PHoS: Persistent Homology for Virtual Screening

Bryn Keller
Intel Corporation
bryn.keller@intel.com

Michael Lesnick
SUNY Albany

Ted Willke
Intel Corporation

August 14, 2018

Abstract

Finding new medicines is one of the most important tasks of pharmaceutical companies. One of the best approaches to finding a new drug starts with answering this simple question: Given a known effective drug X, what are the top 100 molecules in our database most similar to X? Thus the essence of the problem is a nearest-neighbors search, and the key question is how to define the distance between two molecules in the database. In this paper, we investigate the use of *topological*, rather than geometric, or chemical, signatures for molecules, and two notions of distance that come from comparing these topological signatures. We introduce PHoS (for ***P**ersistent **Ho**mology-based virtual **S**creening*), a new system for ligand-based screening using a topological technique known as *multi-parameter persistent homology*. We show that our approach can match or exceed a reasonable estimate of current state of the art (including well-funded commercial tools), even with relatively little domain-specific tuning. Indeed, most of the components we have built for this system are general-purpose tools for data science and will be released soon as open source software.

1 Introduction

In recent years, the search for new drugs has been moving from the wet labs to the computer. The search, when conducted *in silico*, uses a process known as *virtual screening*. Drugs generally work by attaching to certain *binding pockets* on specific proteins. The compound that attaches to the protein (i.e., the drug) is called the *ligand*. There are two main ways of approaching virtual screening: *structure-based* methods consider the crystal structure of the protein and search for ligands that will have a good fit with a known binding pocket of that particular protein, by considering the shape and chemical properties of the binding pocket, and trying various conformations of all the potential ligands in the search database. The *ligand-based* approach instead focuses on a known ligand (either an existing drug, or a naturally occurring substance such as a

neurotransmitter), and searches the database to find compounds with similar shapes and chemical properties [6].

Ligand-based screening is generally faster, and has the advantage of working even when the exact binding site is unknown, or indeed even when the protein the drug acts on is unknown. Structure-based approaches are more computationally intensive, and require more information, though they can be more conclusive as a result. In either case, only after likely candidate drugs have been found and validated in virtual screening are they ever tested in a wet lab. [63]

Shape, particularly 3D shape, turns out to be extremely important to predicting the effectiveness of a drug-protein interaction [39]. Much of the difficulty in virtual screening involves handling multiple *conformations* of compounds that have the same molecular formula, but important structural differences, such as differences in bond angles. Even small conformational differences can have up to a 100x impact on the effectiveness of a drug [39].

While there are *many* approaches to ligand-based virtual screening, several of which will be discussed in the Related Work section, our approach focuses on the application of *topological data analysis* to this problem. Topology is the branch of mathematics that deals with properties of a geometric object that are invariant under *continuous deformations* (informally, continuous deformations include bending, twisting and stretching, but not puncturing or tearing) [35, 59]. Topological data analysis (TDA) uses the techniques and tools of topology to study the shape (i.e., coarse-scale, global, geometric properties) of data [16].

One of the most popular TDA tools is *persistent homology* [71, 28]. This paper focuses on the application of an extension of this called *multi-parameter* persistent homology [17].

Specifically, we make extensive use of a new software tool for working with 2-parameter persistent homology called RIVET (the Rank Invariant Visualization and Exploration Tool) [48], codeveloped by the the first two authors. Mathematicians have been refining the tools of topological data analysis for more than a decade, and there is now a large literature on the theoretical aspects of multi-parameter persistence. However, RIVET is the first publicly available software package for working with multi-parameter persistent homology in data analysis applications.

The methodology we use to compare our approach to existing approaches requires explanation: We do not have access to the (expensive) commercial software licenses needed to run leading virtual screening tools like OpenEye's ROCS [55]. While benchmarks of the performance of such tools on publicly available data have been published, for technical reasons[1] we are not able to exactly run the same benchmarks on our tool. Thus, to compare our tool with these commercial tools, we run a slightly modified version of one benchmark on our tool, and then derive an estimate of performance of our tool on the unmodified benchmark by using the performance of an open source tool as a

[1]The tool needed to process the data is also commercial and expensive, so we used the open source RDKit [46] instead, which failed to load some of the molecules in the dataset.

reference. We explain this in more detail in the Results section.

We find that the estimated performance of our tool on the benchmark is slightly better than ROCS and others, with many potential avenues remaining for future improvements.

2 Related Work

There are many other techniques for ligand-based virtual screening. Shin et al. provide a useful survey [63], which we briefly summarize:

One dimensional approaches, such as SMILES [67] or SMARTS [1], represent the molecule as a string of characters, and attempt to judge similarity using text distance, Tanimoto correlation, or similar metrics. These methods are cheap, but not very effective, since much of ligand-protein interaction depends on the 3D shape of the ligand, which such methods fail to capture.

Two dimensional approaches, such as BCI [8], RASCAL [58], MOLPRINT2D [10], and SIMCOMP [36], represent the molecule as a graph similar to those found in any chemistry textbook. These are more useful than 1D approaches, but still fail to capture much of the important structure of the molecule. For example, two different conformations of phosphodiesterase 4D that match exactly (100%) with a 2D similarity method, have only 98% similarity in 3D, and the difference between the two means a 10x-100x difference in efficacy [39], so it can't be ignored.

Three dimensional approaches, such as USR [7], PL-PatchSurfer [38], and ROCS [37], use features like volume, atomic distances, surfaces, or fields.

- USR [7] works by calculating statistical moments from various reference points in the molecule — for example, it calculates every atom's distance from the center of mass, and then averages those distances, and calculates the variance and skewness. Similarly it measures the distance between every atom and the farthest atom from the center, and the closest, and also the atom that is farthest from the atom that is farthest from the center. The inverse of the Manhattan distance between these 12 moments in each of the two molecules is then the similarity. The weakness relative to our method is that small changes (such as one that changes which atom is farthest from the center of mass) can significantly and undesirably change the signature. Also, this is a purely structural measure, so there is no way to encode chemical or electrostatic properties in this signature. However, later variations [6, 50] do make some alterations to support additional properties.

- PL-PatchSurfer [38] takes a different approach. Rather than computing a global fingerprint for each ligand, it computes the surface of each molecule and breaks it down into local patches, so that it can compare the molecules by looking for similar patches.

- ROCS [55] uses a spherical Gaussian function centered on each heavy atom, and then tries to orient the two molecules such that they have the

best possible overlap when superimposed. Once aligned, ROCS calculates the "Shape Tanimoto" [31], which is a measure of the difference between the volume of the two shapes. More recent implementations add "atom color" to the similarity calculation, where color refers to role or type of the atom in the overall molecule. Its performance depends heavily on having a large number of conformations for each compound in the database. It also depends on finding a good alignment of the two molecules, which the algorithm does not guarantee to be optimal.

Ours is not the first application of topology to medicinal chemistry: One of the early applications of persistent homology was to protein docking [3], which continues to be an area of active exploration [69, 70, 44], though the details of the approach taken in those work are rather different than ours. More recently, there have been results in structure-based screening using standard (single-parameter) persistent homology together with machine learning [14]. In contrast, our approach achieves state of the art performance in ligand-based screening, without any need for training a machine learning model. Multi-parameter persistence allows us to both capture the important properties of the shapes of molecules, and to incorporate non-shape information such as electrostatics in a coherent and effective manner.

More recently, there have been efforts to leverage deep learning [41], and there has even been a system that combined (single-parameter) persistent homology with deep learning [14, 13] in a docking-based approach.

As far as we know, no one has published results on using multi-parameter persistent homology for virtual drug screening. However, Anthony Bak has given talks beginning in 2013 [4] and most recently in 2015 [5] on (hitherto unpublished) work he did using the persist homology software Dionysus [52] and the topological data analysis tool Mapper [64] on the problem of drug similarity. He reports that he was able to find new chemical similarities that experts in the field had previously not known about, and that these discoveries were primarily driven by analysis of 2^{nd} homology (voids or cavities). Bak's analysis uses standard 1-parameter persistent homology, not multi-parameter persistent homology, though he suggests that the latter should be well suited to this problem. He also repeatedly cited computational limitations as obstacles to better results, a sentiment that we echo. We have made progress on this issue by using cluster computing, but much future work remains in the area of computational efficiency for this technique.

