When is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and Its Extreme Values

Pierre Baldi and Ramzi Nasr*

School of Information and Computer Sciences, Institute for Genomics and Bioinformatics, University of California, Irvine, Irvine, California 92697-3435

Received January 6, 2010

As repositories of chemical molecules continue to expand and become more open, it becomes increasingly important to develop tools to search them efficiently and assess the statistical significance of chemical similarity scores. Here, we develop a general framework for understanding, modeling, predicting, and approximating the distribution of chemical similarity scores and its extreme values in large databases. The framework can be applied to different chemical representations and similarity measures but is demonstrated here using the most common binary fingerprints with the Tanimoto similarity measure. After introducing several probabilistic models of fingerprints, including the Conditional Gaussian Uniform model, we show that the distribution of Tanimoto scores can be approximated by the distribution of the ratio of two correlated Normal random variables associated with the corresponding unions and intersections. This remains true also when the distribution of similarity scores is conditioned on the size of the query molecules to derive more fine-grained results and improve chemical retrieval. The corresponding extreme value distributions for the maximum scores are approximated by Weibull distributions. From these various distributions and their analytical forms, *Z*-scores, *E*-values, and *p*-values are derived to assess the significance of similarity scores. In addition, the framework also allows one to predict the value of standard chemical retrieval metrics, such as sensitivity and specificity at fixed thresholds, or receiver operating characteristic (ROC) curves at multiple thresholds, and to detect outliers in the form of atypical molecules. Numerous and diverse experiments that have been performed, in part with large sets of molecules from the ChemDB, show remarkable agreement between theory and empirical results.

1. INTRODUCTION

As chemical repositories of molecules continue to grow and become more open,[1−5] it becomes increasingly important to develop the tools to search them efficiently. In one of the most typical settings, a query molecule is used to search millions of other compounds, not only for exact matches, but also frequently for approximate similarity matches. In a drug discovery project, for instance, one may be interested in retrieving all the commercially available compounds that are “similar” to a given lead, with the objective of finding compounds with better physical, chemical, biological, or pharmacological properties.

Of course, the idea of searching for molecular “cousins” is not new; it constitutes one of the pillars of bioinformatics, where one routinely searches for homologues of nucleotide or amino acid sequences. Search tools such as BLAST[6] and its significance “E-scores” have become de facto standards of modern biology, and they have driven the exponential expansion of bioinformatics methods in the life sciences.

In chemoinformatics, several approaches have been developed for chemical searches, including different molecular representations and different similarity scores. However, no consensus tool such as BLAST has emerged, for several reasons. Some of the reasons involve the cultural differences between the two fields, especially in terms of openness and data sharing. However, there are also more technical and fundamental reasons—in particular, there has been no systematic derivation of a theory that can account for molecular similarity scores and their distributions and significance levels. As a result, many existing search engines do not return a score with the molecules they retrieve, let alone any measure of significance.

Examples of fundamental questions one would like to address include: What threshold should one use to assess significance in a typical search? For instance, is a Tanimoto score of 0.5 significant or not? And how many molecules with a similarity score above 0.5 should one expect to find? How do the answers to these questions depend on the size of the database being queried, or the type of queries used? Clear answers to these questions are important for developing better standards in chemoinformatics and unifying existing search methods for assessing the significance of a similarity score, and ultimately for better understanding the nature of chemical space.

These questions are addressed here systematically, by conducting a detailed empirical and theoretical study of chemical similarity scores and their extreme values. Surprisingly rare previous work related to these questions include an interesting study by Keiser et al.,[7] which used the empirical fitting of distributions to extreme chemical similarity scores but does not derive a predictive mathematical theory of chemical scores and their extremes values, and a short preliminary report of some of our own results.[8] Here,

* Author to whom correspondence should be addressed. E-mail: pfbaldi@ics.uci.edu.

10.1021/ci100010v © 2010 American Chemical Society
Published on Web 06/14/2010

we provide a more general, complete, and self-contained treatment of these questions, including both new theoretical and new simulation results. In particular, we extend the previous work by studying several different ways of assessing the significance of chemical similarity scores, by analyzing, in detail, how the results depend on the parameters of the query molecule, as well as the size of the database being searched, by applying the general framework to the analysis and prediction of ROC curves for molecular retrieval, by applying the general framework to the detection of outlier molecules, and by providing a more complete and predictive mathematical theory of the distribution of similarity scores and its extreme values.

The rest of this paper is organized as follows. Section 2 defines the molecular representations and similarity scores that are used throughout the study. Sections 3 and 4 develop the probabilistic models required to both approximate empirical distributions of similarity scores and create random background models against which significance can be assessed. Section 5 presents the main theory for the distribution of chemical similarity scores, followed by section 6, which presents the theory for the distribution of the extreme values of the score distributions. Corresponding experimental results to illustrate and corroborate the theory are described in sections 7 and 8, followed by a discussion and conclusions section (section 9). To improve the readability, the details of the mathematical derivations are given in A1 and B1.

2. MOLECULAR REPRESENTATIONS AND SIMILARITY SCORES

Many different representations and similarity scores have been developed in chemoinformatics. The methods to be described here are broadly applicable; however, for exposition purposes, we illustrate the theory using the framework that is most commonly used across many different chemoinformatics platforms, namely, binary fingerprint representations with Tanimoto similarity scores. When appropriate, we also briefly describe how the same approach can be extended to other implementations and settings.

2.1. Molecular Representations: Fingerprints. Multiple representations have been developed for small molecules, from one-dimensional (1D) SMILES strings to three-dimensional (3D) pharmacophores,[9] and different representations can be used for different purposes. To search large databases of compounds by similarity, most modern chemoinformatics systems use a fingerprint vector representation[9−15] whereby a molecule is represented by a vector whose components index the presence/absence, or the number of occurrences, of a particular functional group, feature, or substructure in the molecular bond graph. Because binary fingerprints are used in the great majority of cases, here, we present the theory for these fingerprints, but it should be clear that the theory can readily be adapted to fingerprints based on counts. We use **A** to denote a molecule and $\vec{A} = (A_i)$ to denote the corresponding fingerprint. We let **A** denote the number of 1-bits in the fingerprint $\vec{A}$ ($A = |\vec{A}|$).

In early chemoinformatics systems, fingerprint vectors were relatively short, containing typically a few dozen components selected from a small set of features, handpicked by chemists. In most modern systems, however, the major trend is toward the combinatorial construction of extremely long feature vectors with several components N that can vary over the range of 10^3-10^6, depending on the set of features. Examples of typical features include all possible labeled paths or labeled trees, up to a certain depth. The advantage of these longer, combinatorially based representations is 2-fold. First, they do not require expert chemical knowledge, which may be incomplete or unavailable. Second, they can support extremely large numbers of compounds containing both existing and unobserved molecular structures, such as those that are starting to become available in public repositories and commercial catalogs, as well as the recursively enumerable space of virtual molecules.[16] The particular nature of the fingerprint components is not essential for the theory to be presented. To illustrate the principles, in the simulations, we have used both fingerprints based on labeled paths and fingerprints based on labeled shallow trees with qualitatively similar results. For completeness, the details of the fingerprints used in the simulations are given below in the Data subsection. For the sake of brevity and consistency, the examples reported in the results are derived primarily using fingerprints based on paths.

2.2. Fingerprint Compression. In many chemoinformatics systems, the long sparse fingerprint vectors are often compressed to much shorter and denser binary fingerprint vectors. The most widely used method of compression is a lossy compression method based on the application of the logical OR operator to the binary fingerprint vector after modulo wrapping to 512, 1024, or 2048 bits.[12] Other more-efficient lossless methods of compression have recently been developed.[15] With the proper and obvious adjustments, our results are applicable to both lossy compressed and uncompressed fingerprints. Because these are widely used, the majority of the simulation examples that we report are obtained using modulo-OR compressed binary fingerprints of length $N = 1024$. Because of their shorter length, these fingerprints also have the advantage of accelerating Monte Carlo sampling simulations.

2.3. Similarity Scores. Several similarity measures have been developed for molecular fingerprints.[17,18] Given two molecules $\mathcal{A}$ and $\mathcal{B}$, the Tanimoto similarity score is given by

$$S(\mathcal{A}, \mathcal{B}) = S(\vec{A}, \vec{B}) = (A \cap B)/(A \cup B) \qquad (1)$$

Here, $(A \cap B)$ denotes the size of the intersection (i.e., the number of 1-bits common to $\vec{A}$ and $\vec{B}$) and $(A \cup B)$ denotes the size of the union (i.e., the number of 1-bits in $\vec{A}$ or $\vec{B}$). Because the Tanimoto similarity is, by far, the most widely used, the theory and experimental results reported here are based on the Tanimoto similarity. However, we also briefly describe how the same theory can be extended to other measures. Because Tanimoto similarity scores are built from intersections and unions, it will be natural to begin the theoretical analysis by studying the distribution of these intersections and unions, in particular, their means, variances, and covariances.

2.4. Data. In the simulations, we illustrate the methods using fingerprints that are either randomly generated using one of the stochastic models described in section 3, or randomly selected from the 5 M molecules or so available in the ChemDB database.[1] In the case of the actual molecules, we use fingerprints associated with two schemes:[15] labeled paths of length up to eight (i.e., 9 atoms and 8 bonds), or labeled circular substructures of depth up to two, with Element (E) and Extended Connectivity (EC) labeling. In the first scheme, which will be referred to as *paths* throughout the paper, for each chemical, we extract all labeled paths of length up to eight, starting from each vertex and using depth-first traversal of the edges in the corresponding molecular graph. For this scheme, each vertex is labeled by the element (C, N, O, etc.) of the corresponding atom and each edge is labeled by the type (single, double, triple, aromatic, and amide) of the corresponding bond. This scheme is closely related to the scheme used in many existing chemoinformatics systems, including the Daylight system.[12] In the second scheme, for each chemical, we extract every circular substructure of depth up to two from the corresponding molecular graph. Circular substructures (see Hert et al.,[19] Bender et al.,[20] and Hassan et al.[21]) are fully explored labeled trees of a particular depth, rooted at a particular vertex. For this scheme, molecular graphs are labeled as follows: each vertex is labeled by the element (C, N, O, etc.) and degree (1, 2, 3, etc.) of the corresponding atom, and each edge is labeled as previously described. The degree of a vertex is given by the number of edges incident to that vertex or, equivalently, the number of atoms bonded to the corresponding atom.

Both in the case of randomly generated fingerprints and actual molecular fingerprints, we used both uncompressed fingerprints, corresponding also to lossless compressed fingerprints, as well as lossy compressed fingerprint obtained using the standard modulo-OR compression algorithm to generate fingerprint vectors with a length of 1024. For both the randomly generated fingerprints and actual molecular fingerprints, we run typical simulations using a sample of $n = 100$ queries against background sets, which range from 5000 to 1 million fingerprints, to study the effects associated with database size.

3. PROBABILISTIC MODELS OF FINGERPRINTS

One of the main goals of this work is to derive good statistical models and approximations for the distribution of similarity scores. At the most fundamental level this can be addressed by building probabilistic models of fingerprints. Statistical models of fingerprints are essential for a variety of tasks. For instance, in fingerprint compression, fingerprints can be viewed as "messages" produced by a stochastic source and understanding the statistical regularities of the source is essential for deriving efficient compression algorithms that use short codewords for the most frequent events. Here, statistical models are essential in at least two different ways: (1) to model and approximate the distribution of statistical scores, and (2) to assess significance against a random background. Of course, similar observations can be made in bioinformatics, for instance, to assess the probability of observing a particular sequence or alignment score against a random generative or evolutionary model of protein or DNA sequences. Note that, as a default, we assume that the distribution over the queries is the same as the distribution over the molecules in the database. However, these statistical models can also be used to model particular distributions over the space of queries that may differ from the overall background distribution.

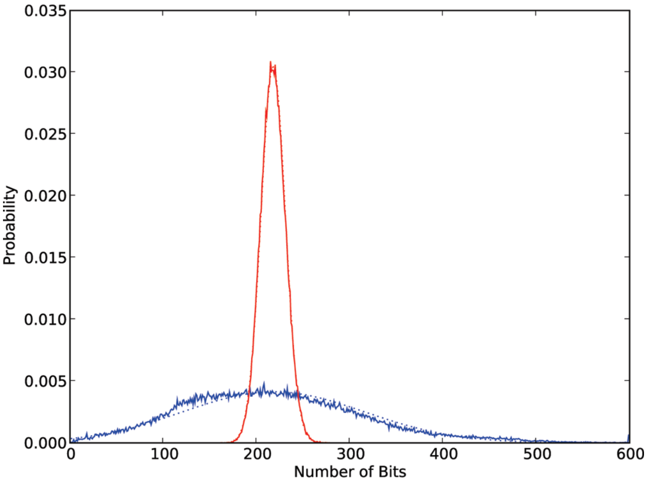


Figure 1. Distributions of the number of 1-bits in fingerprints from the ChemDB (solid blue line) and fingerprints from the matching single-parameter Bernoulli model (solid red line) with $p \approx 205/1024$). Both distributions are constructed using a random sample of 100 000 fingerprints. Although both distributions have similar means, the standard deviations differ significantly. The distributions are also fit using two normal distributions, which approximate the data well (dotted lines).

