.; Hooft, J. J. J. van der; Huber, F. Bridging Polarities in Metabolomics: Cross-Ionization Mode Chemical Similarity Prediction between Tandem Mass Spectra. bioRxiv January 24, 2025, p 2024.03.25.586580. https://doi.org/10.1101/2024.03.25.586580.
(33) Landrum, G. *RDKit: Open-source cheminformatics*. http://www.rdkit.org.
(34) Huber, F.; Verhoeven, S.; Meijer, C.; Spreeuw, H.; Castilla, E. M. V.; Geng, C.; Hooft, J. J. j van der; Rogers, S.; Belloum, A.; Diblen, F.; Spaaks, J. H. Matchms - Processing and Similarity Evaluation of Mass Spectrometry Data. *Journal of Open Source Software* **2020**, *5* (52), 2411. https://doi.org/10.21105/joss.02411.
(35) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.
(36) O'Boyle, N. M.; Sayle, R. A. Comparing Structural Fingerprints Using a Literature-Based Similarity Benchmark. *Journal of Cheminformatics* **2016**, *8* (1), 36. https://doi.org/10.1186/s13321-016-0148-0.

(37) Lam, S. K.; Pitrou, A.; Seibert, S. Numba: A LLVM-Based Python JIT Compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*; LLVM '15; Association for Computing Machinery: Austin, Texas, 2015; pp 1–6. https://doi.org/10.1145/2833157.2833162.
(38) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 819–828. https://doi.org/10.1021/ci034001x.
(39) Huber, F.; Van Der Burg, S.; Van Der Hooft, J. J. J.; Ridder, L. MS2DeepScore: A Novel Deep Learning Similarity Measure to Compare Tandem Mass Spectra. *J Cheminform* **2021**, *13* (1), 84. https://doi.org/10.1186/s13321-021-00558-4.
(40) Bohde, M.; Manjrekar, M.; Wang, R.; Ji, S.; Coley, C. W. DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra. arXiv May 27, 2025. https://doi.org/10.48550/arXiv.2502.09571.
(41) Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
(42) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *Journal of Cheminformatics* **2016**, *8* (1), 61. https://doi.org/10.1186/s13321-016-0174-y.
(43) Hong, Y.; Ye, Y.; Tang, H. Machine Learning in Small-Molecule Mass Spectrometry. *Annual Review of Analytical Chemistry* **2025**, *18* (Volume 18, 2025), 193–215. https://doi.org/10.1146/annurev-anchem-071224-082157.
(44) de Jonge, N. F.; Mildau, K.; Meijer, D.; Louwen, J. J. R.; Bueschl, C.; Huber, F.; van der Hooft, J. J. J. Good Practices and Recommendations for Using and Benchmarking Computational Metabolomics Metabolite Annotation Tools. *Metabolomics* **2022**, *18* (12), 103. https://doi.org/10.1007/s11306-022-01963-y.
(45) Dalke, A. The Chemfp Project. *Journal of Cheminformatics* **2019**, *11* (1), 76. https://doi.org/10.1186/s13321-019-0398-8.
(46) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model.* **2010**, *50* (5), 771–784. https://doi.org/10.1021/ci100062n.
(47) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (1), 163–166. https://doi.org/10.1021/ci990316u.
(48) Zhong, S.; Guan, X. Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive Regression Models for Water Contaminants' Activities and Properties. *Environ. Sci. Technol.* **2023**, *57* (46), 18193–18202. https://doi.org/10.1021/acs.est.3c02198.