Finally, multi-parameter persistent homology has been used to classify hepatic lesions in computed tomography images [2].

3 Methods

3.1 The PHoS Pipeline

Our PHoS approach to ligand-based virtual screening begins, as any ligand-based virtual screening system must, with a database of candidate ligand molecules.

Such a database may either contain one representative conformation for each ligand, or multiple conformations per ligand. We first outline our ligand screening pipeline in the case that there is just one conformation per ligand: Using 2-parameter persistent homology, we associate to each ligand several topological signatures; we work with two basic types of signatures: *fibered barcodes* and *Hilbert functions*. We consider a dissimilarity measure on fibered barcodes called the *matching distance* [], and a dissimilarity measure on Hilbert functions called the ℓ^2*-distance*. These dissimilarity measures on ligand signatures give us dissimilarity measures on the ligands themselves, in the obvious way.

Virtual screening then works in the following simple fashion: Given a dissimilarity measure d on ligands obtained as above and a query ligand M, we find and return the k nearest neighbors of M in the database, with respect to d. In what follows, we explain the topological signatures and the metrics on them: First, we introduce the requisite ideas from ordinary and two-parameter persistent homology. Then, using these ideas, we discuss the PHoS pipeline in more detail.

To extend our pipeline to the case where we have multiple conformations per ligand, a number of straightforward approaches are possible; the simplest is to take the dissimilarity between two molecules to be the minimum dissimilarity between any two conformations. In this paper, our computational experiments focus on the single confirmation case.

3.2 Betti numbers and Persistent Homology

Persistent homology provides simple, efficiently computable invariants of data called *barcodes*. By *data*, we will mean a finite set of points in Euclidean space $\mathbb{R}^n$, though persistent homology works with other data types as well. A barcode is simply a finite collection of pairs $(x, y) \in \mathbb{R} \times (\mathbb{R} \cup \{\infty\})$, with $x < y$. Here we give a brief informal explanation of persistent homology, emphasizing the basic geometric intuition. Along the way, we also explain closely related topological invariants called *Betti numbers*, which also play an important role in our work. For a fully formal account of persistent homology, we refer the reader to the literature [26, 57, 33]; we also highly recommend Matthew Wright's short and very clear video introduction[68].

To construct a barcode from data, we first construct a nested sequence of geometric objects called a *Vietoris-Rips filtration*. To explain this, we will need several simple definitions.

Neighborhood Graphs First, given a data set $X = \{x_1, \ldots, x_k\} \subset \mathbb{R}^n$, and $r \geq 0$ we define $N(X)_r$, the *r-neighborhood graph of* X, to be the undirected graph with vertices X and an edge $[x_i, x_j]$ if and only if $|x_i - x_j| \leq r$. For $r < 0$, we define $N(X)_r$ to be the empty graph, i.e., the graph with no vertices and no edges.

Simplicial Complexes A *simplicial complex* is a higher-dimensional generalization of an undirected graph, where we allow not only vertices and edges,

but also triangles, tetrahedra, and their higher-dimensional analogues (called simplices). If S and T are simplicial complexes such that $S = T$ or T is obtained from S by adding more simplices, we say that S is a *subcomplex* of T and write $S \subset T$.

Clique Complexes For G an undirected graph with vertex set V, a *k-clique* of G is a subset σ of V of size k such that any two elements of σ are connected by an edge in G. We can extend G to a simplicial complex $C(G)$, called *clique complex* of G, by adding in a k-dimensional simplex with vertices σ for each $(k+1)$-clique σ, for all $k \geq 2$. Thus, we add a triangle into the graph for each 3-clique, add a tetrahedron for each 4-clique, and so on.

Betti numbers The primary example of a property of a geometric object that is invariant under continuous deformations is the presence and number of *holes* in the object. As such, topology is largely concerned with the study of holes. In fact, topologists distinguish between holes of different *degree*. In the language of topology, 0-degree holes are connected components. Thus, to a topologist, the symbol + has a single degree-zero hole, the symbol = has two degree-0 holes, and the symbol ÷ has three degree-0 holes. A degree-1 hole is a "tunnel" or hole you can see through, like the center of a roll of paper towels. A degree 2-hole is a "hollow space," like the empty space in an inflated balloon. Objects in 3-dimensional space cannot have any other kinds of holes, but higher dimensional objects can also have i-dimensional holes, for $i \geq 3$.

A standard construction associates to any simplicial complex S a sequence of non-negative numbers

$$\beta_0(S), \beta_1(S), \beta_2(S), \ldots$$

called the *Betti numbers* of S. Informally speaking, we interpret $\beta_i(S)$ as the number of i-dimensional holes in S.

Importantly, the Betti numbers of a simplicial complex can be efficiently computed in practice via linear algebra [26].

Vietoris-Rips Filtrations A *filtration* is a collection of simplicial complexes $F = \{F_r\}_{r \in \mathbb{R}}$ such that $F_r \subset F_s$ whenever $r \leq s$. Given a data set X, taking the clique complex of each neighborhood graph of X gives us a filtration

$$VR(X) := \{C(N(X)_r)\}_{r \in \mathbb{R}}.$$

This is called the *Vietoris-Rips* filtration of X; it is arguably the most popular construction of a filtration from data in applied topology, though others are commonly considered as well. Note that if X is finite then are only finitely many different simplicial complexes in $VR(X)$.

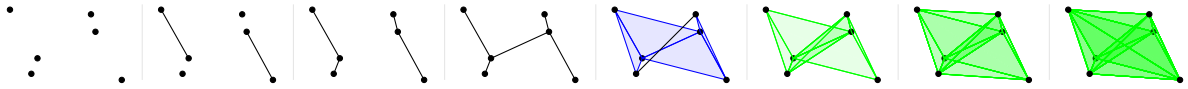

Figure 1: The Vietoris-Rips filtration of the heavy atoms of 2,3-butanediol in $\mathbb{R}^3$ (slightly simplified to save space). Blue is used for triangles and green for tetrahedra.

Figure 2: The barcodes of the filtration in Figure 1 of the heavy atoms of 2,3-butanediol. 0^{th} barcodes are shown in black. Intervals in this barcode represent connected components: each interval ends when the component it represents connects to another component. The 1^{st} barcode, which consists of a single interval, is shown in red: the interval represents the cycle that is born in step 5 and dies in step 6. This example has an empty i^{th} barcode for all $i > 1$.

Barcodes As noted above, a *barcode* is a finite collection of intervals $[x, y)$ on the real line; see Fig. 2.

Under a very mild finiteness condition on a filtration F which is always satisfied in practice, one obtains a barcode $\mathcal{B}_i(F)$ for each $i \in \{0, 1, 2, \ldots\}$, via the application of standard ideas from algebraic topology and abstract algebra [26, 22]. These barcodes are sometimes called the *persistent homology of* F. Given a data set X, we thus obtain a barcode $\mathcal{B}_i(VR(X))$ for each i. The intervals of $\mathcal{B}_0(VR(X))$ are always of the form $(0, y)$ for some $y > 0$.

Informally, the geometric interpretation of persistent homology is as follows: Each pair (x, y) in the i^{th} barcode of a filtration corresponds to an i-dimensional hole in the filtration. x is the index at which the hole forms, and y is the index at which the hole closes up. As this interpretation suggests, the barcode $\mathcal{B}_i(F)$ in fact determines the Betti number $\beta_i(F_r)$ for each $r \geq 0$: $\beta_i(F_r)$ is simply the number of intervals in $\mathcal{B}_i(F)$ which contain r. However, the barcode contains additional information about how holes in different simplicial complexes of F are related to one another.

The Vietoris-Rips filtration on the centers of the heavy atoms of the molecule 2,3-butanediol is shown in Figure 1, and its associated 0^{th} and 1^{st} barcodes are shown together in Figure 2.