3.1. Single-Parameter Bernoulli and Binomial Model. The simplest statistical model for binary fingerprints is a sequence of independent identically distributed Bernoulli trials (coin flips) with a probability p of producing a 1-bit, and a probability$q$ ($q = 1 - p$) of producing a 0-bit. This model can be applied to both long fingerprints with a very low p value or to the modulo-OR compressed fingerprints with a higher p value. The coin flip model corresponds to fingerprint features that are randomly ordered and statistically exchangeable, in fact, even independent, and leads to a binomial model $\mathcal{B}(N, p)$ with only two parameters N and p, for the total number of 1-bits in the corresponding fingerprints. The single-parameter Bernouilli model is a weak model of real fingerprints for two reasons. First, the probabilities of the individual components are not identical: some features are more likely to occur than others. Second, the components are not strictly independent. These shortcomings are further addressed in the more-complex models that are described below. Nevertheless, the single-parameter Bernouilli model remains useful, because of its simplicity and tractability, and it provides a point of reference or baseline for other models.

The Bernoulli model can be used to approximate the distribution of fingerprints in an entire database such as ChemDB by setting p equal to the average fingerprint density in the database. If one then compares the behavior of the number A of 1-bits in the Bernoulli generated fingerprints and in the actual database, one typically observes that the average of A is the same in both cases, via the construction of p, but the variance is quite different. The variance A in the Bernoulli-generated fingerprints is given by Npq and is always, at the most, equal to the expectation Np, whereas in large databases of compounds, one typically observes a larger variance (see Figure 1). Generally, a better model for A is provided by a normal distribution $\mathcal{N}(\mu, \sigma^2)$, where the mean $\mu = Np$ and variance $\sigma^2 \neq Npq$ are fitted empirically to the data.

In some analyses, it is useful to consider fingerprints that contain A 1-bits. These can be modeled using Bernoulli coin flips with $p = A/N$, although this is, at best, an approximation, because in the resulting fingerprints, the number of 1-bits is not constant and varies around the mean value A, introducing some additional variability, with respect to the case where A is held fixed (see the section "Conditional Distribution Uniform Model"). Finally, a distribution over queries that is different from the overall database distribution can be modeled using two Bernoulli models: one with parameter r for the queries, and one with parameter p ($p \neq r$) for the database.

3.2. Multiple-Parameter Bernoulli Model. While the coin flip model is useful to derive several approximations, chemical fingerprints clearly have a more-complex structure and their components are not exactly exchangeable, because the individual feature probabilities p_1, ..., p_N are not identical and equal to p but vary significantly. In particular, when the fingerprint components are reordered in decreasing frequency order, they typically follow a power-law distribution,[15] especially in the uncompressed case. The probability of the j-ranked component is given approximately by $p_j = Cj^{-\alpha}$, resulting in a line of slope $-\alpha$ in a log−log plot. Thus, the statistical model at the next level of approximation is that of a sequence of nonstationary independent coin flips where the probability p_j of each coin flip varies. This multiple-parameter Bernoulli model has N parameters: p_1, p_2, ..., and p_N. In this case, using the independence, the expectation of the total number A of 1-bits is given by $\sum_i p_i$ and its variance is given by $\sum_{i=1}^{N} p_i q_i$. Generally, this variance is still an underestimate of the variance observed in actual large databases, despite the larger number of parameters, compared to the single-parameter Bernoulli model (not shown). Similar to the case of the single-parameter Bernouilli model, a distribution over queries different from the overall distribution could be modeled using a multiple-parameter Bernoulli model with a different set of parameters r_1, ..., r_N.

3.3. Conditional Distribution Uniform Model. Both the single-parameter and multiple-parameter Bernoulli models consider the fingerprint components as independent random variables. The Conditional Distribution Uniform Model is an exchangeable model where the components are weakly coupled and, therefore, are not independent. To generate a fingerprint vector under this model, one first samples the value A corresponding to the total number of 1-bits in the fingerprint, using a given distribution, typically a normal distribution (see Figure 1). The model then assumes that, conditioned on the value of A, all fingerprints with A 1-bits are equally probable (uniform distribution). Thus, for example, the Conditional Normal Uniform Model has only three parameters: the mean (μ), the standard deviation (σ^2), and N. Compared to the binomial model, the additional parameter in the Conditional Normal Uniform Model allows for a better fit of the variance of A in the data. As we shall see, for the questions to be considered here, the Conditional Normal Uniform Model performs the best, despite the fact that it does not model the probability differences between different fingerprint components.

3.4. Spin Models. More-complex, second-order models are possible but will not be considered here. These models are essentially spin models from statistical physics; they are also known as Markov Random fields or Boltzmann machines.[22,23] In these models, one also would have to take into account the correlations between pairs of features that can be superimposed over the multiple-parameter Bernoulli model. In real data, these correlations are often (although not always) weak, but not exchangeable, and thus behave differently from those introduced in the Conditional Distribution Uniform Model. The slight improvements in modeling accuracy that may result from spin models generally come at a significant computational cost, because these cannot be solved analytically and therefore cannot be used in a straightforward manner to derive the probability distribution of the similarity scores. The study of spin models is left for future work.

4. PROBABILISTIC MODELS OF DISTRIBUTION SCORES

Although, in the following sections and the Appendix, we show how the distribution of similarity scores can be estimated somehow from "first principles", i.e., from the corresponding probabilistic model of fingerprints, it is also possible to model or approximate the score distribution directly, for instance, by assuming that the scores are approximately normally distributed and obtaining the mean and standard deviation by sampling methods. In a similar way, one can also use a gamma distribution model to completely avoid negative scores, or a beta model to insist on bounded scores between 0 and 1, to model the overall distribution of scores. Another intermediary alternative, which is less direct but still avoids modeling the fingerprints themselves, is to model the intersections and unions that are used to derive the Tanimoto scores, and then try to derive the distribution of the scores from those models. For example, one could consider modeling both the intersections and corresponding unions using two different normal distributions and derive the means and standard deviations of these normal distributions by sampling methods. The commonalities, differences, and tradeoffs between these various modeling and approximation approaches to the distribution of chemical similarity scores will become clear in the following sections. The most complex case, where everything is derived from the probabilistic models of fingerprints, is treated in detail in the Appendix.

5. THE DISTRIBUTION OF THE SIMILARITY SCORES

With these preliminaries in place, we are now prepared to analyze the distribution of Tanimoto scores under the various probabilistic models.

5.1. Main Result. Since the Tanimoto score is the ratio of an intersection over an union, the basic strategy is to first study the distribution of the corresponding intersection and union, as well as their means and variances. Note that the intersection and union generally are not two independent random variables, but have a nonzero correlation that must be estimated analytically or through simulations. In turn, from these results, one can derive a closed-form approximation for the distribution of the Tanimoto scores and its extreme properties. This analysis can be performed using empirical fingerprint data, as well as fingerprints generated by the probabilistic models. Furthermore, the analysis can be conducted by conditioning on the total number of 1-bits contained in the query molecule (A) and the molecules being searched (B), by conditioning only one of these quantities—typically, the number A of 1-bits in the query molecule—and integrating over the other, or with no conditioning at all by integrating over both the fingerprints in the queries and the fingerprints in the database being searched. These forms of conditioning are practically relevant, especially conditioning on A, which will be shown to lead to much better retrieval results.

In all these cases, one generally finds that.

(1) The intersection and union have an approximately normal distribution, with means and variances that can be estimated empirically or computed analytically in the case of the probabilistic models.

(2) The intersection and union have a nonzero (positive) covariance that can be estimated empirically or computed analytically in the case of the probabilistic models.

(3) As a consequence, the distribution of the corresponding Tanimoto scores can be modeled and approximated by the distribution of the ratio of two correlated normal random variables.

These facts are demonstrated in the Results section using simulations. In A1, mathematical proofs are provided for the probabilistic models of fingerprints, together with analytical formulas for the means, variances, and covariances of the intersections and unions.

5.2. Ratio of Two Correlated Normal Random Variables Approximation. Whether one uses the single-parameter Bernoulli/binomial model, multiple-parameter Bernoulli model, or the Conditional Density Uniform Model, or the empirical intersection and union data, in the end, the Tanimoto score distribution can be approximated by the distribution of two correlated normal random variables, approximating the numerator and the denominator. The different models will yield different estimates of the mean, variance, and covariance of the respective normal distributions.

The density of the ratio of two correlated normal random variables has been studied in the literature and can be obtained analytically, although its expression is somewhat involved.[24−27] The probability density for $Z = X/Y$, where $X \approx \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \approx \mathcal{N}(\mu_Y, \sigma_Y^2)$, and $\rho = \mathrm{Corr}(X, Y) \neq \pm 1$ is given by the product of two terms:

$$P_Z(z) = \frac{\sigma_X\sigma_Y\sqrt{1-\rho^2}}{\pi\left(\sigma_Y{}^2z^2 - 2\rho\sigma_X\sigma_Y z + \sigma_X{}^2\right)}\left[\exp\left(-\frac{1}{2}\sup R^2\right) \times \left(1 + \frac{R\Phi(R)}{\phi(R)}\right)\right] \quad (2)$$

or

$$P_Z(z) = \frac{\sigma_X\sigma_Y\sqrt{1-\rho^2}}{\pi\left(\sigma_Y{}^2z^2 - 2\rho\sigma_X\sigma_Y z + \sigma_X{}^2\right)}\left\{\exp\left(-\frac{1}{2}\sup R^2\right) + \sqrt{2\pi}R\Phi(R)\left[\exp-\frac{1}{2}(\sup R^2 - R^2)\right]\right]\right\} \quad (3)$$

where

$$R = R(z) = \frac{(\sigma_Y{}^2\mu_X - \rho\sigma_X\sigma_Y\mu_Y)z - \rho\sigma_X\sigma_Y\mu_X + \sigma_X{}^2\mu_Y}{\sigma_X\sigma_Y\sqrt{1-\rho^2}\sqrt{\sigma_Y{}^2z^2 - 2\rho\sigma_X\sigma_Y z\sigma_X{}^2}} = \frac{\left(\frac{\mu_X}{\sigma_X} - \rho\frac{\mu_Y}{\sigma_Y}\right)z - \left(\rho\frac{\mu_X}{\sigma_X} - \frac{\mu_Y}{\sigma_Y}\right)\frac{\sigma_X}{\sigma_Y}}{\sqrt{1-\rho^2}\sqrt{z^2 - 2\rho\frac{\sigma_X}{\sigma_Y}z + \left(\frac{\sigma_X}{\sigma_Y}\right)^2}} \quad (4)$$

$$\sup R^2 = \frac{\sigma_Y{}^2\mu_X{}^2 - 2\rho\sigma_X\sigma_Y\mu_X\mu_Y + \sigma_X{}^2\mu_Y{}^2}{\sigma_X{}^2\sigma_Y{}^2(1-\rho^2)} = \frac{\left(\frac{\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{\mu_X\mu_Y}{\sigma_X\sigma_Y} + \left(\frac{\mu_Y}{\sigma_Y}\right)^2}{1-\rho^2} \quad (5)$$

and

$$\sup R^2 - R^2 = \frac{(\mu_X - \mu_Y z)^2}{\sigma_Y{}^2z^2 - 2\rho\sigma_X\sigma_Y z + \sigma_X{}^2} = \frac{\left(\frac{\mu_X\sigma_X}{\sigma_X\sigma_Y} - \frac{\mu_Y}{\sigma_Y}z\right)^2}{z^2 - 2\rho\frac{\sigma_X}{\sigma_Y}z + \left(\frac{\sigma_X}{\sigma_Y}\right)^2} \quad (6)$$

Thus, anytime one can approximate the intersections and the unions using two correlated normal random variables, the distribution of the Tanimoto scores can be approximated using eqs 2−6 with $X = I$ and $Y = U$. This approach can be used, for instance, to derive the mean and standard deviation of the Tanimoto scores under various assumptions, including (1) the single- and multiple-parameter Bernoulli models with $p = r$ (or $p_i = r_i$) for the average Tanimoto scores across all queries; (2) the single- and multiple-parameter Bernoulli models with $p \neq r$ (or $p_i \neq r_i$) for queries modeled by a Bernoulli model different from that used for the database being searched; (3) the Conditional Distribution Uniform Model with A fixed, or A integrated over the database distribution, or a distribution over queries; and (4) the empirically derived normal models for the union and intersection averaged over the entire database, or focused on a particular class of molecules.

A Python code implementation for the density of the ratio of two correlated normal random variables (eqs 2−6) is available from the ChemDB chemoinformatics portal (cdb.ics.uci.edu), under "Supplements".

5.3. Extensions to Other Measures. Although we have described the theory for the Tanimoto similarity scores, the same theory can readily be adapted to most other fingerprint similarity measures.[17,18] To observe this, one must note that the other measures consist of algebraic expressions built from $A \cup B$ and $A \cap B$, as well as other obvious terms (such as A, B, and sometimes N). For example, the Tversky measure[28,29] is an important generalization of the Tanimoto measure, which is defined by

$$S_{\alpha\beta}(\vec{A}, \vec{B}) = \frac{A \cap B}{\alpha A + \beta B + (1 - \alpha - \beta)(A \cap B)} \quad (7)$$

where the parameters α and β can be used to tune the search toward substructures or superstructures of the query molecule. The numerator and denominator in the Tversky measure can again be modeled by two correlated normal random variables. The only difference is observed in the mean and variance of the denominator, and its covariance with the numerator. The new mean, variance, and covariance can be computed empirically. They can also be derived analytically for the simple probabilistic models, as described in A1. Similar considerations apply to all the other measures described in refs 17 and 18. Thus, the distribution and statistical properties of all of the other similarity measures[17,18] can readily be derived from the general framework described and presented here.

