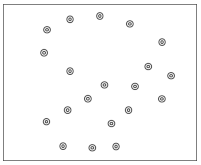
Topological Methods in Data Analysis

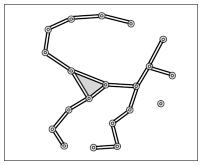
Vidit Nanda

Summary. *I develop algebraic-topological theories, algorithms and software for the analysis of non-linear data and complex systems arising in various scientific contexts. In particular, I employ discrete Morse-theoretic techniques to substantially compress cell complexes built around the input data without modifying their core topological properties. Recently, I have generalized discrete Morse theory itself by recasting it in terms of 2-categorical localization.*

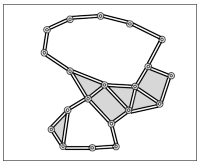
The Multi-Scale Homology of Data



While statistical methods such as regression analysis are extremely efficient tools for analyzing data whose underlying shape is known a-priori, recent use of algebraic topology has had striking success in estimating that underlying shape [4, 24]. The vanguard technique in topological data analysis is *persistent homology* [14]. One begins by constructing a cell complex (filtered by subcomplexes) where vertices coincide with the data points and higher cells are introduced when metrically appropriate.

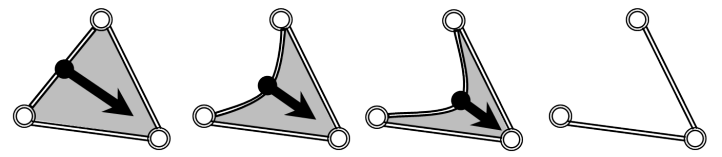


As more cells enter at larger scales, various homological features (such as connected components, tunnels, cavities and so forth) appear and disappear. It follows from rudimentary representation theory that one can unambiguously associate to each such feature an interval $[b,d)$ containing those scales at which it persisted across the filtration. These intervals comprise the *persistence diagram* of our dataset; aside from being a perfect descriptor of the filtered homology of the resulting persistence module (at least over field coefficients), these diagrams are stable with respect to perturbations of the original data [6].



The task of computing persistence diagrams from data reduces to linear algebra: incidence relations among cells of adjacent dimensions produce matrix representations of boundary operators. Standard row and column operations put these matrices in *Smith normal form*, from which persistence intervals can be read off directly [35]. The complexity is cubical in the total number of cells, but the cell count might be exponential in the number of underlying data points[1]. The inevitable burden incurred by keeping track of higher-order metric proximity (rather than pairwise distances used in single-linkage clustering) compels us to try and create smaller filtered complexes with isomorphic homology. This reduction is a primary focus of my research.

Discrete Morse Theory: Homotopy and Homology



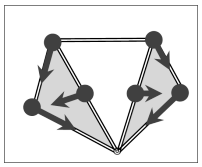
The operation illustrated above removes two adjacent cells from the depicted cell complex. If the smaller cell is a free face of the larger one[2], then this operation preserves homotopy type (and hence, homology) and is called an *elementary collapse*. It dates back to the fundamental work of JHC Whitehead on simple homotopy equivalence [33], which in turn constitutes perhaps the earliest topological motivation for algebraic K-theory. In my research, this operation plays a central role for the somewhat more visceral reason outlined above: it is an engine for homologically faithful data compression.

R Forman's combinatorial analogue of Morse theory [12] provides a principled method to perform several such collapses simultaneously. In this discrete universe, Riemannian manifolds are replaced by cell complexes while *partial pairings* of adjacent cells (subject to a global acyclicity condition) serve as Morse functions. The combinatorial vector field is given flowing from cells down to their boundaries, only making

[1] One can create filtered simplicial complexes with only forty vertices and over a trillion cells!

[2] That is, no other cell contains the smaller cell in its boundary.

exceptions to flow against dimension when paired cells are encountered. The cells left unpaired are called *critical* because they play a similar role to stationary points in smooth Morse theory.



The payoff is gratifying, perhaps even to those already familiar with smooth Morse theory: for each partial pairing on a cell complex X, there is a *Morse complex* M lying in the homotopy class of X whose cells correspond (in number and dimension) to the critical cells. Although the attaching maps of M are not precisely known in general, one can exploit *gradient paths* of the form

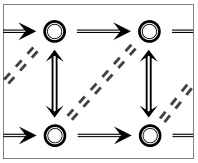
$$y_0 < x_0 > y_1 < x_1 > \cdots > y_k < x_k$$

(where each $y_\bullet$ is paired with $x_\bullet$) to compute the degrees of co-dimension one attaching maps, and hence calculate the Morse homology $H_\bullet(M;R)$ with coefficients in any ring R. By the homotopy-invariance of homology, this Morse homology is isomorphic to $H_\bullet(X;R)$. It is also considerably easier to compute when the number of critical cells is relatively small.

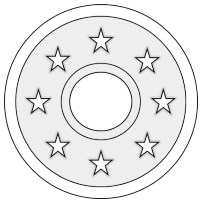
With S Harker, K Mischaikow and M Mrozek, I devised the first discrete Morse-theoretic algorithms for computing homology of cell complexes as well as morphisms induced on homology by maps arising from correspondences of top-dimensional cells [16]. The resulting software library **CHomP** [15] has now found extensive use in computational dynamics.

Algorithms for Filtrations and Local Systems

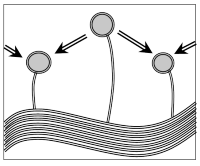
My dissertation work [19, 21] focused on adapting discrete Morse theory to the filtered setting without loss of persistent homology. A partial pairing on a cell complex is *subordinate* to a filtration by subcomplexes when it satisfies the following (additional) property: if cells x and y are paired, then both must have entered the filtration in the same subcomplex. With this crucial modification, the entire discrete Morse-theoretic machinery may be brought to bear on the task of reducing large filtered complexes without altering their persistent homology.