Importantly, these barcodes are readily computable for data sets contain-

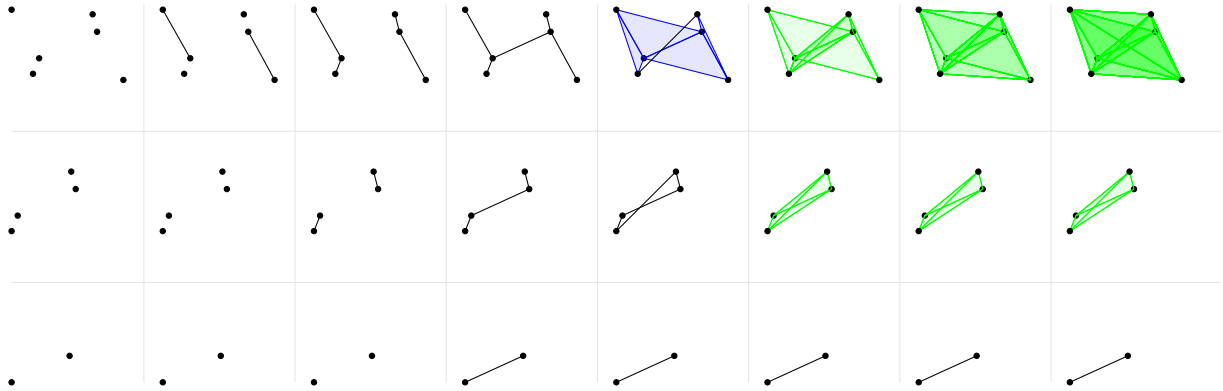

Figure 3: The Vietoris-Rips *bi*filtration of the heavy atoms of 2,3-butanediol in $\mathbb{R}^3$ (slightly simplified to save space), using partial charge as the second parameter. Blue is used for triangles and green for tetrahedra. Note the top row is identical to Figure 1, while the lower rows provide new information.

ing thousands of points (or even more points for low-dimensional data, if one uses a suitable approximation scheme [61, 25] or an alternative construction called the alpha-filtration [26, 27]). The standard algorithms use a variant of Gaussian elimination [26]. Moreover, these barcodes are known to be stable to perturbations of the data [20].

3.3 Bifiltrations and Two-Parameter Persistent Homology

In many data-analytic situations, a single filtration is not sufficient to encode the structure of interest in our data. For example, if our data points X represent atom centers in a molecule, the construction of barcodes outlined above is not sensitive to the partial charge of the atoms. This motivates the consideration of *two-parameter persistent homology*.

Two parameter persistent homology associates to a data set a *bifiltration*, a 2-parameter analogue of a filtration. A bifiltration is a collection of simplicial complexes $\{F_{r,s}\}_{(r,s)\in\mathbb{R}^2}$ such that $F_{r,s} \subset F_{r',s'}$ whenever $r \leq r'$ and $s \leq s'$.

Vietoris-Rips Bifiltration As with the one-parameter case, there are several ways one can associate a bifiltration to data. We will consider the *Vietoris-Rips Bifiltration*[17], which has the advantage of being simple and computationally tractable. Given a data set $X = \{x_1, \ldots, x_k\} \subset \mathbb{R}^n$ and any function $\gamma : X \to \mathbb{R}$, we define a bifiltration $VR(X, \gamma)$ by taking $VR(X, \gamma)_{r,s} = C(N(X^s)_r)$, where X^s is the subset of X consisting of just those points whose function value is at

most s. For instance, we may take X to be the set of atom centers of a molecule and γ to be the function specifying the partial charge of each atom. Figure 3 illustrates this bifiltration for the same molecule considered in Figure 1.

At the end of this section, we consider in more detail the problem of associating a bifiltration to a ligand.

Barcodes of Bifiltrations? Perhaps surprisingly, the natural analogue of a barcode for bifiltrations is (in general) an extremely complicated object—so complicated in fact that one cannot hope to use this object in practical data analysis applications, except in special cases [57, 17, 56]. In particular, contrary to what one might naively hope, a barcode of a bifitration generally *cannot* be defined in any reasonable way as a collection of regions in the plane.

Nevertheless, one can define simple topological signatures of bifiltrations that are useful for data analysis. We focus on two here, the *fibered barcode* and the *truncated Hilbert function*.[2]

Fibered Barcodes We define a *fibered* barcode to be a map which associates each line L in $\mathbb{R}^2$ of non-negative slope to a barcode. Given a bifiltration F, we now define an associated fibered barcode $\mathcal{B}_i(F)$ for each $i \geq 0$, as follows:

For L a line in $\mathbb{R}^2$ with non-negative slope, the restriction of F to L gives a 1-parameter filtration F^L. We define $\mathcal{B}_i(F)$, the i^{th} *fibered barcode* of F, to be the map $L \mapsto \mathcal{B}_i(F^L)$; this is well defined under very mild conditions on F.

The fibered barcodes $\mathcal{B}_i(F)$ encode important information about the topological structure of F. For example, if L is a diagonal line—say, a line of slope 1—a long interval in $\mathcal{B}_i(F^L)$ corresponds to a topological feature in F that persists over a large range of *both* bifiltration parameters.

Truncated Hilbert Functions Let $\mathbb{N}$ denote the non-negative integers. For $i \geq 0$, the i^{th} *Hilbert function* of a bifiltration F, denoted Hil^i_F, is the function $\mathbb{R}^2 \to \mathbb{N}$ given by $\text{Hil}^i_F(a) := \beta_i(F_a)$. That, is $\text{Hil}^i_F(a)$ is the i^{th} Betti number of the simplicial complex in F at index a. A visualization of some Hilbert functions of bifiltrations arising from molecular data, provided by RIVET, is shown in Figure 4; darker shading indicates a higher function value. It is easy to check that the i^{th} fibered barcode of a bifiltration determines the i^{th} Hilbert function.

We work with a variant of the Hilbert function, which takes the function to be zero outside of a certain rectangle: Given a bifiltration F, let $R_i(F)$ denote the minimal rectangle $[a, b) \times [c, d)$ containing any index where an i-dimensional hole forms or closes up. This rectangle is well defined and of finite area under

[2] Both of these signatures are are defined in terms of an algebraic object called *a persistent homology module*; this is basic object of study in multidimensional persistence[17]. To minimize the amount of formalism we need to introduce, we do not discuss persistence modules here.

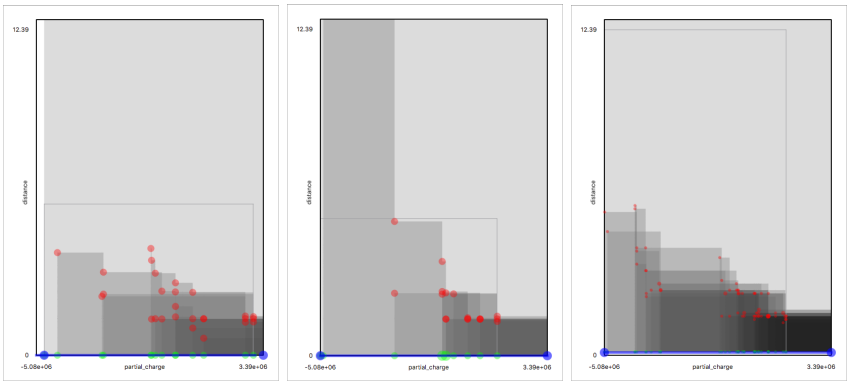

Figure 4: The 0^{th} Hilbert function on the bifiltrations of three different molecules, taking the function γ to be the partial charge function. From left to right: acetylsalicylic acid, acetaminophen, and doxorubicin.

mild assumptions on F, e.g., when F is a Vietoris-Rips bifiltration. A formal definition of $R_i(F)$ requires technical language that we have not introduced.[3]

Define $\mathrm{RH}_F^i : \mathbb{R}^2 \to \mathbb{N}$, *the* i^{th} *restricted Hilbert function of* F, by

$$\mathrm{RH}_F^i(a) := \begin{cases} \mathrm{Hil}_F^i(a) & \text{for } a \in R_i(F), \\ 0 & \text{otherwise.} \end{cases}$$