5.4. Alternatives and Related Approaches. Because the intersections and unions always have positive values, it is also possible, in some cases, to approximate their distributions using gamma distributions. The distribution of Tanimoto scores can then be modeled using the distribution of the ratio of two correlated gamma distributions, for which some theory exists.[30−32] Similarly, in regimes where the finite [0, N] range of the intersections and unions becomes important, the intersection and union can be rescaled by $1/N$ and the corresponding distributions modeled using beta distributions. In this case, the distribution of Tanimoto scores can be modeled using the distribution of the ratio of two correlated beta distributions.[33] Finally, as already mentioned, it is also possible to model or approximate a distribution of Tanimoto scores *directly*, using a normal, gamma, or beta distribution (or a mixture of these distributions) without having to first consider the intersections and unions.

It is not possible to give a general prescription regarding which approach may work best in a practical application, because this may depend on the details and goals of a particular implementation. However, the theory presented provides a general framework for predicting and modeling the distribution of Tanimoto scores that can be adapted to any particular implementation. And using this distribution, it is possible to derive measures and visualization tools to assess the quality and significance of the molecules being retrieved with a given query and the corresponding rates of false positives and false negatives, as described in the next sections.

6. THEORY: Z-SCORES, E-VALUES, P-VALUES, OUTLIERS, AND ROC CURVES

There are various computational approaches for determining the significance of similarity scores. All these approaches are derived from the distribution of similarity scores. Significance scores include Z-scores, E-values, and p-values associated with the extreme value distribution[34−36] of similarity scores. The distribution of similarity scores can also be used to detect outliers and predict Receiver Operating Characteristic (ROC) curves in chemical retrieval. As we shall see, these significance analyses can be performed and yield better results when conditioned based on the size of the queries.

6.1. *Z*-Scores. In the Z-score approach, one simply looks at the distance of a score from the mean of the corresponding family of scores, in terms of the number of standard deviations. Therefore, the Z-score is given by

$$Z = \frac{t - \mu}{\sigma} \tag{8}$$

The parameters μ and σ can be determined either empirically from a database of fingerprints, or using the statistical models described above. While Z-scores can be useful, their focus is on the global mean and standard deviation of the distribution of the scores, not on the tail of extreme values. Thus, we next consider two measures that are more focused on the extreme values.

6.2. *E*-Values. When considering a particular similarity value or selecting a similarity threshold t for a given query, an important consideration is the expected number of hits in the database above that threshold. To use terminology similar to that used for BLAST, we refer to this number as the E-value. From the distribution of scores in a database of size D, the E-value that corresponds to a Tanimoto threshold t is estimated by

$$E = [1 - F(t)] \times D \tag{9}$$

where $F(t)$ is the cumulative distribution of the corresponding similarity scores, which can be approximated using the methods described above.

6.3. Extreme Value Distributions and *P*-Values. The second approach, which focused on extreme values, corresponds to computing p-values. For a given score t, its p-value is the probability of finding a score equal to or greater than t under a random model. Thus, in this case, one is interested in modeling the tail of the distribution of the scores, and, more precisely, the distribution of the maximum score.[34−36] This distribution depends on the size of the database being searched, because, for a given query, and everything else being equal, we can expect the maximum similarity value to increase with the database size.

Consider a query molecule $\mathscr{A}$ used to search a database containing D molecules, yielding D similarity scores t_1, ..., t_D. The cumulative distribution of the maximum (max) is given by

$$F_{\max}(t) = P(max \leq t) = P(t_1 \leq t)...P(t_D \leq t) = F(t)^D \tag{10}$$

under the usual assumption that the scores are independent and identically distributed. Here, $F(t)$ is the cumulative distribution of a single score. A p-value is obtained by computing the probability

$$p = 1 - F_{\max}(t) \tag{11}$$

that the maximum score be larger than t under F. The density of the maximum is obtained by differentiation:

$$f_{\max}(t) = Df(t)[F(t)]^{D-1} \tag{12}$$

where $f(t)$ is the density of a single score. In the case of Tanimoto similarity scores, $f(t)$ can be approximated by the ratio of two correlated normal random variables approach described above, and $F(t)$ is obtained from $f(t)$ by integration. $F(t)$ can also be approximated by[25]

$$F(t) \approx \Phi\left(\frac{\mu_Y t - \mu_X}{\sigma_X \sigma_Y a(t)}\right) \tag{13}$$

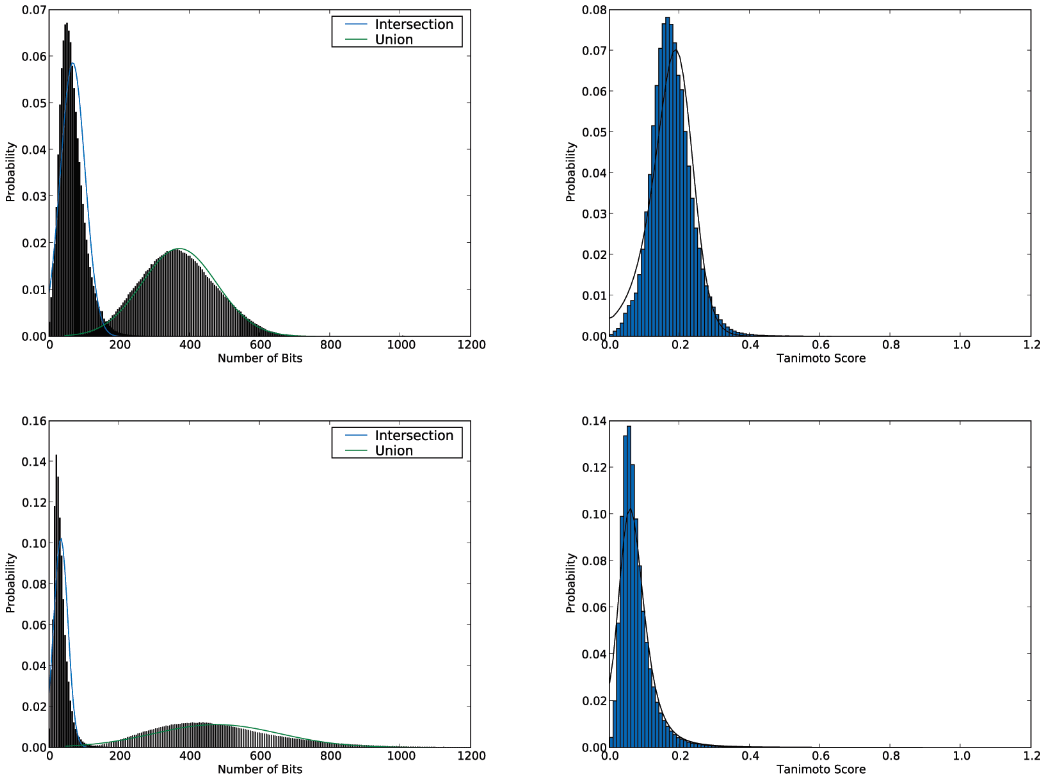


Figure 2. Results obtained with 100 molecules randomly selected from ChemDB used as queries against a sample of 100 000 molecules randomly selected from ChemDB. The two upper figures correspond to fingerprints with lengths of 1024 with modulo OR lossy compression, while the two lower figures correspond to fingerprints with lossless compression (equivalent to uncompressed fingerprints). The figures in the left column display histograms of the sizes of the intersections and unions and their direct normal approximations, in blue and green, respectively. The figures in the right column display histograms of the Tanimoto scores (blue bars), while the solid black line shows the corresponding approximation derived using the ratio of correlated normal random variables approach.

where $\Phi(u) = \int_{-\infty}^{u} (2\pi)^{-1/2} e^{-x^2/2}\, dx$ is the cumulative distribution of the normalized normal distribution and the term $a(t)$ is given by

$$a(t) = \left(\frac{t^2}{{\sigma_X}^2} - \frac{2\rho t}{\sigma_x \sigma_Y} + \frac{1}{{\sigma_Y}^2}\right)^{1/2} \tag{14}$$

This approximation is good when the denominator of the ratio of two correlated normal random variables is positive, with its standard deviation much larger than its average. In any case, by combining eqs 2, 12, and 13, we get

$$f_{\max}(t) \approx D \frac{\sigma_X \sigma_Y \sqrt{1-\rho^2}}{\pi({\sigma_Y}^2 t^2 - 2\rho\sigma_X\sigma_Y t + {\sigma_X}^2)}\left[\exp\left(-\frac{1}{2}\sup R^2\right) \times \left(1 + \frac{R\Phi(R)}{\varphi(R)}\right)\right]\left[\Phi\left(\frac{\mu_Y t - \mu_X}{\sigma_X \sigma_Y a(t)}\right)\right]^{D-1} \tag{15}$$

Finally, because the Tanimoto scores are bounded by one, the theory of extreme value distributions shows that the cumulative distribution of the normalized maximum score n_D, normalized linearly in the form $n_D = a_D max + b_D$ using appropriate sequences a_D and b_D of normalizing constants, converges to a type-III extreme-value distribution, or a Weibull distribution function, of the form

$$F(x) = P(n_D \leq x) = \exp\left[-\left(\frac{\mu - x}{\sigma}\right)^{\xi}\right] \tag{16}$$

The linear normalization can be ignored, because it is absorbed into the parameters of the Weibull distribution. The advantage of the Weibull formula is its suitability for representing $F_{\max}$ in a closed form that can be easily and efficiently computed. How to fit the Weibull distribution to the data, in practice, is described in B1.

6.4. Outliers. The framework allows us to detect molecules that are atypical, within their group, in the following sense. From the framework, we can predict the typical (average) distribution of Tanimoto scores for a given query size S, or the expected number of hits above any given threshold t, given S. Clearly, if we are dealing with an actual query molecule $\mathcal{A}$ with a fingerprint containing $A = S$ 1-bits, if the distribution of observed scores for $\mathcal{A}$ differs from the typical distribution given A, then the molecule $\mathcal{A}$ can be viewed as being atypical within the class of molecules in the database containing A 1-bits in their fingerprints. The difference between the typical distribution of scores for molecules with A 1-bits and the distribution of scores generated by the actual query $\mathcal{A}$ can be measured in many ways, for instance, by using the relative entropy or Kullback Leibler divergence between the two distributions.[37] Similar considerations can be made using the expected number of scores above a given threshold for molecules with A 1-bits versus the actual number observed for molecule $\mathcal{A}$.

6.5. Receiver Operating Characteristic (ROC) Curves. Finally, the general framework can be used to predict false positive and false negative rates, as well as standard Receiver Operating Characteristic (ROC) curves. For conciseness, let us describe the approach for ROC curves, which plot false positive rates on the x-axis versus true positive rates on the y-axis. Consider a set of molecules (e.g., a set of estrogen

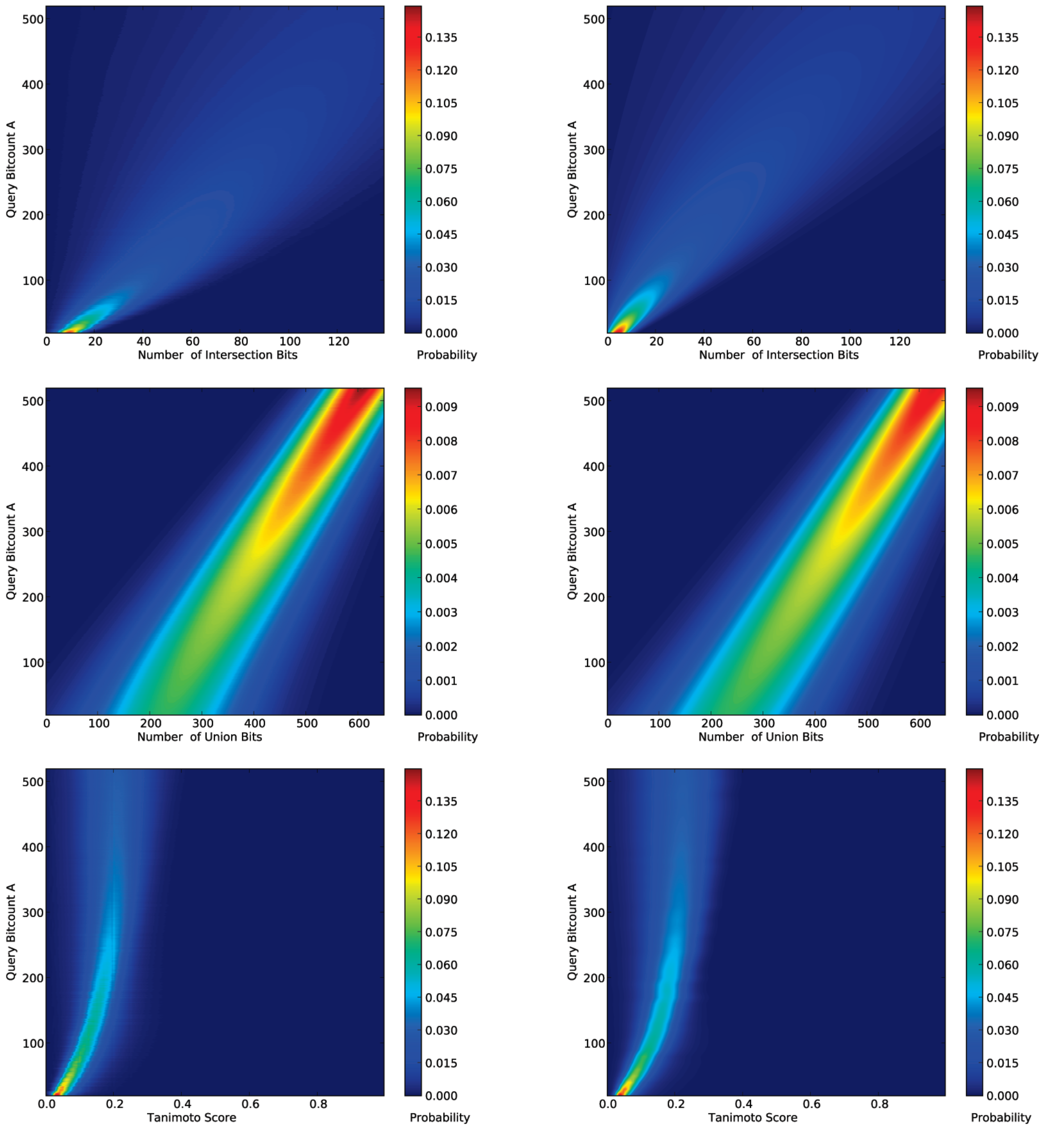


Figure 3. Empirical (left) and predicted (right) heat maps, corresponding to the distribution of the intersections (top), unions (middle), and Tanimoto scores (bottom). The distribution is conditioned on the size of the query molecule, *A*, shown on the vertical axis. The empirical results are obtained using, for each *A*, 100 molecules randomly selected from the molecules in ChemDB with size *A*. The theoretical results of the intersection and union distributions use the Conditional Normal Uniform Model. At each value of *A*, the mean and variance of the intersection and union are obtained from eqs A12, A13, A16, and A19, respectively. The theoretical score distribution is a result of the ratio of correlated normal random variables approximation given by eqs 2−6.

receptor binding molecules) as a set of positive examples used to search a large database for similar molecules. Empirically, or using the ratio of correlated normal random variables approach, one can derive a density *f* and a corresponding cumulative distribution *F* for the similarity scores of the positive examples, and a density *g* and a corresponding cumulative distribution *G* for the similarity scores of the negative examples provided by the overwhelming majority of the molecules in the large database. Thus, for a given threshold *t* on the Tanimoto similarity, the corresponding point on the ROC is easily obtained and given by

$$x = 1 - G(t) \quad \text{and} \quad y = 1 - F(t) \qquad (17)$$

In other words, using continuous approximations, the equation of the ROC curve is given by $y = 1 - F[G^{-1}(1 - x)]$. Similarly, other measures, such as specificity or sensitivity, can be estimated at any given threshold.

Now, armed with this theoretical framework, we can proceed with simulations to demonstrate how the framework can be applied and assess the quality of the corresponding predictions. The following sections describe experimental results obtained using actual molecules from the ChemDB. A large number of experiments were performed and only a sample of the main results is described here, for the sake of brevity. Unless otherwise specified, the results reported here are obtained using path fingerprints compressed to 1024 bits, using lossy modulo-OR compression.

7. SIMULATION RESULTS: THE DISTRIBUTION OF SIMILARITY SCORES

We first examine the quality of the ratio of correlated normal random variables approximation. Figure 2, in the left column, shows the empirical distributions of the sizes of the intersections and unions averaged across the entire database and obtained using Monte Carlo methods, for both lossy compressed fingerprints (upper plots) and uncompressed fingerprints (lower plots), together with their normal approximations. The positive covariance between the intersec-

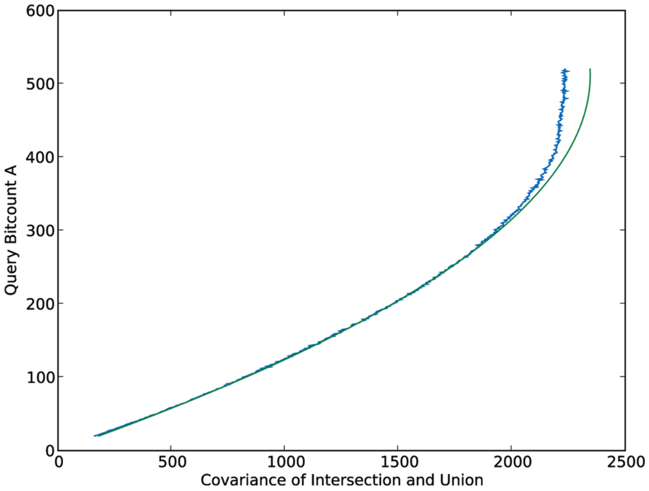


Figure 4. Empirical and theoretical covariance Cov(*I*, *U*) between the intersection and the union, conditioned on the size *A* of the query molecule, shown in blue and green, respectively. Empirical results are obtained by using, for each *A*, 100 molecules randomly selected from the molecules in ChemDB with size *A*. *A* is shown on the vertical axis, for the sake of consistency with the previous heat map figures. Theoretical predictions are derived with the Conditional Normal Uniform Model conditioned based on *A* (see eq A22).

tions and unions is $\mathrm{Cov}(I, U) = 3048.5$ (with corresponding correlation $\mathrm{Corr}(I, U) = 0.82$) for the lossy compressed fingerprints, and $\mathrm{Cov}(I, U) = 1253.2$ ($\mathrm{Corr}(I, U) = 0.35$) for the uncompressed fingerprints. In the right column of Figure 2, one can see the corresponding histogram of Tanimoto scores and the ratio of correlated normal random variables approximation. Overall, the ratio of correlated normal random variables approach approximates the histograms very well in this case, where one is applying the process of averaging over all molecules.

To test whether the ratio of correlated normal random variables approximation works well at a finer-grained level, we repeat a similar experiment but with conditioning based on the size *A* of the query molecules. In fact, in this experiment, we use an even more stringent theoretical model. Instead of fitting normal distributions to the intersections and unions (as in Figure 2), we assume that the data is generated by the Conditional Normal Uniform Model with only two parameters that are fit to the mean and variance of *B* across the entire ChemDB. As described in the Appendix, this gives us analytical formulas for the means and variances of the intersections and unions *for each value* of *A*, as well as their covariances. Figure 3 provides heat maps showing the corresponding empirical and predicted distributions of the intersections (first row) and unions (second row), as a function of *A*. The last row compares the observed Tanimoto score distribution to the predicted Tanimoto score distribution, using the ratio of correlated normals approach. Overall, there is remarkable agreement between the theoretical predictions and the corresponding empirical observations at all values of *A* and at all Tanimoto scores, especially considering that the Conditional Normal Uniform Model used in these heat maps has only two parameters that are fit to the actual data (the mean and variance of *B* in ChemDB).

Similarly, Figure 4 shows how, for each value of *A*, the covariance between the union and the intersection is well-predicted by the Conditional Normal Uniform Model, with a small deviation observed for molecules with a high bit count, where the covariance is slightly smaller than predicted by the theory, probably as a result of a decrease in the variability of the size of the union for queries associated with molecules from ChemDB with a large *A* (the size of the union tends to be close to *A*, because the components in the complement have exceedingly small probabilities).

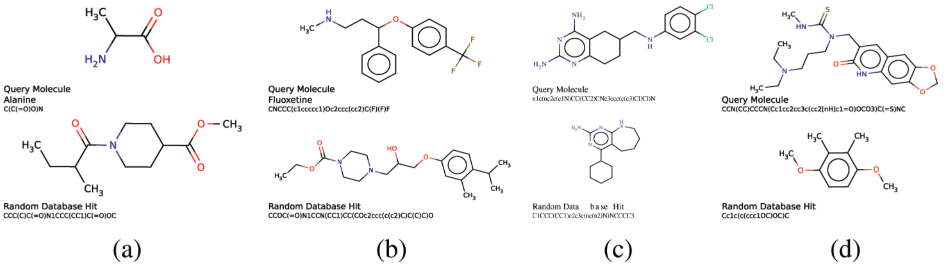
In summary, these results show that the distribution of Tanimoto scores can be modeled, predicted, or approximated accurately with the framework proposed here. Among the simplest models, the Conditional Normal Uniform Model performs best. Conditioning on the size *A* of the query can play an important role, since there are significant variations in the distribution of the scores as *A* varies.

8. ASSESSMENT OF SIGNIFICANCE: *Z*-SCORES, *E*-VALUES, *P*-VALUES, OUTLIERS, AND ROC CURVES

We now turn to the assessment of significance using *Z*-scores, *E*-values, extreme-value distributions and *p*-values, and ROC curves. Figure 5 provides four examples, one in each column, of pairs of molecules where the top molecule can be viewed as the query, and the bottom molecule can be viewed as a potential “hit” retrieved while searching a random subset of 100 000 molecules taken from the ChemDB. The four queries have different sizes corresponding to $A = 16$, 109, 199, and 258. The corresponding four Tanimoto similarity scores are 0.200, 0.400, 0.571, and 0.233. Columns a and d correspond to similar Tanimoto scores, although they should be viewed quite differently, because of the disparity in the size *A* of the corresponding queries, as shown in the following analyses.

8.1. *Z*-Scores. The *Z*-score is the distance of a Tanimoto score from the mean measured in terms of the number of standard deviations. As usual, for a given query of size *A*, the mean and standard deviation can be computed over all molecules, or over molecules of size *A* only. As expected, Figure 5 shows that the *Z*-scores computed over all molecules are not very informative and would indicate that all four pairs of molecules have *Z*-scores above 9 and therefore are significantly very similar. The *Z*-scores computed by conditioning based on *A* are slightly more informative: while they still return three of the matches, corresponding to columns a, b, and c, as highly significant (*Z*-scores above 5) compared to random, they begin to separate these cases from the case of column d, which is scored as not being very significant (*Z*-score of 0.810).

8.2. *E*-Values. As described previously, the *E*-value for a Tanimoto score *t* represents the expected number of “hits” above *t* and can be estimated empirically or predicted from the distribution of the scores and the size *D* of the database (see eq 9). Figure 6 shows again that theoretical *E*-values obtained by conditioning based on *A* are more useful and in the range of the *E*-values observed empirically (here, $D = 100\,000$). The empirical *E*-values, or the predicted *E*-values conditioned on *A*, now clearly separate the similarity in column d as being nonsignificant. The empirical and predicted *E*-value show that, when the size of the molecular fingerprints is taken into consideration, the *E*-value in column d of Figure 6 is considerably *less* significant that the *E*-value in column a of Figure 6, despite the fact that the Tanimoto score in column d (0.233) is higher than the Tanimoto score in column a (0.200). In addition, the *E*-values identify column


Query Molecule
Alanine
C(C(=O)O)N
Random Database Hit
CCC(C)C(=O)N1CCC(CC1)C(=O)OC
(a)
Query Molecule
Fluoxetine
CNCCC(c1ccccc1)Oc2ccc(cc2)C(F)(F)F
Random Database Hit
CCOC(=O)N1CCN(CC1)CC(COc2ccc(c(c2)C)C(C)C)O
(b)
Query Molecule
n1c(nc2c(c1N)CC(CC2)CNc3ccc(c(c3)Cl)Cl)N
Random Data base Hit
C1CCC(CC1)c2c3c(nc(n2)N)NCCCC3
(c)
Query Molecule
CCN(CC)CCCN(Cc1cc2cc3c(cc2[nH]c1=O)OCO3)C(=S)NC
Random Database Hit
Cc1c(c(ccc1OC)OC)C
(d)

Column of Molecules		a	b	c	d
Query Bit Count A		16	109	199	258
Tanimoto	Raw Score	0.200	0.400	0.571	0.233
	Z-Score	9.757	83.566	146.624	22.032
	Z-Score | A	5.630	5.536	8.027	0.810
E-value	Empirical	994	221	0	12594
	Predicted	37890.1	32.4	4.2	17332.7
	Predicted | A	916.3	253.3	0.1	19935.7
p-value	Empirical	1.0	0.88	0.0	1.0
	Predicted	1.0	0.992	0.031	1.0
	Predicted | A	1.0	0.839	0.000	1.0

Figure 5. Schematic depictions of query molecules. The first row shows four query molecules; the second row considers four corresponding potential "hits" in the corresponding columns. The table shows the size A of the four query molecules, followed by the corresponding Tanimoto scores, Z-scores, E-scores, and p-values observed empirically or predicted from the theory with and without conditioning on the size A of the query molecule. Molecules are represented by Daylight-style fingerprints of length 1024 with OR lossy compression.

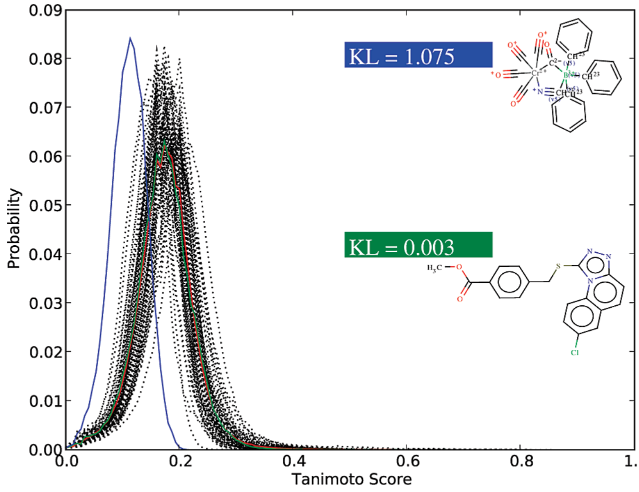


Figure 6. Empirical score distributions for 100 query molecules satisfying $A = 220$. Each black curve is associated with one of the molecules and is obtained by scoring the molecule against a random sample of 100 000 molecules from the ChemDB. The red curve corresponds to the mean of the 100 curves and is essentially identical to the predicted distribution of scores conditioned on $A = 220$. The green curve corresponds to a molecule in the group that is typical and the blue curve corresponds to a molecule that is atypical. The difference between the distributions is measured here, in terms of Kullback–Leibler (KL) divergence or relative entropy.

c as being the only one with really significant similarity (E-value of 0.1).