Remark 3.1. For $X \subset \mathbb{R}^n$ finite, and $\gamma : X \to \mathbb{R}$ any function, $\mathrm{RH}^i_{VR(X,\gamma)}$ is piecewise constant; indeed, it can be shown that there is a rectangular grid on $R_i(F)$, consisting of cells of the form $[a, b) \times [c, d)$, such that $\mathrm{RH}^i_{VR(X,\gamma)}$ is constant on each cell. $\mathrm{RH}^i_{VR(X,\gamma)}$ is also square integrable, i.e.,

$$\int (\mathrm{RH}^i_{VR(X,\gamma)})^2 \, dA < \infty.$$

3.4 Metrics on Topological Signatures of Bifiltrations

In both theory and applications of topological data analysis, dissimilarity measures on topological signatures play an essential role. Most often, one works with a *pseudometric*, i.e., a dissimilarity measure which is non-negative, symmetric, and satisfies the triangle inequality. There is a substantial literature on pseudometrics in the multi-parameter persistence setting, and many pseudometrics

[3] For $i \geq 0$, let F^j denote the j^{th} module in a minimal free resolution for the i^{th} homology module of F. $R_i(F)$ is the minimal rectangle containing all bigrades of elements in bases for F^0 and F^1; see, e.g., the work of Lesnick and Wright[48].

have been proposed; see for example [47, 60, 45, 12]. Here, we consider one pseudometic on fibered barcodes, the *matching distance* [45], and one pseudometric on Hilbert functions, the ℓ^2*-distance*. As mentioned above, these pseudometrics each induce pseudometrics on ligands that we will use in our PHoS ligand screening pipeline.

Matching Distance on Fibered Barcodes The matching distance on fibered barcodes is a multiparameter extension of a popular metric on barcodes called the *bottleneck distance*. Roughly, the bottleneck distance $d_b(\mathcal{B}, \mathcal{C})$ between two barcodes $\mathcal{B}$ and $\mathcal{C}$ is the magnitude of a perturbation of $\mathcal{B}$ required to transform $\mathcal{B}$ into $\mathcal{C}$ (or vice versa); here, *magnitude* is defined as the maximum distance an endpoint of any interval moves in the perturbation.[21]. This distance has very good theoretical properties, and plays a important role in the theoretical foundations of topological data analysis [20, 18, 9, 62, 30]. The bottleneck distance is also readily computed [42], and together with its variant the Wasserstein distance, is commonly used in applications[29, 2].

The matching distance d_M between fibered barcodes $\mathcal{B}$ and $\mathcal{C}$ is defined in terms of d_b by taking

$$d_M(\mathcal{B}, \mathcal{C}) := \max_L w_L \, d_b(\mathcal{B}(L), \mathcal{C}(L)),$$

where L ranges over all affine lines of positive slope, and w_L is a weight depending only on the slope of L. The weights w_L are chosen to ensure that d_M satisfy a natural stability property [45].

ℓ^2-Distance on Restricted Hilbert Functions For $f, g : \mathbb{R}^2 \to \mathbb{N}$ any square-integrable functions, the ℓ^2*-distance* between f and g is given by

$$d^2(f, g) := \sqrt{\int (f - g)^2 \, dA}.$$

By Remark 3.1, when f and g are restricted Hilbert functions of Vietoris-Rips bifiltrations, $d_R^2(f, g)$ is well defined and finite.

3.5 Computation of Invariants and Metrics of Two-Parameter Persistence

The RIVET software allows for computationally efficient handling of both fibered barcodes and restricted Hilbert functions. RIVET precomputes a data structure on which efficient queries of a fibered barcode $\mathcal{B}$ can be performed: Given a line L in $\mathbb{R}^2$ of non-negative slope, this data structure returns $\mathcal{B}(L)$. For details about this, including statements about computational complexity, see the RIVET paper [48]. A matrix reduction algorithm can be used to compute the restricted Hilbert functions [49]. For the bifiltrations we consider in our virtual screening applications, these computations are very fast in practice (see Figure 11).

The stability of persistent homology implies that the matching distance $d_M(\mathcal{B}, \mathcal{C})$ between fibered barcodes $\mathcal{B}$ and $\mathcal{C}$ can be computed approximately, up to arbitrary accuracy, by computing the bottleneck distance $d_b(M^L, N^L)$ for a finite number of lines L [11]. The computational complexity of the best known algorithm for computing the bottleneck distance is $O(n^{1.5} \log n)$ where n is the number of intervals [42] in the two barcodes. Thus, if we approximate a matching distance by computing S bottleneck distances, the runtime of the computation is $O(S(n^{1.5} \log n))$, where n is the maximum number of intervals of any barcode encountered in the computation. In our setting, we take S to be at least 225 to get sufficiently good approximations to the matching distance, and n is typically on the order of $20 - 40$. We find that these matching distance computations are practically feasible in our setting, but rather costly. In contrast, the time cost of computing the ℓ^2-distance between two restricted Hilbert functions, is linear in the number of cells in the grid over which the Hilbert function is piecewise constant (see Remark 3.1), and considerably faster than the matching distance computation in practice.

3.6 The PHoS Pipeline Revisited

As we have explained above, to define a topological signature of a ligand conformation using two parameter persistence, we first associate to the ligand a bifiltration, and then take a topological signature of the bifiltration. Building on the discussion of two-parameter homology above, we now provide some additional details about these steps.

As we have already explained, given a data set $X \subset \mathbb{R}^3$ and a function $\gamma : X \to \mathbb{R}$, we can construct the Vietoris-Rips bifiltration $VR(X, \gamma)$. Given a ligand conformation M, it is natural to choose X to the be set of centers of atoms of L. A number of reasonable choices are available for γ. For example, as mentioned above, we can take γ to be a partial charge function. Alternatively we can take γ to measure mass, or hydrogen donor/acceptor status, among others.[4] Fig. 5 shows a Vietoris-Rips bifiltration on the atom centers of the same 2,3-butanediol molecule considered in Fig. 3, but this time taking γ to measure mass rather than partial charge.

In our experiments we work with the i^{th} fibered barcodes and restricted Hilbert functions for $i = 0, 1, 2$; this captures information about clusters, tunnels, and voids formed by the atoms of a ligand. When comparing two signatures, regardless of whether we use fibered barcodes or restricted Hilbert functions, we consider the distance between two signatures to be the sum of the i^{th} distances for $i = 0, 1, 2$.

[4]Note that it is possible, and quite reasonable in fact, to take γ to be the *negative* of any of these functions; the Vietoris Rips filtration $VR(X, -\gamma)$ carries different information about the pair (X, γ).

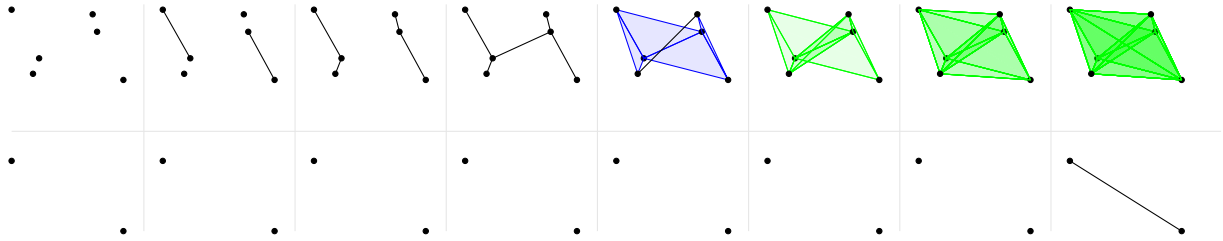

Figure 5: The Vietoris-Rips filtration of the heavy atoms of 2,3-butanediol in $\mathbb{R}^3$ (slightly simplified to save space), using atomic mass for the second parameter. Compare with Figure 3, which is based on partial charge instead of mass. Blue is used for triangles and green for tetrahedra.

3.7 Example

We illustrate the topological approach to ligand screening, by considering its behavior on 4 substances: acetylsalicylic acid, acetaminophen, sucrose, and doxorubicin. Since acetylsalicylic acid and acetaminophen work on the same protein target, we would hope for them to be quite similar in our system. Doxorubicin and sucrose are more complex molecules that work with different receptors, so we would expect them to be quite different from the first two, and perhaps more similar to each other, though not nearly so similar as acetylsalicylic acid and acetaminophen are. In Figure 6, we see qualitative similarities as we expect in the Hilbert functions of these molecules, and in Figure 7 we see the actual calculated distances are also in line with these expectations. It is interesting to note that in this work we only consider the sum of the distances for connected components, holes, and voids, but it is possible that a more sophisticated (perhaps discovered via machine learning?) scheme for combining theses distances could be even more effective.

3.8 Implementation

Our implementation must process a large number of molecules, and relate them using one of our chosen distances measures. For each molecule, we calculate and store topological signatures for Vietoris-Rips bifiltrations that capture connected components, holes, and voids. The computations of the bifiltrations and their signatures are somewhat expensive, but the distance computations are the bottleneck: While our ℓ^2 distance computations are considerably faster than our approximate matching distance computations, both approaches are costly, especially compared to lightweight virtual screening approaches like USR [7], which can analyze a dataset of the size we used in a few seconds.

Thus, to make the problem computationally feasible, we distribute the work using a cluster of 50 Intel® Xeon® E5-2699 v4 servers at 2.20 GHz (88 cores

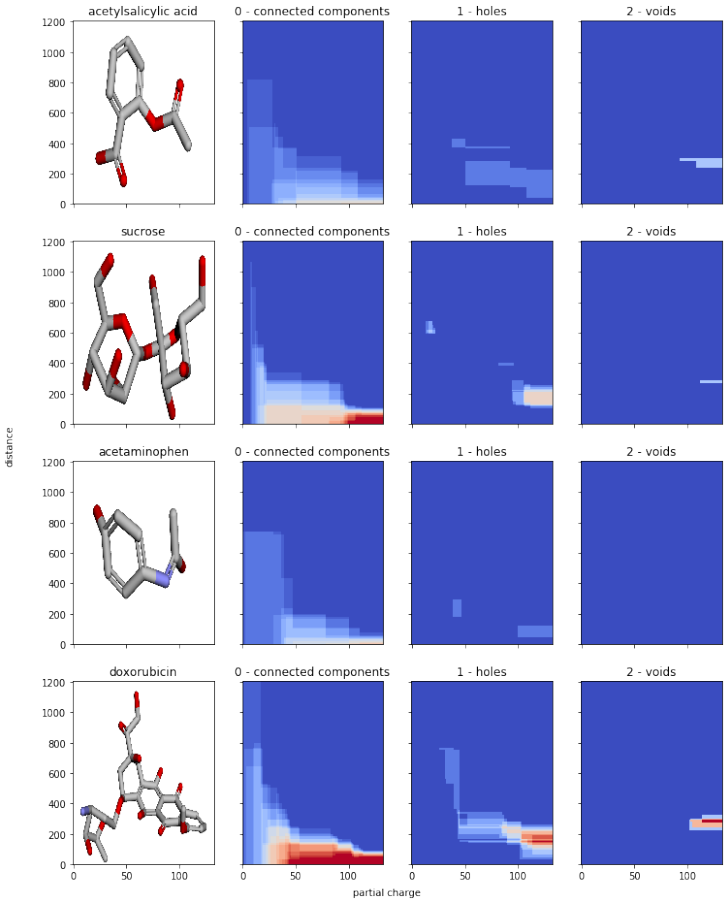

Figure 6: Four example substances, along with their i^{th} Hilbert functions for $i = 0, 1, 2$ (connected components, holes, and voids). The axis units are discrete value steps rather than physical values. The molecules are shown with the partial charge indicated, which ranges from red for low partial charge to blue for higher partial charges. The colors of the Hilbert functions range from blue for lower values, red for higher values, and the actual color values are comparable only within each column. The red areas in the connected components column shows that sucrose and (especially) doxorubicin have more atoms than the other substances. All of these molecules have some holes that result from aromatic rings, and the more complex the molecule, the more likely it is to have additional temporary holes and voids as well.

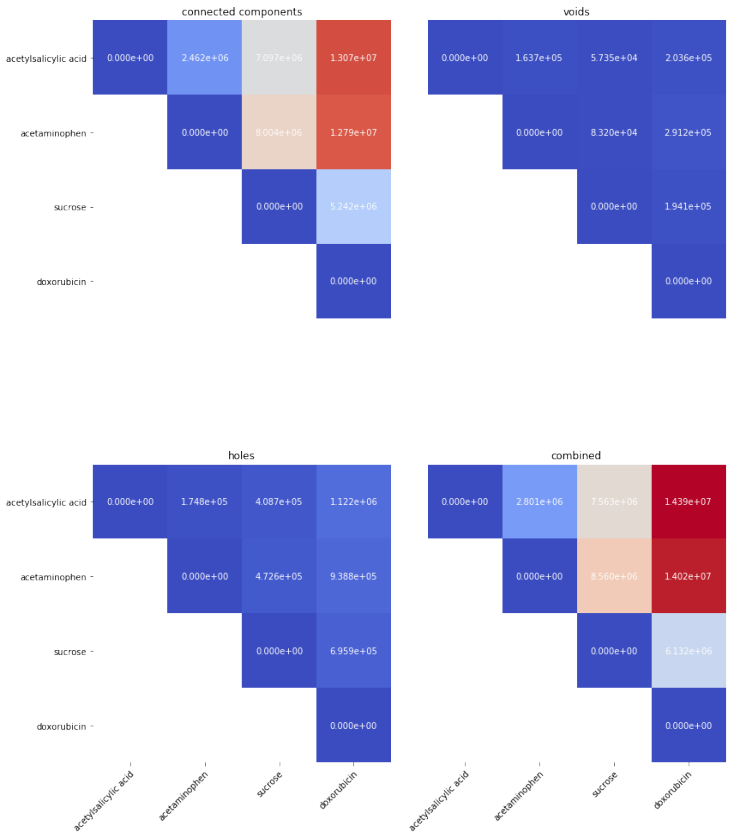

Figure 7: Distance matrices for the four example substances, showing the distance for each of the three i^{th} Hilbert functions, as well as the summed total distance. As we would hope, acetylsalicylic acid and acetaminophen are quite a bit closer than the other substances. We also see that the total distance is largely determined by the 0^{th} Hilbert function, with higher numbers making relatively smaller contributions.

per server), with 128 GB of memory. In the present work we generally used between 8 and 16 machines for each run. At this point we have not attempted to optimize any of the tools we use for these calculations, so we do not present performance data, leaving that for future work. The present implementation uses Python [66], including SciPy [40], Pandas [51], and HDF5 [34] for the pre-processing stages that generate the bifiltrations for each molecule, and the metric database is written in Rust [24]. Both parts use MPI [32, 23, 65] for cluster communication. As part of this work, C++, C, and Python APIs for RIVET [48] (available as part of the RIVET distribution), and C and Python APIs were created for Hera [43], which we used to compute the bottleneck distance.

4 Results

4.1 Datasets

Our experiments were conducted with the Cleves and Jain [19] dataset, as well as with a large (approximately 1.5 million substances) subset of the DUD-E [53] database, which we further resampled into small (approximately 1 thousand substances) samples, each with a single target (i.e., each dataset contains only ligands that bind to the same single protein binding pocket), and stratified so that the ratio of actives (true or potential drugs) to decoys (non-drugs with similar chemical properties to drugs) was the same for that target as in the larger sample. In the Cleves and Jain dataset, each target has only a very few actives, and all the decoys are used regardless of query target. In DUD-E, each target has its own decoys.

4.2 Evaluation

We tested both USR [7] and PHoS in the manner described by Shin et al. [63], using the Enrichment Factor (EF). As in their case, we define $EF_\alpha\%$ as follows:

$$EF_\alpha = \frac{N_{actives,\alpha\%}/N_{database,\alpha\%}}{N_{actives}/N_{database}}, \tag{1}$$

where $N_{actives}$ is the total number of actives in the entire database, $N_{database}$ is the total number of all compounds in the database including decoys, and $\alpha\% \in \mathbb{R}$ is the percentage of the query results we wish to inspect. Thus, $N_{actives,\alpha\%}$ is the number of actives in the top α percent of the results, and $N_{database,\alpha\%}$ is α percent of $N_{database}$.