Figure 7 provides further evidence of the utility of E-values and conditioning on A. This figure is obtained using 55 estrogen-receptor binding molecules,[38] together with a random sample of 100 000 molecules extracted from the ChemDB. For each one of the 55-estrogen receptor ligands and any threshold, the figure essentially plots the number of molecules that have a score above that threshold. Thus, for example, for $t = 0$, all 100 000 molecules have a score above t, corresponding to 55 superimposed dots on the graph. Likewise, for $t = 1$, there are 55 superimposed dots with a vertical value equal to 0, because no molecule scores higher than 1. Note how the number of hits varies greatly in the threshold interval [0.1, 0.3]. The solid red line represents the predicted E-values obtained using the Conditional Normal Uniform Model and the corresponding ratio of two correlated normal random variables approximation integrated over *all* the molecules. The red curve is slightly shifted, with respect to the empirical points, because the molecules in the estrogen receptor dataset have an average size of 143 bits, while molecules in ChemDB have an average size of 205 bits. Deviations from the red line are observed in the actual data. The right side of Figure 7 shows how this can be corrected by looking at individual molecules, in this case, the molecules with the smallest ($A = 64$) and largest ($A = 305$) size among the dataset of estrogen receptor ligands. The predicted curves obtained by conditioning based on the corresponding values of A are in excellent agreement with the corresponding empirical values.

8.3. Extreme-Value Distributions and p-Values. The p value for a score t is computed from the extreme-value distribution and corresponds to the probability of observing a maximum score above t. Therefore, it is given by the complement (eq 11) of the cumulative distribution of the maximum scores (eq 10). It can again be measured empirically by Monte Carlo sampling or predicted from the distribution of the scores and its extreme-value distribution, in particular, using the Weibull form of eq 16 (see also B1). Figure 5 again demonstrates examples of p-value results for actual molecule searches in ChemDB. Each search yields one binary result of whether or not the maximum score is

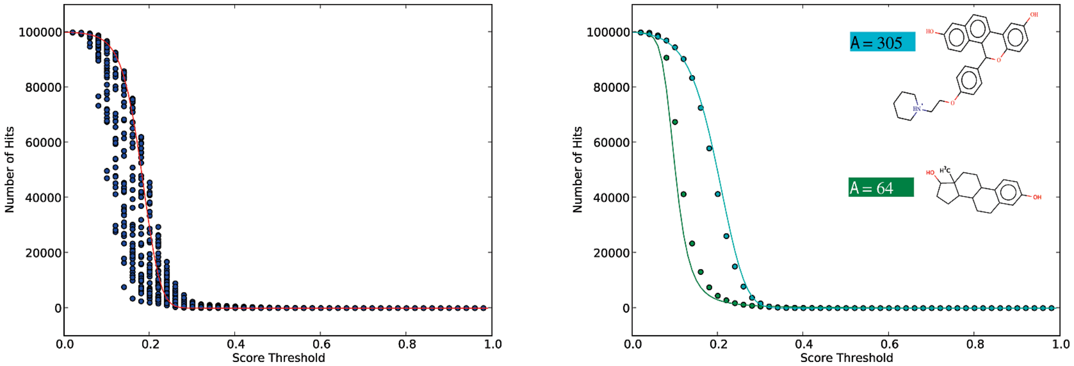


Figure 7. Left: 55 Estrogen receptor ligands are used to query a sample of 100 000 molecules randomly selected from the ChemDB. Horizontal axis represents Tanimoto threshold scores. Vertical axis represents number of scores above the threshold (hits). Each dot represents a query's number of hits above the corresponding threshold on the horizontal axis. Superimposed dots are indistinguishable (see text). The solid red line represents the predicted *E*-values based on the ratio of two correlated normal random variables approximation integrated over all values of *A* in the sample. Right: Dots associated with the estrogen receptor ligand with the largest *A* (cyan) and the smallest *A* (green) are isolated. The solid lines show predicted *E*-values based on the ratio of two correlated normal random variables conditioned based on the size of the two query molecules: $A = 305$ (cyan) and $A = 64$ (green).

greater than a threshold. Thus, multiple searches of the query molecule against different samples are needed to derive a probability that directly compares to the computed *p*-value. As in the case of *E*-values, the figure shows that the *p*-values obtained by conditioning based on *A* closely approximate the *p*-values obtained by Monte Carlo simulations. These *p*-values very clearly identify column c in the figure as the only column corresponding to a significant Tanimoto similarity, with respect to ChemDB. Unlike the *E*-value above, the *p*-value is not effective with regard to separating columns a and d in the figure. This is because the *p*-value is useful for assessing Tanimoto scores that are in the tail of high scores and does not work well on average scores. Additional results showing good agreement between predicted and empirical *p*-values, as well as additional technical details, are given in B1 (see Figures B1 and B3).

8.4. Outliers. The notion of outliers, similar to the notion of significance, is relative to a particular background distribution. For instance, we can apply the general framework to easily detect molecules that are outliers, or behave atypically, with respect to the rest of the molecules in a database such as ChemDB. This is illustrated in Figure 6 with an example focusing on molecules satisfying $A = 220$ showing the distribution of the scores for 100 such molecules. The red curve represents essentially the predicted distribution conditioned on $A = 220$. The green and blue curves identify two different molecules in this group, with very typical and very atypical behavior, as measured in terms of Kullback−Leibler divergence. The KL divergence is given by $D_{\mathrm{KL}}(P||M) = \int_0^1 P(t)\log\, P(t)/M(t)$ and can be used to measure the dissimilarity between any distribution $P(t)$ of scores and the expected distribution $M(t)$. The typical molecule has a KL divergence of 0.003 (green), whereas the atypical molecule has a KL divergence of 1.075 (blue).

8.5. ROC Curves. Figure 8 compares empirical and predicted ROC curves for six diverse data sets of molecules taken from the literature: (1) 55 estrogen receptor ligands;[38] (2) 17 neuraminidase inhibitors;[38] (3) 24 p38 MAP kinase inhibitors;[38] (4) 40 gelatinase A and general MMP ligands;[38] (5) 36 androgen receptor ligands;[39] and (6) 28 steroids with corticosteroid binding globulin (CBG) receptor affinity[40] against a random background of 100 000 molecules selected from the ChemDB and used as negative examples. The distribution of positive scores is obtained empirically by deriving all the pairwise scores in each dataset. The distribution of negative scores is obtained in each case by all the pairwise scores between molecules in the corresponding positive sets and the 100 000 molecules in the negative set. The distribution of negative scores is modeled here as a ratio of two correlated normal random variables, or as a single normal, gamma, or beta (after rescaling) distribution. In all cases, the predicted ROC curves approximate the empirical ROC curves well. Thus, in short, the framework presented here allows one to predict both true and false, as well as positive and negative, rates at all possible thresholds quite accurately and estimate retrieval measures such as specificity, sensitivity, precision, and recall at a given threshold, or estimate ROC curves over the entire set of possible score thresholds.

9. DISCUSSION AND CONCLUSIONS

This paper develops the statistical theory for modeling, predicting, approximating, and understanding the distributions of chemical similarity scores and their extreme values. The framework allows one to answer the questions raised in the Introduction and yields simple guidelines to determine the significance of chemical similarity scores by computing *Z*-scores, *E*-values, and *p*-values. To demonstrate the advantages of *Z*-scores, *E*-values, and *p*-values, consider, for example, a Tanimoto score of 0.5. The significance of this score depends on many considerations, including the size of the database, the types of molecules represented in the database, and the molecular representations used. For instance, the significance of a 0.5 score varies when it is obtained in a database containing 1024 modulo-OR compressed path fingerprints versus one containing circular substructure fingerprints with lossless compression. This makes the 0.5 Tanimoto score very specific to the particular implementation. In contrast, the *Z*-score, *E*-value, and *p*-value corresponding to a Tanimoto score of 0.5 take into account the global distribution of the scores and are more intrinsic and comparable across different implementations and experiments.

The parameters describing the score distributions can be derived from various models of fingerprints or they can be

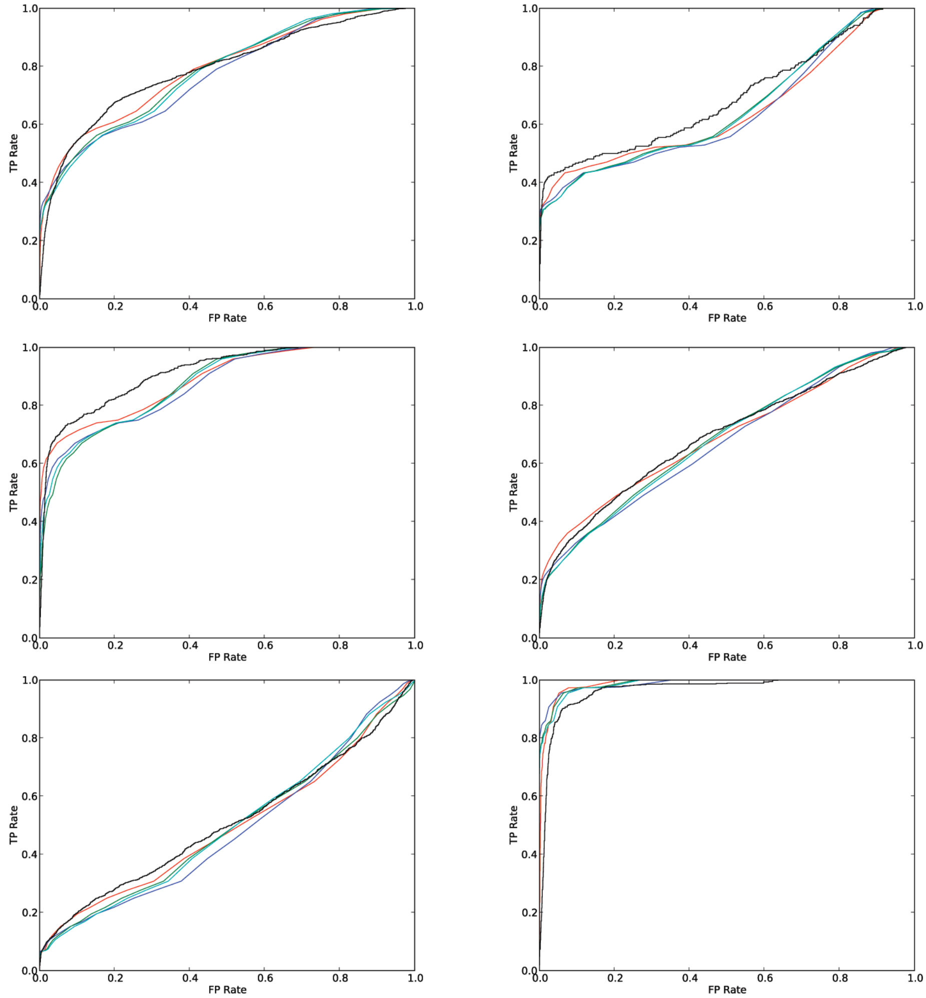


Figure 8. ROC curves for six datasets of active molecules (from left to right and top to bottom): (1) 55 estrogen receptor ligands; (2) 17 neuraminidase inhibitors; (3) 24 p38 MAP kinase inhibitors; (4) 40 gelatinase A and general MMP ligands; (5) 36 androgen receptor ligands; and (6) 28 steroids with corticosteroid binding globulin (CBG) receptor affinity. Empirical ROC curves are shown in black. Various approximations of the negative molecule scores distribution are used to get the theoretical curves, including a ratio of two correlated normal random variables distribution (red), a single normal distribution (blue), a single gamma distribution (green), and a single beta distribution (cyan), using a random sample of 100 000 molecules from the ChemDB.

learned empirically. The detailed derivation in A1 demonstrates how the models can be conditioned based on the size of the query molecule (*A*) and/or database molecule (*B*), providing multiple sets of parameters specific to those sizes. Parameters learned from empirical data can also be conditioned on molecule size, by sampling correspondingly from molecule fingerprints containing *A* and/or *B* 1-bits. Conditioning the parameters on both *A* and *B* greatly increases the number of parameters. For instance, in a typical implementation using 1024 modulo-OR compressed path fingerprints, the value of *A* and *B* could span the 1−500 range, requiring many parameters in the 500^2 range. A large number of parameters may increase the look-up time, thus adding complexity to the search. The results presented here show that conditioning on the query molecule size *A* alone offers a good tradeoff, because of the manageable parameter size (∼500) and considerably improved retrieval results, compared to no conditioning at all. Furthermore, using probabilistic models such as the Conditional Normal Uniform Model, the number of parameters can be further reduced to a very small number (∼2), although it may still be desirable to precompute and store the parameters of the score distributions at each possible value of *A*. Similarly, to condition the parameters of the Weibull extreme-value distribution on *A*, B1 demonstrates a simple approach where the parameters can be computed economically using simple polynomial functions of *A*. For several applications, it is also possible to condition the distribution of scores or extreme values based on groups of related molecules (for instance, molecules known to have the same biological function or bind to the same receptor). In this case, theoretical distributions such as the distribution of the ratio of two correlated normal random variables or the Weibull distribution for extreme values must be fitted to empirical data.[7]

This work has been in part inspired by analogies with the field of bioinformatics and the problem of searching large databases of nucleotide or amino acid sequences using standard tools such as BLAST. However, in considering future applications of the theory to chemoinformatics, it is important also to take into consideration some of the differences between the two fields, including differences in culture with respect to data sharing, openness, and standardization. In addition, BLAST was originally created to detect homology due to evolution. While natural evolution is different from the process that has led to the small molecules found in chemoinformatics databases today, we do not believe that this alone results in a fundamental difference, especially in light of the fact that increasing numbers of synthetic biological sequences are being bioengineered. Perhaps more significant is the fact that simple die toss models generally are better at modeling biological sequences than simple coin flips are at modeling small molecule fingerprints. This is due to the sequential nature of biological sequences and the nonsequential nature of molecular fingerprints. For instance, a small die toss perturbation of a biological sequence results in another sequence, whereas a coin flip perturbation of a fingerprint generally does not correspond to a valid fingerprint. However, even such a difference seems to be more quantitative than qualitative and implies that different probabilistic models may be used in different domains.