4.3 Parameter Search

Since we use 2-parameter persistence, with the first parameter chosen to range over the inter-atom Euclidean distance, we must choose a function γ (see Bifiltrations and Two-Parameter Persistent Homology) over the atoms, which offers

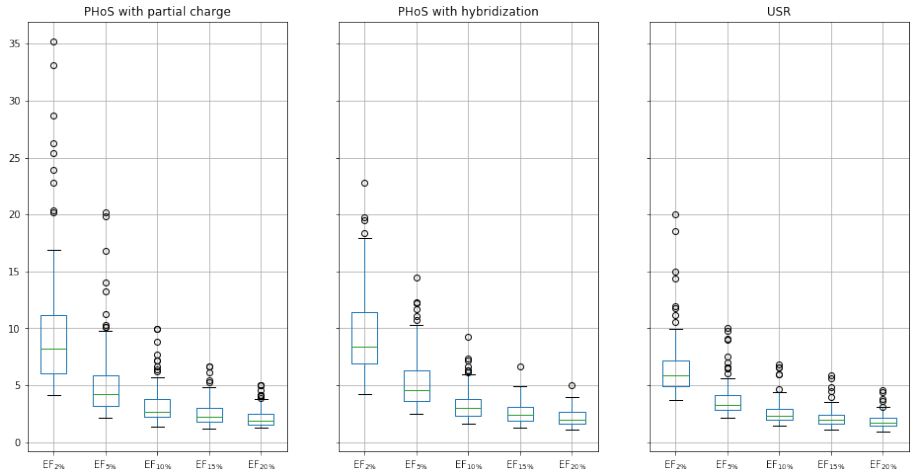

Figure 8: Performance on DUD-E vs. USR, 1 conformation, using matching distance. Each column is a box and whisker plot representing the given EF_x averaged across all queries, with the circles representing outliers. We can see that PHoS with partial charge and PHoS with hybridization both greatly outperform USR. The two PHoS methods perform similarly, with a slight edge to partial charge due to a few high-performing outliers.

the greatest insight into the structure of the molecule. There are also a number of hyperparameters to consider, such as the granularity of the bifiltrations that we calculate.

We first ran the pipeline interactively several times to determine reasonable starting values for the parameters of bifiltration resolution and slice resolution, settling on values of 11 and 15, respectively. Then to determine which grouping parameter would offer the best performance, we ran each of 15 different possible choices of γ on one 1000-element subsample of DUD-E per target (that is, 15 functions $\times$ 90 targets $=$ 1350 runs on our compute cluster). Interestingly, we found that there was no one winner; different grouping parameters worked better on different targets, though there were some that were clearly more important than others. For the best performing grouping parameters, we repeated the experiment with bifiltration resolution of 100, and slice resolution of 30. Values higher than these either led to computational difficulties or did not appreciably improve the results.

4.4 Performance with Best Parameters

Next, we tested the best parameters obtained against the DUD-E dataset versus USR, so as to help place our method in that context as well. Unfortunately, to compare with other important methods such as ROCS [55] or PL-

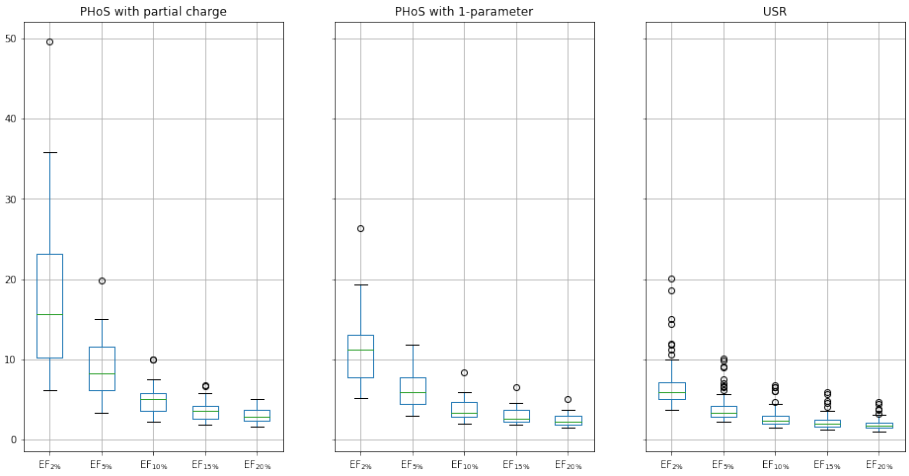

Figure 9: Performance on Cleves and Jain dataset with 1-parameter persistence, 1 conformation, using matching distance. Each column is a box and whisker plot, representing the given EF_x averaged across all queries, with the circles representing outliers. In this plot, we see that 2-parameter persistence with partial charge outperforms both traditional 1-parameter persistence and USR on this dataset.

PatchSurfer[38], we would need access to expensive licenses for ROCS and OMEGA [54] that our lab does not have. We refer the interested reader to the work of Shin et al. [63] for a comparison between USR and these other methods. Nevertheless, the comparison with USR on the Cleves and Jain dataset in Table 1 should serve to provide a first approximation of where PHoS might place in a broader comparison of ligand based virtual screening tools.

For reference, USR [7] (using an open source Python implementation [15]) was also run against the Cleves and Jain dataset. Summary results for both PHoS and USR are displayed in Table 1, and more detailed information for PHoS with partial charge and USR is included in the Supplemental Information.

We note that since we used RDKit [46] to load the molecules, there were differences in the number of decoys that could be loaded (65 failed in our dataset, versus 8 in Shin et al.'s case), which we expect explains the difference between USR's performance in their study (mean $EF_{2\%}$ of 8.8 across all targets) and in ours (mean $EF_{2\%}$ of 10.16).

Assuming the difference in USR's scores (a factor of $10.16/8.8 \approx 1.155$) translates to the difference in the score that PHoS would get on Shin et al.'s dataset, we would expect PHoS to achieve a score of $18.598/1.155 \approx 16.10$.

Based on the estimate of our systems's performance (16.10) versus the measured performance of ROCS (15.9) in the Shin et al. study, we expect that if we did have access to the ROCS and OMEGA software, we would find that this

Table 1: Mean performance on Cleves-Jain dataset, single conformation: $EF_{2\%}$ for PHoS (with various parameters) vs. USR. The best overall mean was PHoS with partial charge, 18.598. USR overall mean was 10.159. The best result in each target is in bold. Partial charge and atomic number are self-explanatory. Total degree is the number of bonds each atom has, aromatic is a flag that indicates whether an atom is in an aromatic ring or not, hybrid represents the hybridization type of each atom (S, SP, etc.), and 1-param represents single-parameter persistence calculated on the Euclidean distance between atoms.

Target	partial charge	atomic no.	total degree	aromatic	hybrid	1-param	USR
a	**19.064**	15.374	14.144	14.144	15.989	14.144	10.402
b	**21.138**	14.553	11.545	11.301	16.504	11.951	5.339
c	**23.867**	16.952	23.198	8.253	22.306	7.807	9.544
d	**32.593**	12.914	15.989	17.834	16.604	11.684	16.521
e	15.21	**20.281**	19.267	15.21	13.182	11.154	11.098
f	**35.84**	27.569	19.299	13.785	33.083	19.299	16.458
g	**32.449**	12.168	17.239	17.239	16.224	11.154	9.08
h	**49.688**	14.196	27.379	18.253	28.393	12.168	11.098
i	13.182	11.154	**15.21**	13.182	14.196	11.154	10.089
j	6.79	**10.097**	6.964	**10.097**	8.182	5.919	7.622
k	6.093	6.441	**11.49**	6.615	6.093	6.441	7.622
l	13.33	13.034	11.849	**23.698**	9.183	7.702	6.484
m	15.278	**21.181**	14.583	20.486	**21.181**	15.625	7.946
n	7.639	6.944	7.292	**10.417**	9.028	7.639	9.674
o	12.204	11.356	11.45	11.921	**14.136**	8.01	9.519
p	**7.902**	5.569	4.139	4.064	5.72	5.644	4.643
q	**27.108**	19.692	26.085	20.459	14.066	13.043	16.031
r	7.655	23.162	11.188	12.759	**27.873**	12.759	16.797
s	9.484	10.573	**14.149**	10.106	9.173	5.131	6.034
t	17.239	14.196	**29.407**	18.253	26.365	26.365	14.125
u	19.444	**25.0**	20.139	19.792	19.792	13.889	7.946
v	15.96	**23.94**	17.456	17.456	18.454	10.474	9.429

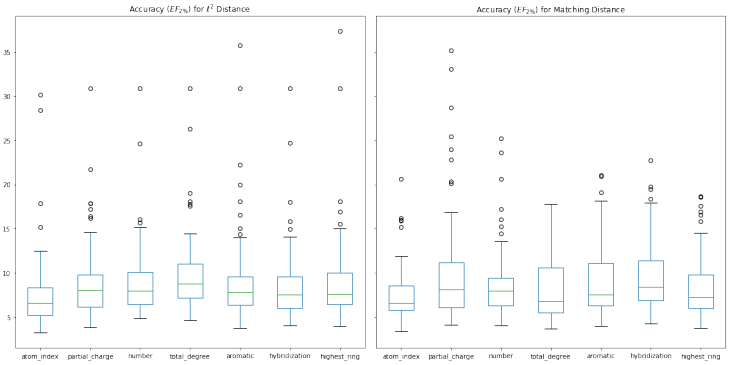

Figure 10: Comparison of $EF_{2\%}$ results on DUD-E dataset, using ℓ^2 distance and matching distance (with slice resolution 30), using box and whisker plots, with circles representing outliers. Both methods used filtration resolution of 100. The best mean value across all targets was achieved with matching distance on partial charge, with a mean single conformation $EF_{2\%}$ of 9.92. The best mean result for the ℓ^2 distance was an $EF_{2\%}$ of 9.66 using total degree, which counts the number of bonds attached to each atom.

initial version of PHoS has mean performance approximately equal to that of ROCS, with many open options for improving performance in future studies.

Finally, we compared the two distance pseudometrics used in PHoS. In Figure 10, we compare the difference in accuracy, and in Figure 11 we examine the relative times for comparisons using the two methods. Neither one has been extensively optimized, but it is clear that while the matching distance obtains slightly better accuracy, the ℓ^2 distance will be the likely choice as we scale to larger datasets in the future.

4.5 Discussion

It is important to note that while the results of Shin et al.'s study (and the original USR paper [7]) present summary means of results across all targets, we found the results of both PHoS and USR varied substantially across targets in both the DUD-E and Cleves & Jain datasets. We believe this variation deserves greater study, regardless of the computational methods used. What is it, exactly, that makes one method work well with one target, and poorly with another? PHoS has greater variance and range than USR, though the bottoms of both ranges are typically comparable, so there are many queries for which PHoS does much better than the mean, and in fact PHoS has more (and higher) high-performing outliers across targets than USR does. If we can better understand what makes PHoS perform so well on these high-performing outliers, we may be able to leverage that understanding to further improve PHoS.

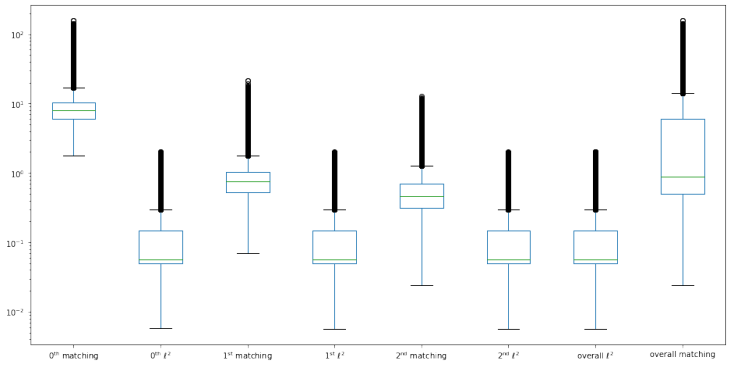

Figure 11: Comparison of the query runtimes (note the log scale) against DUD-E dataset, using matching distance (with slice resolution 30) and ℓ^2 distance, using box and whisker plots, with circles representing outliers. Both methods used filtration resolution of 100. While 0^{th} (connected components) comparisons dominate the overall expense for the matching distance, there are many outliers in all columns. Meanwhile the ℓ^2 distance runtimes are far lower, and show little variation across columns.

Additionally, we noted that in both the Cleves & Jain dataset and the DUD-E dataset, different choices of the second parameter (e.g. partial charge, mass, etc.) led to considerable difference in the score for PHoS on different targets. Naturally, we wondered if there could be some way to combine more than one of these parameters, which could be more effective than any one of them alone. This will be an area of future work.

Note we did not study the performance on a database with more than one conformer per compound. At this stage, the results would have been uninteresting since they would only have consisted of testing each conformation for one molecule against all conformations of another, and choosing the closest. We believe there are more interesting possiblities for multi-conformation comparison that do not amount to treating each conformation as a separate compound, but will have to defer that point to future work.

5 Conclusions

We have introduced PHoS, a new tool for ligand-based virtual screening based on multi-parameter persistent homological signatures of molecules, with drug similarity determined by distances in metric spaces of topological signatures. We believe the experimental evidence strongly shows that this new application of multiparameter persistent homology has great potential for helping researchers and pharmaceutical companies find new drugs faster.

In the future, we hope to continue this work by improving the computational efficiency of the tool to make it practical for users with databases with between 10-100 million compounds. The current work serves as a baseline implementation against which future algorithmic improvements can be judged. We also plan to further analyze the wide range of performance on different protein binding pocket targets (between approximately 3x and 36x more effective than chance, depending on the target). Finally, we will continue to improve the accuracy of the method by combining information from multiple types of topological signatures or taking better advantage of multiple conformations.

6 Acknowledgmements

The authors thank the Intel Corporation and the Princeton Neuroscience Institute for forming the Mind's Eye collaboration, where the authors met, and whose efforts to use topological data analysis for functional MRI analysis in neuroscience somehow eventually led to the present work. We also thank the RIVET developers, particularly Matthew Wright for creating the initial implementation of RIVET, and Simon Segert for improvements to the user interface that improved the presentation of this work. Thanks also to the Hera developers for their efficient implementation of the bottleneck distance, and particularly to Arnur Nigmetov for speedy resolution of reported issues.

References

[1] SMARTS - A Language for Describing Molecular Patterns.

[2] Adcock, A., Rubin, D., and Carlsson, G. Classification of hepatic lesions using the matching metric. *Computer Vision and Image Understanding 121* (2014), 36–42.

[3] Agarwal, P. K., Edelsbrunner, H., Harer, J., and Wang, Y. Extreme Elevation on a 2-Manifold. *Discrete & Computational Geometry 36* (2006), 553–572.

[4] Bak, A. Spaces of Shapes: Creating Moduli Spaces of Chemical Compounds for Drug Discovery, 2013.

[5] Bak, A. Feature Generation for Drug Discovery Learning- Using Persistent Homology to Create Moduli Spaces of Chemical Compounds, 2015.

[6] Ballester, P. J., Finn, P. W., and Richards, W. G. Ultrafast shape recognition: Evaluating a new ligand-based virtual screening technology. *Journal of Molecular Graphics and Modelling 27*, 7 (2009), 836–845.

[7] Ballester, P. J., and Richards, W. G. Ultrafast shape recognition for similarity search in molecular databases. *Proceedings of the Royal Soci-*

ety A: Mathematical, Physical and Engineering Sciences 463, 2081 (2007), 1307–1321.

[8] Barnard, J. M., and Downs, G. M. Chemical Fragment Generation and Clustering Software. *Journal of Chemical Information and Computer Sciences 37*, 1 (1997), 141–142.

[9] Bauer, U., and Lesnick, M. Induced Matchings and the Algebraic Stability of Persistence Barcodes. 162–191.

[10] Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *Journal of Chemical Information and Computer Sciences 44*, 5 (2004), 1708–1718.