As far as other domains are concerned, it is worth noting that the general methods presented here are not limited to chemoinformatics: they could be applied to other areas of information retrieval, particularly text retrieval, which is formally very similar to chemical retrieval. Text retrieval methods often represent documents precisely by binary fingerprints, similar to molecular fingerprints, using the well-known "bag of words" approach. In this approach, each document is viewed as a bag of words, and the components of the corresponding fingerprint index the presence or absence of each word from the vocabulary in the document. In addition, the similarity between documents is often computed from the corresponding fingerprints using precisely the same Jaccard-Tanimoto similarity measure.

In the short term, however, it is likely that users of chemoinformatics search engines, as well as users of Google and other text search engines, will continue to inspect the top hits returned by a search manually and rely on the raw Jaccard-Tanimoto scores or the corresponding *Z*-scores, *E*-values, and *p*-values to assist them with their inspections. It is in high-throughput data mining applications with a large number of queries applied to one or multiple databases, possibly orders of magnitude larger than those available today, and when manual inspection becomes impossible, that the framework developed here may find its most fruitful applications.

ACKNOWLEDGMENT

Work supported by NIH Biomedical Informatics Training (Grant No. LM-07443-01) and NSF (Grant Nos. EIA-0321390 and IIS-0513376) to P.B. We would like also to acknowledge the OpenBabel Project and OpenEye Scientific Software for their free software academic license, as well as additional support from the Camille and Henry Dreyfus Foundation. We also thank Dr. R. Benz for useful discussions.

APPENDIX A. PROBABILISTIC MODELS OF FINGERPRINTS

In this appendix, we show how the probabilistic models can be treated analytically and how the means, standard deviations, and covariances of the intersections and unions can be computed.

A1. Single-Parameter Bernoulli Model. For this model, we assume that the fingerprints $\vec{B}$ in the database are generated by N coin flips with a constant probability p of producing a 1-bit. The distribution of the number of 1-bits in the database fingerprints is then given by a binomial distribution $\mathscr{B}(N, p)$, which can be approximated by a normal distribution $\mathscr{N}(Np, Npq)$ when N is large. Similarly, we can assume that the query fingerprints are produced by coin flips with a constant probability r of producing a 1-bit. The case where $r \neq p$ can be treated at no extra cost. Consider a fingerprint query $\vec{A}$, with A 1-bits with a binomial distribution $\mathscr{B}(N, r)$, which can be approximated by $\mathscr{N}(Nr, Nrs)$ ($s = 1 - r$) when N is large. Let $\vec{I} = (I_i)$ and $\vec{U} = (U_i)$ denote the intersection and union fingerprints, respectively. The intersection size $I = A \cap B = \sum_i I_i = \sum_i (A_i \cap B_i)$ then is a random variable with a binomial distribution $\mathscr{B}(N, pr)$, which can be approximated by a normal distribution $\mathscr{N}(Npr, Npr(1 - pr))$ when N is large, as well as a Poisson distribution $\mathscr{P}(Npr)$ when N is large and pr is very small. Similarly, the union size $U = A \cup B = \sum_i U_i = \sum_i (A_i \cup B_i)$ is a random variable with a binomial distribution $\mathscr{B}(N, 1 - qs) = \mathscr{B}(N, p + r - pr)$, which can be approximated by a normal distribution for N large $\mathscr{N}(N(1 - qs), N(1 - qs)qs)$, and a Poisson distribution $\mathscr{P}(N(p + r - pr))$ when N is large and $p + r - pr$ is small.

Under the binomial model, we can get an exact expression for the distribution of the Tanimoto scores. The Tanimoto score $T = I/U$ can only take rational values t between 0 and 1. Assuming that n and m are irreducible, with $0 \leq n \leq m$ and $t = n/m$, the probability $P(T = t)$ is given exactly by

$$P(T = t) = P\left(\frac{I}{U} = \frac{n}{m}\right) = \sum_{k=1}^{K} P\left(\begin{matrix} I = kn \\ U = km \end{matrix}\right) =$$
$$\sum_{k=1}^{K} \binom{N}{kn} p^{kn} q^{kn} \binom{N - kn}{km - kn} (ps + qr)^{km-kn} q^{N-km} s^{N-km} \quad \text{(A1)}$$

where K is the largest integer such that $Km \leq N$, i.e., $K = \lfloor N/m \rfloor$. Clearly, if t is not rational, this probability is 0. Thus, from this distribution, we can derive all the properties of the score distribution, including its mean and variance, under the assumptions of the binomial model.

To further simplify matters, by approximating the numerator I and denominator U by normal distributions as described above, we can view the Tanimoto score as the ratio of two correlated normal random variables. Thus, we next need to compute the covariance between I and U. Noticing that the components I_i and U_j are independent for $i \neq j$, we have

$$\mathrm{Cov}(I, U) = \sum_i \mathrm{Cov}(I_i, U_i) = N\mathrm{Cov}(I_i, U_i) \quad \text{(A2)}$$

A direct calculation gives

$$\mathrm{Cov}(I_i, U_i) = E(I_i U_i) - E(I_i)E(U_i) = pr - pr(p + r - pr) = pr(1 - p - r + pr) \quad \text{(A3)}$$

so that

$$\mathrm{Cov}(I, U) = Npr(1 - p - r + pr) \quad \text{(A4)}$$

Thus, we can approximate the distribution of the Tanimoto scores under the simple Bernoulli model by studying the ratio of two correlated normal random variables approximating the numerator I and denominator U, with means, variances, and covariances as described above.

In Figure 1, the distributions of the number of 1-bits in actual fingerprints contained in the ChemDB and fingerprints generated by the single-parameter Bernoulli model, with p

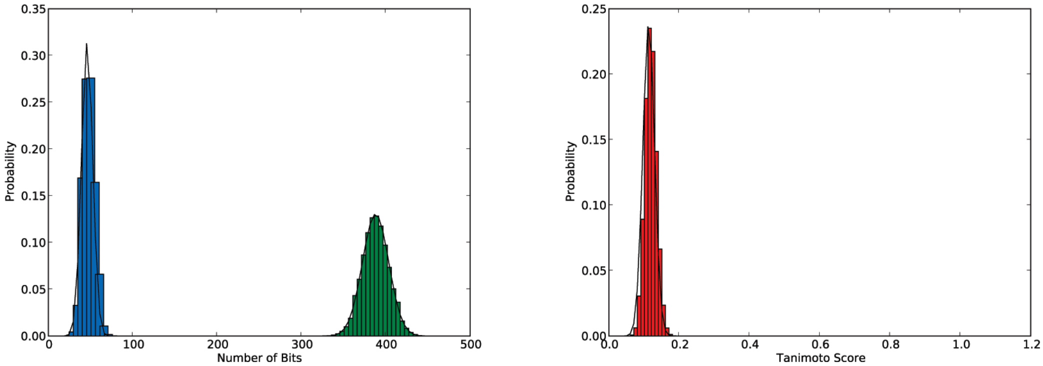


Figure A1. Results obtained using 100 query fingerprints to search 100 000 fingerprints. All fingerprints have length $N = 1024$ and are generated using a single-parameter Bernoulli model with $p = 205/1024$ to fit the average values in the actual ChemDB fingerprints. Left: histograms for the size of the intersections (blue) and the unions (green), together with their normal approximations (solid black lines). Right: histogram for the corresponding Tanimoto scores (red), together with the corresponding ratio of correlated normal random variables approximation (solid black line).

chosen to fit the average, are compared. Although both distributions have the same mean by construction, the variance of the ChemDB distribution is significantly larger. Both distributions are also well-approximated by normal distributions with $\mathcal{N}(\mu = 218, \sigma^2 = 9552)$ for the empirical distribution, and $\mathcal{N}(\mu = 218, \sigma^2 = 172)$ for the binomial model. The additional width parameter of the normal model helps to capture the diverse sizes of molecules represented in the empirical fingerprints and can be used effectively in a Conditional Normal Uniform Model.

Figure A1 shows the distributions of the intersections and unions of fingerprints generated using a single-parameter Bernoulli model with p fit to approximate the mean of the fingerprints in ChemDB ($p \approx 205/1024$) for both the queries and the database. Normal approximations to these distributions are superimposed on the two histograms. Figure A1 also shows the empirical distribution of the scores, together with the corresponding ratio of correlated normal random variables approximation.

A2. Multiple-Parameter Bernoulli Model. The analysis above for the single-parameter Bernoulli model can easily be extended to the multiple-parameter Bernoulli model by using similar expressions for the mean, variance, covariance of the individual variables I_i and U_i and combining them using the linearity of the expectation and the independence of components associated with different indices. In this case, we let $p_1, p_2, ..., p_N$ be the vector of probabilities for the database and $r_1, r_2, ..., r_N$ the vector of probabilities for the queries. The mean and variance of I are given by $\sum_i p_i r_i$ and $\sum_i p_i r_i(1 - p_i r_i)$, respectively. Thus, I can be approximated by a normal distribution $\mathcal{N}(\sum_i p_i r_i, \sum_i p_i r_i(1 - p_i r_i))$. Similarly, the mean and variance of U are given by $\sum_i(1 - q_i s_i)$ and $\sum_i(1 - q_i s_i)q_i s_i$, respectively. Thus, U can be approximated by a normal distribution $\mathcal{N}(\sum_i(1 - q_i s_i), \sum_i(1 - q_i s_i)q_i s_i)$. Finally, for the individual covariance terms, we have $\mathrm{Cov}(I_i, U_i) = p_i r_i(1 - p_i - r_i + p_i r_i)$ and $\mathrm{Cov}(I_i, U_j) = 0$ for $i \neq j$. Therefore, the full covariance is given by the sum $\mathrm{Cov}(I, U) = \sum_i p_i r_i(1 - p_i - r_i + p_i r_i)$. Thus, one can effectively proceed with the ratio of two correlated normal random variables approximation, as for the single-parameter Bernoulli model.

Despite its many parameters, the multiple-parameter Bernoulli model suffers from some of the same weaknesses as the single-parameter Bernoulli model, compared to empirical fingerprints (result not shown). In particular, when using the empirical probabilities p_i, the model can easily fit the average of A but tends to underestimate the variance of A, where A is the number of 1-bits in the empirical fingerprints.

A3. Conditional Distribution Uniform Model. With the Conditional Distribution Uniform Model, we can first fit a distribution—typically, a normal distribution—to the size A of the fingerprints in the database or the set of queries. This generally provides a better fit than what can be obtained using the Bernoulli models. Second, this model allows exact conditioning based on the size A of the queries or the size of the molecules being searched. In the single-parameter Bernoulli approach, conditioning based on A is not implemented exactly but instead is approximated using $p = A/N$. Finally, once A is fixed, the uniform portion of the model ensures exchangeability without independence of the components.

A3.1. Conditioning on A and B. In the Conditional Distribution Uniform Model, it is easy to see that, for fixed A and B, the intersection $I = A \cap B$ has a hypergeometric distribution, with probabilities given by

$$P(I = k|A, B) = \frac{\binom{A}{k}\binom{N - A}{B - k}}{\binom{N}{B}} = \frac{\binom{B}{K}\binom{N - B}{A - k}}{\binom{N}{A}} \tag{A5}$$

for $A + B - N \leq k \leq \min(A, B)$; otherwise, $P(I = k|A, B) = 0$.

To study the Tanimoto scores directly, we have the conditional density

$$P\left(T = \frac{I}{U} = t|A, B\right) = P\left(\frac{I}{A + B - I} = t|A, B\right) = P\left(I = \frac{t(A + B)}{1 + t}|A, B\right) \tag{A6}$$

and the conditional cumulative distribution

$$P(T \leq t|A, B) = P\left(\frac{I}{A + B - I} \leq t|A, B\right) = P\left(I \leq \frac{t(A + B)}{1 + t}|A, B\right) \tag{A7}$$

Therefore, the probability distribution for the similarity T can be derived from the hypergeometric distribution of I, given A, B, and N. In particular, we have the conditional distribution

$$P(T = t|A) = \sum_{B=0}^{B=N} P(t|A, B)P(B) \qquad \text{(A8)}$$

where the sum is over the distribution $P(B)$. This approach is thus consistent with the Conditional Distribution Uniform Model, which depends on the model for $P(B)$. To model this distribution, we can use the binomial model

$$P(B) = \binom{N}{B} p^{B}(1 - p)^{N-B}$$

However, it is often preferable, as previously discussed, to use a more-flexible normal model, with

$$P(B) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(B - \mu)^2}{2\sigma^2}\right] \qquad \text{(A9)}$$

where the mean and standard deviation are fitted to the empirical values. The unconditional distribution of Tanimoto scores is given by a second integration over the distribution $P(A)$ of queries

$$P(T = t) = \sum_{A=0}^{A=N} P(T = t|A)P(A) = \sum_{A=0}^{A=N} \sum_{B=0}^{B=N} P(t|A, B)P(B)P(A) \qquad \text{(A10)}$$

For the Conditional Distribution Uniform Model, we can derive the means, variances, and covariances of the intersections and unions at all levels of conditioning. First, conditioned based on A and B, the mean of this hypergeometric distribution is $E(I|A, B) = AB/N$. The variance is given by $\mathrm{Var}(I|A, B) = A(B/N)(1 - B/N)(N - A)/(N - 1)$. The union can be studied from the intersection by writing $U = A + B - I$, so that $P(U = k|A, B) = P(I = A + B - k|A, B)$. Therefore, conditioned based on A and B, the expectation of U is given by $E(U) = A + B - E(I)$, using the linearity of the expectation. Similarly, $\mathrm{Var}(U|A, B) = \mathrm{Var}(I|A, B)$, i.e., the variance of U is equal to the variance of I. In the same way, we can also estimate the covariance by writing $\mathrm{Cov}(I, U) = E(IU) - E(I)E(U)$, and writing $U = A + B - I$ yields

$$\mathrm{Cov}(I, U) = E(I(A + B - I)) - E(I)E(U) = (A + B)E(I) - E(I^2) - E(I)(A + B - E(I)) = -\mathrm{Var}(I) \qquad \text{(A11)}$$

A3.2. Conditioning on *A* Only (or *B* Only). We can now condition over A only (or B only, *mutatis mutandis*), i.e., integrating over B. For this, we assume that B has a distribution with mean μ_B and variance σ_B^2. This distribution could be normal but does not have to be so. Only the mean and variance of this distribution are used in the following calculations, and similarly for the distribution of A. Integrating over B, the mean of the intersection and the union are given by

$$E(I|A) = \int_B \frac{AB}{N} g(B)\, \mathrm{d}B = \frac{A\mu_B}{N} \qquad \text{(A12)}$$

and

$$E(U|A) = \int_B A + B - \frac{AB}{N} g(B)\, \mathrm{d}B = A + \mu_B - \frac{A\mu_B}{N} \qquad \text{(A13)}$$

To compute the corresponding variances, we write

$$\mathrm{Var}(I|A) = \int_{\vec{B}} \left(I - \frac{A\mu_B}{N}\right)^2 g(\vec{B})\, \mathrm{d}\vec{B} = \int_B \int_{|\vec{B}|=B} \left(I - \frac{AB}{N} + \frac{AB}{N} - \frac{A\mu_B}{N}\right)^2 g(\vec{B})\, \mathrm{d}\vec{B} \qquad \text{(A14)}$$

Here, by $\vec{B}$, we wish to denote all the fingerprints in the database, and $B = |\vec{B}|$ is the bit counting function. By expanding the square and integrating first over molecules satisfying $B = |(\vec{B})|$, and then over B, we get

$$\mathrm{Var}(I|A) = \int_B \frac{AB(N - A)(N - B)}{(N - 1)N^2} g(B)\, \mathrm{d}B + \int_B \left(\frac{AB}{N} - \frac{A\mu_B}{N}\right)^2 g(B)\, \mathrm{d}B \qquad \text{(A15)}$$

These integrals can easily be calculated and yield

$$\mathrm{Var}(I|A) = \frac{A(N - A)}{(N - 1)N^2}(N\mu_B - \sigma_B{}^2 - \mu_B{}^2) + \left(\frac{A^2}{N^2}\right)\sigma_B{}^2 \qquad \text{(A16)}$$

This is an example of the law of total variance or variance decomposition formula,

$$\mathrm{Var}(X|Y) = E(\mathrm{Var}(X, YZ)) + \mathrm{Var}(E(X|Y, X) \qquad \text{(A17)}$$

which will be used again in the following calculations without further mention. The same decomposition for the union yields

$$\mathrm{Var}(U|A) = \frac{A(N - A)}{(N - 1)N^2}(N\mu_B - \sigma_B{}^2 - \mu_B{}^2) + \int_B \left(\left(A + B - \frac{AB}{N}\right) - \left(A + \mu_B - \frac{A\mu_B}{N}\right)\right)^2 g(B)\, \mathrm{d}B \qquad \text{(A18)}$$

and, finally,

$$\mathrm{Var}(U|A) = \frac{A(N - A)}{(N - 1)N^2}(N\mu_B - \sigma_B{}^2 - \mu_B{}^2) + \left(1 - \frac{A}{N}\right)^2 \sigma_B{}^2 \qquad \text{(A19)}$$

Similarly, to calculate the covariance, we have

$$\begin{aligned} \mathrm{Cov}(I|A, U|A) &= \int_B \int_{|\vec{B}|=B} \left(I - \frac{A\mu_B}{N}\right) \times \left(U - A - \mu_B + \frac{A\mu_B}{N}\right) g(\vec{B})\, \mathrm{d}\vec{B} \\ &= \int_B \int_{|\vec{B}|=B} \left(I - \frac{AB}{N} + \frac{AB}{N} - \frac{A\mu_B}{N}\right) \times \left[U - \left(A + B - \frac{AB}{N}\right) + \left(A + B - \frac{AB}{N}\right) - A - \mu_B + \frac{A\mu_B}{N}\right] g(\vec{B})\, \mathrm{d}\vec{B} \end{aligned} \qquad \text{(A20)}$$

which gives

$$\mathrm{Cov}(I|A, U|A) = \int_B -\frac{AB(N-A)(N-B)}{(N-1)N^2} + \left[\frac{A}{N}\frac{(N-A)}{N}(B-\mu_B)^2\right]g(B)\,\mathrm{d}B \tag{A21}$$

After integration over B, we finally get

$$\mathrm{Cov}(I|A, U|A) = -\frac{A(N-A)}{(N-1)N^2}(N\mu_B - {\sigma_B}^2 - {\mu_B}^2) + \frac{A}{N}\left(1-\frac{A}{N}\right){\sigma_B}^2 \tag{A22}$$

A3.3. No Conditioning. If now we integrate with respect to the query molecules A, the means are given by

$$E(I) = \int_A \frac{A\mu_B}{N}g(A)\,\mathrm{d}A = \frac{\mu_A\mu_B}{N} \tag{A23}$$

and

$$E(U) = \int_A A + \mu_B - \frac{A\mu_B}{N}g(A)\,\mathrm{d}A = \mu_A + \mu_B - \frac{\mu_A\mu_B}{N} \tag{A24}$$

To compute the variances, we apply again the law of total variance to obtain

$$\mathrm{Var}(I) = \int_A \left[\frac{A(N-A)}{(N-1)N^2}(N\mu_B - {\sigma_B}^2 - {\mu_B}^2) + \frac{A^2}{N^2}{\sigma_B}^2\right]g(A)\,\mathrm{d}A + \int_A \left(\frac{A\mu_B}{N} - \frac{\mu_A\mu_B}{N}\right)^2 g(A)\,\mathrm{d}A \tag{A25}$$

which yields

$$\mathrm{Var}(I) = \frac{(N\mu_B - {\sigma_B}^2 - {\mu_B}^2)}{(N-1)N^2}(N\mu_A - {\sigma_A}^2 - {\mu_A}^2) + {\sigma_B}^2\left(\frac{{\sigma_A}^2 + {\mu_A}^2}{N^2}\right) + {\sigma_A}^2\left(\frac{\mu_B}{N}\right)^2 \tag{A26}$$

Similarly, for the union

$$\mathrm{Var}(U) = \int_A \left[\frac{A(N-A)}{(N-1)N^2}(N\mu_B - {\sigma_B}^2 - {\mu_B}^2) + \left(1-\frac{A}{N}\right)^2{\sigma_B}^2\right]g(A)\,\mathrm{d}A + \int_A \left[\left(A + \mu_B - \frac{A\mu_B}{N}\right) - \left(\mu_A + \mu_B - \frac{\mu_A\mu_B}{N}\right)\right]^2 g(A)\,\mathrm{d}A \tag{A27}$$

which yields

$$\mathrm{Var}(U) = \frac{N\mu_B - {\sigma_B}^2 - {\mu_B}^2}{(N-1)N^2}(N\mu_A - {\sigma_A}^2 - {\mu_A}^2) + \left(1 - \frac{2\mu_A}{N} + \frac{{\sigma_A}^2 + {\mu_A}^2}{N^2}\right){\sigma_B}^2 + \left(1 - \frac{\mu_B}{N}\right)^2{\sigma_A}^2 \tag{A28}$$

Finally, for the covariance, we have

$$\mathrm{Cov}(I, U) = \int_{\vec{A}} \left(I_{|A} - \frac{\mu_A\mu_B}{N}\right)\left(U_{|A} - \mu_A - \mu_B + \frac{\mu_A\mu_B}{N}\right)g(\vec{A})\,\mathrm{d}\vec{A} \tag{A29}$$

which can again be expanded as

$$\mathrm{Cov}(I, U) = \int_A \int_{|\vec{A}|=A} \left(I_A - \frac{A\mu_B}{N} + \frac{A\mu_B}{N} - \frac{\mu_A\mu_B}{N}\right) \times \left[U_A - \left(A + \mu_B - \frac{A\mu_B}{N}\right) + \left(A + \mu_B - \frac{A\mu_B}{N}\right) - \mu_A - \mu_B + \frac{\mu_A\mu_B}{N}\right]g(\vec{A})\,\mathrm{d}\vec{A} \tag{A30}$$

from which

$$\mathrm{Cov}(I, U) = \int_A \left[-\frac{A(N-A)}{(N-1)N^2}(N\mu_B - {\sigma_B}^2 - {\mu_B}^2) + \frac{A}{N}\left(1 - \frac{A}{N}\right){\sigma_B}^2\right]g(A)\,\mathrm{d}A + \int_A \left(\frac{A\mu_B}{N} - \frac{\mu_A\mu_B}{N}\right) \times \left[\left(A + \mu_B - \frac{A\mu_B}{N}\right) - \mu_A - \mu_B + \frac{\mu_A\mu_B}{N}\right]g(A)\,\mathrm{d}A \tag{A31}$$

which yields

$$\mathrm{Cov}(I, U) = -\frac{N\mu_B - {\sigma_B}^2 - {\mu_B}^2}{(N-1)N^2}(N\mu_A - {\sigma_A}^2 - {\mu_A}^2) + \left(\frac{\mu_A}{N} - \frac{{\mu_A}^2 + {\sigma_A}^2}{N^2}\right){\sigma_B}^2 + \frac{\mu_B}{N}\left(1 - \frac{\mu_B}{N}\right){\sigma_A}^2 \tag{A32}$$

When the queries come from the database itself with the same distribution, we can simplify the last set of formulas for the mean, variance, and covariance using $\mu_A = \mu_B$ and $\sigma_A = \sigma_B$.

In short, we have derived general analytical formulas for the means, variances, and covariances of the intersections and unions of fingerprints under the Conditional Distribution Uniform Model, with various degrees of conditioning. These formulas can be used directly to derive corresponding ration of correlated normal random variables approximations to the corresponding Tanimoto score distributions.

APPENDIX B. EXTREME-VALUE DISTRIBUTION

This appendix describes a complementary approach to the extreme-value distribution (EVD) of the similarity scores using a Poisson distribution. It also describes how the parameters of the Weibull approximation to the EVD (eq 16) can be fit to the data.