[11] Biasotti, S., Cerri, A., Frosini, P., and Giorgi, D. A new algorithm for computing the 2-dimensional matching distance between size functions. *Pattern Recognition Letters 32*, 14 (2011), 1735–1746.

[12] Blumberg, A. J., and Lesnick, M. Universality of the homotopy interleaving distance. *arXiv preprint arXiv:1705.01690* (2017).

[13] Cang, Z., Mu, L., and Wei, G. W. *Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening*, vol. 14. 2018.

[14] Cang, Z., and Wei, G.-W. TopologyNet: Topology based deep convolutional neural networks for biomolecular property predictions. 1–27.

[15] Canno, E. Python implementation of Pedro Ballester USR descriptor (J Comput Chem 28:1711-23), 2010.

[16] Carlsson, G. *Topology and data*, vol. 46. 2009.

[17] Carlsson, G., and Zomorodian, A. The theory of multidimensional persistence. *Discrete and Computational Geometry 42*, 1 (2009), 71–93.

[18] Chazal, F., de Silva, V., Glisse, M., and Oudot, S. *The structure and stability of persistence modules*. 2012.

[19] Cleves, A. E., and Jain, A. N. Robust ligand-based modeling of the biological targets of known drugs. *Journal of Medicinal Chemistry 49*, 10 (2006), 2921–2938.

[20] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. Stability of Persistence Diagrams. *Discrete and Computational Geometry 37* (2006), 103–120.

[21] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. Stability of persistence diagrams. *Discrete and Computational Geometry 37*, 1 (Jan. 2007), 103–120.

[22] Crawley-Boevey, W. Decomposition of pointwise finite-dimensional persistence modules. *Journal of Algebra and Its Applications 14*, 05 (2015), 1550066.

[23] Dalcin, L. D., Paz, R. R., Kler, P. A., and Cosimo, A. Parallel distributed computing using Python. *Advances in Water Resources 34*, 9 (2011), 1124–1139.

[24] Developers, T. R. P. The Rust Programming Language, 2011.

[25] Dey, T. K., Shi, D., and Wang, Y. SimBa: An Efficient Tool for Approximating Rips-filtration Persistence via Simplicial Batch-collapse. In *European Symposium on Algorithms* (2016), no. 206, pp. 1–16.

[26] Edelsbrunner, H., and Harer, J. Computational topology: An introduction. *Effective Computational Geometry for Curves and Surfaces* (2006), 277–312.

[27] Edelsbrunner, H., Kirkpatrick, D., and Seidel, R. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory 29*, 4 (jul 1983), 551–559.

[28] Edelsbrunner, H., Letscher, D., and Zomorodian, A. Topological persistence and simplification. *Discrete and Computational Geometry 28*, 4 (2002), 511–533.

[29] Edelsbrunner, H., and Morozov, D. Persistent homology: theory and practice. *6th European Congress of Mathematics* (2012), 123–142.

[30] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. Confidence sets for persistence diagrams. *Annals of Statistics 42*, 6 (2014), 2301–2339.

[31] Fontaine, F., Bolton, E., Borodina, Y., and Bryant, S. H. Fast 3D shape screening of large chemical databases through alignment-recycling. *Chemistry Central Journal 1*, 12 (2007), Published online.

[32] Forum, M. P. MPI: A Message-Passing Interface Standard. Tech. rep., Knoxville, TN, USA, 1994.

[33] Ghrist, R. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society 45*, 1 (2008), 61–75.

[34] Group, T. H. hdf5.

[35] Hatcher, A. *Algebraic Topology*, vol. 1. Cambridge University Press, 2001.

[36] Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc 125*, 39 (2003), 11853–11865.

[37] Hawkins, P. C. D., Skillman, A. G., and Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *Journal of Medicinal Chemistry 50*, 1 (2007), 74–82.

[38] Hu, B., Zhu, X., Monroe, L., Bures, M. G., and Kihara, D. PL-Patchsurfer: A novel molecular local surface-based method for exploring protein-ligand interactions. *International Journal of Molecular Sciences 15*, 9 (2014), 15122–15145.

[39] Hu, Y., Furtmann, N., Gütschow, M., and Bajorath, J. Systematic identification and classification of three-dimensional activity cliffs. *Journal of Chemical Information and Modeling 52*, 6 (2012), 1490–1498.

[40] Jones, E., Oliphant, T., Peterson, P., and Others. SciPy: Open Source and Scientific Tools for Python, 2001.

[41] Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design 30*, 8 (2016), 595–608.

[42] Kerber, M. Persistent Homology – State of the art and challenges. *Internat. Math. Nachrichten 231*, 231 (2016), 15—-33.

[43] Kerber, M., Morozov, D., and Nigmetov, A. Geometry Helps to Compare Persistence Diagrams. 103–112.

[44] Kovacev-Nikolic, V., Bubenik, P., Nikolić, D., and Heo, G. Using cycles in high dimensional data to analyze protein binding. 21.

[45] Landi, C. The rank invariant stability via interleavings. *arXiv preprint arXiv:1412.3374* (2014).

[46] Landrum, G. RDKit: Open-source cheminformatics.

[47] Lesnick, M. The theory of the interleaving distance on multidimensional persistence modules. *Foundations of Computational Mathematics 15*, 3 (2015), 613–650.

[48] Lesnick, M., and Wright, M. Interactive Visualization of 2-D Persistence Modules. 1–75.

[49] Lesnick, M., and Wright, M. Computing Bigraded Betti Numbers in Cubic Time. *In Preparation* (2018).

[50] Li, H., Leung, K.-S., Wong, M.-H., and Ballester, P. J. USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques. *Nucleic Acids Research 2*, 32 (2016), gkw320.

[51] McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (2010), S. van der Walt and J. Millman, Eds., pp. 51–56.

[52] Morozov, D. Dionysus.

[53] Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry 55*, 14 (2012), 6582–6594.

[54] OpenEye Scientific Software. OMEGA.

[55] OpenEye Scientific Software. ROCS.

[56] Oppermann, S. Some background from representation theory, 2017.

[57] Oudot, S. Y. *Persistence Theory: From Quiver Representations to Data Analysis:*. Mathematical Surveys and Monographs. American Mathematical Society, 2015.

[58] Raymond, J. W., Gardiner, E. J., and Willett, P. RASCAL: Calculation of graph similarity using maximum common edge subgraphs. *Computer Journal 45*, 6 (2002), 631–644.

[59] Rotman, J. *An Introduction to Algebraic Topology*. Springer-Verlag New York, New York, 1988.

[60] Scolamiero, M., Chachólski, W., Lundman, A., Ramanujam, R., and Öberg, S. Multidimensional persistence and noise. *Foundations of Computational Mathematics 17*, 6 (2017), 1367–1406.

[61] Sheehy, D. R. Linear-size approximations to the vietoris–rips filtration. *Discrete & Computational Geometry 49*, 4 (2013), 778–796.

[62] Sheehy, D. R. The Persistent Homology of Distance Functions under Random Projection. 328–334.

[63] Shin, W.-H., Zhu, X., Bures, M., and Kihara, D. Three-Dimensional Compound Comparison Methods and Their Application in Drug Discovery. *Molecules 20*, 7 (2015), 12841–12862.

[64] Singh, G., Mémoli, F., and Carlsson, G. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Eurograph. Sympos. Point-Based Graphics* (2007), pp. 91–100.

[65] Steinbusch, B. MPI Bindings for Rust, 2015.

[66] van Rossum, G., and Others. Python Programming Language, 1995.

[67] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci. 28*, 1 (1988), 31–36.

[68] Wright, M. Introduction to Persistent Homology, 2016.

[69] Xia, K., and Wei, G. W. Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering 30*, 8 (2014), 814–844.

[70] Xia, K., and Wei, G. W. Multidimensional persistence in biomolecular data. *Journal of Computational Chemistry 36*, 20 (2015), 1502–1520.

[71] Zomorodian, A. *Computing and Comprehending Topology : Persistence and Hierarchical Morse Complexes*. PhD thesis, University of Illinois at Urbana-Champaign, 2001.