B1. Extreme-Value Distribution Using the Poisson Distribution. As described in the main text, the cumulative distribution of the maximum score, *max*, in a database of size D is given by

$$F_{\max}(t) = F(t)^D \tag{B1}$$

$F(t)$ is the cumulative distribution of the similarity scores which can be obtained empirically through Monte Carlo experiments, or analytically using, for instance, the ratio of correlated normal random variables approach. For a large enough Tanimoto threshold t, it is reasonable to assume that obtaining a score

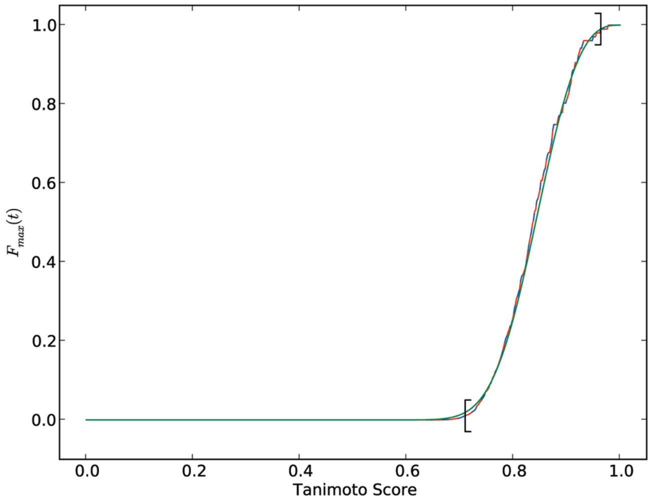


Figure B1. Plot of $F_{\max}(t)$, the cumulative distribution of the maximum score, computed on a random sample of 100 000 molecules from the ChemDB in three different ways. The solid blue curve represents the approach of eq B1. The dashed red line represents the Poisson approach of eq B4. The green solid line shows the Weibull distribution approach of eq B5. The left and right brackets on the curve indicate the acceptable boundary within which t_1 and t_2 should be selected (see eqs B7 and B8).

above t is a rare event which, thus, should follow a Poisson distribution. In other words, the probability of obtaining k similarity scores above t with a database of size D should be approximately given by the Poisson distribution

$$P_{t,D}(k) = \frac{\lambda_{t,D}^{k} e^{-\lambda_{t,D}}}{k!} \tag{B2}$$

with parameter $\lambda_{t,D}$ dependent on the threshold t and the size D of the database being searched. Note that λ could also depend on the size A of the query. λ is also the expectation of the corresponding Poisson distribution, thus $\lambda_{t,D}$ is the expected number of scores above threshold t, which is also called the E-value, and can be computed by eq 9:

$$\lambda_{t,D} = [1 - F(t)]D \tag{B3}$$

The cumulative distribution $F_{\max}(t)$ corresponds to the probability of having no scores above the threshold t and therefore is given by $P_{t,D}(0)$:

$$F_{\max}(t) = P_{t,D}(0) = e^{-[1-F(t)]D} \tag{B4}$$

Equations B1 and B4 give indistinguishable results for $F_{\max}$ (see Figure B1), which also can be expected by looking at the Taylor expansion of $e^{-[1-F(t)]}$.

B2. Fitting the Weibull Distribution Function. As previously described, the probability of *max* can be represented by a Type III EVD, or Weibull distribution function, of the form

$$F_{\max}(t) = \exp\left[-\left(\frac{\mu - t}{\sigma}\right)^{\xi}\right] \tag{B5}$$

where μ is the location parameter, σ the scale parameter, and ξ the shape parameter. The parameter μ is set to a value of 1, because the Tanimoto score (t) is bounded by 1. Parameters σ and ξ are dependent on the underlying cumulative distribution of scores ($F(t)$) and the size of the database (D). Substituting $F_{\max}(t)$ from eq B1 into eq B5 and solving for the parameters, we get the following equation, defined over $F(t) > 0$, $t \neq 1$:

$$\xi \log(1 - t) - \xi \log(\sigma) - \log[-D \log(F(t))] = 0 \tag{B6}$$

To solve for σ and ξ, we substitute two values of t ($t_1 \neq t_2$) in the equation to get the following solutions:

$$\xi = \frac{\log[-D \log(F(t_1))] - \log[-D \log(F(t_2))]}{\log(1 - t_1) - \log(1 - t_2)} \tag{B7}$$

$$\begin{aligned} \sigma &= \exp\left[\frac{\xi \log(1 - t_1) - \log[-D \log(F(t_1))]}{\xi}\right] \\ &= \exp\left[\frac{\xi \log(1 - t_2) - \log[-D \log(F(t_2))]}{\xi}\right] \end{aligned} \tag{B8}$$

Equations B7 and B8 very explicitly show how the parameters can be calculated for a specific database size (D) and Tanimoto cumulative distribution ($F(t)$). $F(t)$ can be computed either empirically, by sampling Tanimoto scores from random fingerprints in the database, or theoretically, by using the derivations presented in this work. In principle, one could use arbitrary values of t_1 and t_2 in eqs B7 and B8. However, values too close to $t_1 = 0$ and $t_1 = 1$ clearly do not work and one must be careful to select values of t_1 and t_2 corresponding to the region where $F_{\max}(t) = F(t)^D$ is not flat (see Figure B1). In practice, values of t_1 and t_2 such that $F(t_1)^D$ and $F(t_2)^D$ are in

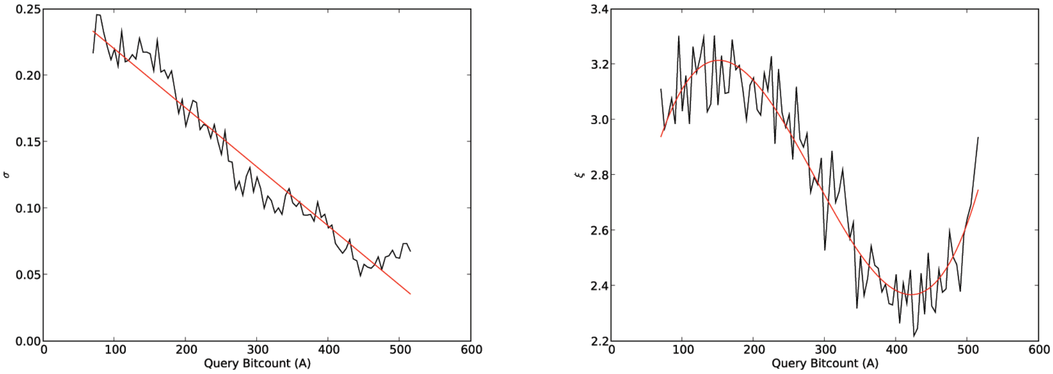


Figure B2. Polynomial fitting of the parameters σ (left) and ξ (right) of the Weibull distribution (eq B5), using a first- and third-degree polynomial, respectively (in red), as a function of the size A of the query. The empirical values (black) are obtained using a random sample of $D = 100\,000$ molecules from the ChemDB. The range of A used for fitting is [70,520]. The polynomials are $\sigma = -0.00044423A + 0.26429116$ and $\xi = 0.00000009A^3 - 0.00007387A^2 + 0.01643368A + 2.12103400$.

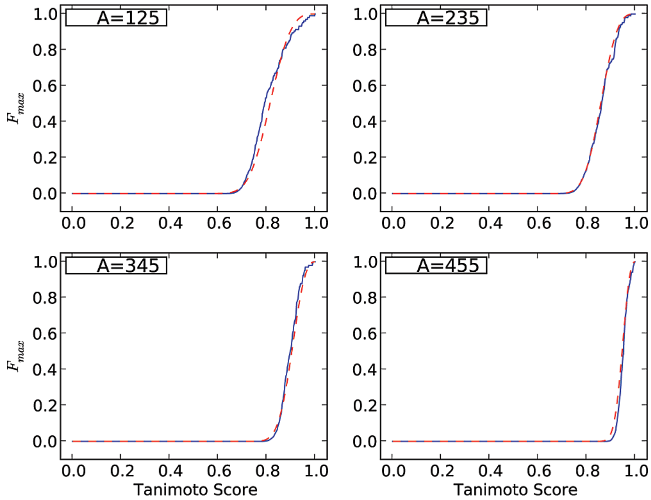


Figure B3. Cumulative EVD $F_{\max}(t)$ computed on a random sample of $D = 100\,000$ molecules from the ChemDB, conditioned on different values of A, using 100 query molecules at each value of A. The solid blue curve represents the values obtained using eq B1 applied with the empirical distribution $F(t)$ of the scores. The dashed red line shows the corresponding Weibull distribution obtained using the polynomial fit for the parameters σ and ξ, as a function of A (solid red line in Figure B2).

the interval [0.01, 0.99] give consistent results and allow for a good fit by the Weibull distribution. Ten shows the fit of $F_{\max}$ by the Weibull distribution and the corresponding [0.01, 0.99] interval. When fitting the Weibull parameters, one can condition $F(t)$ in eqs B7 and B8 based on A. This results in a family of parameters $\xi(A)$ and $\sigma(A)$ for each value of A. These values can be tabulated or one can try to fit them using a simple regression model. Figure B2 shows how the parameter can be fit with simple polynomial curves and Figure B3 shows $F_{\max}$, as well as the Weibull distribution, for different values of A. For each A, the parameters are calculated using the polynomial functions demonstrated in Figure B2.

REFERENCES AND NOTES

(1) Chen, J.; Swamidass, S. J.; Dou, Y.; Bruand, J.; Baldi, P. ChemDB: A public database of small molecules and related chemoinformatics resources. *Bioinformatics* **2005**, *21*, 4133–4139.

(2) Irwin, J. J.; Shoichet, B. K. ZINC-A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 177–182.

(3) Chen, J.; Linstead, E.; Swamidass, S. J.; Wang, D.; Baldi, P. ChemDB Update-Full Text Search and Virtual Chemical Space. *Bioinformatics* **2007**, *23*, 2348–2351.

(4) Wheeler, D.; Barrett, T.; Benson, D.; Bryant, S.; Canese, K.; Chetvernin, V.; Church, D.; DiCuccio, M.; Edgar, R.; Federhen, S.; et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2006**, *35*, D5–D12.

(5) Wang, Y.; Xiao, J.; Suzek, T.; Zhang, J.; Wang, J.; Bryant, S. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633.

(6) Altschul, S.; Madden, T.; Shaffer, A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

(7) Keiser, M.; Roth, B.; Armbruster, B.; Ernsberger, P.; Irwin, J.; Shoichet, B. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197–206.

(8) Baldi, P.; Benz, R. W. BLASTing Small Molecules—Statistics and Extreme Statistics of Chemical Similarity Scores. *Bioinformatics* **2008**, *24*, i357–i365.

(9) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer: Dordrecht, The Netherlands, 2005.

(10) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A Modification of the Jaccard/Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110–119.

(11) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.

(12) James, C. A.; Weininger, D.; Delany, J. Daylight Theory Manual. Available via the Internet at http://www.daylight.com/dayhtml/doc/theory/ (accessed January 5, 2010).

(13) Xue, L.; Godden, J. F.; Stahura, F. L.; Bajorath, J. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218–1225.

(14) Xue, L.; Stahura, F. L.; Bajorath, J. Similarity search profiling reveals effects of fingerprint scaling in virtual screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2032–2039.

(15) Baldi, P.; Benz, R. W.; Hirschberg, D.; Swamidass, S. Lossless Compression of Chemical Fingerprints Using Integer Entropy Codes Improves Storage and Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 2098–2109.

(16) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modelling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.

(17) Holliday, J. D.; Hu, C. Y.; Willett, P. Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screening* **2002**, *5*, 155–166.

(18) Swamidass, S.; Baldi, P. Bounds and Algorithms for Exact Searches of Chemical Fingerprints in Linear and Sub-Linear Time. *J. Chem. Inf. Model.* **2007**, *47*, 302–317.

(19) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.

(20) Bender, A.; Mussa, H.; Glen, R.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Model.* **2004**, *44*, 1708–1718.

(21) Hassan, M.; Brown, R. D.; Varma-O'Brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. *Mol. Diversity* **2006**, *10*, 283–299.

(22) Ackley, D. H.; Hinton, G. E.; Sejnowski, T. J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **1985**, *9*, 147–169.

(23) Frey, B. *Graphical Models for Machine Learning and Digital Communication*; MIT Press: Cambridge, MA, 1998.

(24) Marsaglia, G. Ratios of Normal Variables and Ratios of Sums of Uniform Variables. *J. Am. Stat. Assoc.* **1965**, *60*, 193–204.

(25) Hinkley, D. V. On the Ratio of Two Correlated Normal Random Variables. *Biometrika* **1969**, *56*, 635–639.

(26) Cedilnik, A.; Kosmelj, K.; Blejec, A. The Distribution of the Ratio of Jointly Normal Variables. *Metodoloski Zveki* **2004**, *1*, 99–108.

(27) Pham-Gia, T.; Turkkan, N.; Marchand, E. Density of the Ratio of Two Normal Random Variables and Applications. *Commun. Stat. Theory Methods* **2006**, *35*, 1569–1591.

(28) Tversky, A. Features of similarity. *Psychol. Rev.* **1977**, *84*, 327–352.

(29) Rouvray, D. Definition and role of similarity concepts in the chemical and physical sciences. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 580–586.

(30) Lee, R.-Y.; Holland, B. S.; Flueck, J. A. Distribution of a Ratio of Correlated Gamma Random Variables. *SIAM J. Appl. Math.* **1979**, *36*, 304–320.

(31) Iubbs, J. D.; Smith, O. E. A note on the ratio of positively correlated gamma variables. *Commun. Stat. Theory Methods* **1985**, *14*, 13–23.

(32) Loaiciga, H. A.; Leipnik, R. B. Correlated gamma variables in the analysis of microbial densities in water. *Adv. Water Resour.* **2005**, *28*, 329–335.

(33) Nagar, D. K.; M.Orozco-Castaneda, J.; Gupta, A. K. Product and Quotient of Correlated Beta Variables. *Appl. Math. Lett.* **2009**, *22*, 105–109.

(34) Galambos, J. *The Asymptotic Theory of Extreme Order Statistics*; Wiley: New York, 1978.

(35) Leadbetter, M. R.; Lindgren, G.; Rootzen, H. *Extremes and Related Properties of Random Sequences and Processes*; Springer—Verlag: New York, 1983.

(36) Coles, S. *An Introduction to Statistical Modeling of Extreme Values*; Springer—Verlag: London, 2001.

(37) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley: New York, 1991.

(38) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Funtions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.

(39) Oprea, T.; Davis, A.; Teague, S.; Leeson, P. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

(40) Coats, E. The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspect. Drug Discovery* **1998**, *12—14*, 199–213.

CI100010V