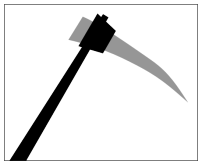
THEOREM A (Mischaikow and Nanda, 2013). *Given a filtered cell complex X with n cells and a subordinate partial pairing, assume that there are m_d critical cells of dimension d. Then, the resulting Morse chain complex is filtered chain-homotopic to the chain complex of X (and hence has the same persistent homology). Moreover, the cost of computing persistent homology reduces from $O(n^3)$ to $O(np\mu_2+\mu_1^3)$, where p bounds the number of co-dimension one neighbors of each cell and $\mu_j=\sum_d (m_d)^j$.*



Computational advantages become apparent when the total number of critical cells μ_1 (and the sum-of-squares μ_2) is considerably smaller than n. When confronting cell complexes built around data points, it is difficult *not* to obtain very few critical cells relative to n even when one resorts to naïve greedy heuristics for constructing the partial pairing. With these considerations in mind, I wrote the **Perseus** software [23] to simplify persistent homology computations. The software was designed to be user-friendly, efficient, and adaptible to a large class of filtered cell complex inputs. Within three years of its release, **Perseus** has been used by several research teams in varied contexts, including (at last count) breast cancer tumor analysis [30], signal processing [29], modeling the spread of contagions [32], the study of granular media [17], and stability analysis of fullerene molecules [34].



Theorem A demonstrates that discrete Morse theory capably adapts to enhancements in the structure of the underlying space (i.e., from a cell complex to a filtration); but it is also handles significantly more general algebraic *coefficients* [31] than those prescribed by a constant ring R. A *local system* or *cosheaf* F over a cell complex X assigns to each cell x its own R-module $F(x)$ and to each face relation $x > y$ a linear map $F(x > y) : F(x) \to F(y)$ subject to a functorial gluing condition [8]. Enriching the coefficient-space from ring elements to modules yields a fruitful and far-reaching generalization: persistent homology across a single scale bears the structure of a cosheaf over the stratified real line.

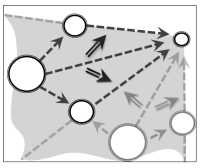


Discrete Morse theory can be used to construct a smaller local system on the critical cells, which provides a shortcut for computing $H_\bullet(X;F)$—the homology of X with F-coefficients. A partial pairing of cells in X is *F-compatible* if for each pair $(x > y)$ the linear map $F(x > y)$ is invertible. With J Curry and R Ghrist, I developed the first algorithm, called **Scythe**, for Morse-theoretic simplification of homology computations with local coefficients [9]. Here is our main result, which assumes that X has n cells and that the rank of each $F(x)$ is bounded above by $r \geq 0$.

THEOREM B (Curry, Ghrist and Nanda, 2015). **Scythe** *constructs F-compatible partial pairings on X; if it produces a pairing with m_d critical cells in dimension d, then the total time complexity of computing $H_\bullet(X;F)$ reduces from $O\left(n^3 r^3\right)$ to $O\left(np\mu_2 r^\omega + \mu_1^3 r^3\right)$. Here the parameters μ_j and p are identical to those from Theorem A, and $\omega < 3$ is the matrix multiplication exponent over R.*

As before, Theorem B confers significant computational benefits when $\mu_j \ll n$. With generous support from the Pacific Northwest National Laboratory's High Performance Data Analytics project, I am currently developing Morse theoretic software to compute homology of cell complexes with local coefficients.

Morse Theory as Categorical Localization



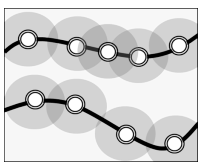
With an eye towards recovering the attaching maps of the discrete Morse complex (and hence an explicit description of its homotopy type rather than homology), I recently became involved in higher-categorical homotopy [28]. The *entrance path category* $\mathbf{Ent}(X)$ of a regular cell complex X has cells of X as objects; morphisms from x to y are given by the poset of *entrance paths*, which are just descending sequences of cells $x > z_0 > \cdots > z_k > y$. Composition is given by concatenation, and the partial order arises from inclusion of sub-sequences. The *classifying space* of this 2-category is homotopy-equivalent to X.

Every partial pairing on the cells of X is precisely a collection $\Sigma = \{(x_\bullet > y_\bullet)\}$ of minimal morphisms in $\mathbf{Ent}(X)$, and every gradient path $y_0 < x_0 > \cdots > y_k < x_k$ of such a pairing is precisely a morphism from y_0 to x_k in the localization $\mathscr{L}_\Sigma\mathbf{Ent}(X)$ of $\mathbf{Ent}(X)$ about Σ. This is no coincidence, as the following result from [20] shows.

THEOREM C (Nanda, 2015). *The canonical localization functor $\mathbf{Ent}(X) \to \mathscr{L}_\Sigma\mathbf{Ent}(X)$ induces a homotopy equivalence of classifying spaces. If we let $\mathbf{Flo}_\Sigma(X)$ denote the full subcategory of $\mathscr{L}_\Sigma\mathbf{Ent}(X)$ spanned by critical cells, then the inclusion $\mathbf{Flo}_\Sigma(X) \hookrightarrow \mathscr{L}_\Sigma\mathbf{Ent}(X)$ also induces homotopy equivalence of classifying spaces.*

Both homotopy equivalences follow from (a 2-categorical version of) D Quillen's Fiber Theorem [27]. We call $\mathbf{Flo}_\Sigma(X)$ the *discrete flow category* associated to the pairing Σ. It has the critical cells as objects, Σ-localized entrance paths as morphisms, and a classifying space which is homotopy-equivalent to X. Thus, $\mathbf{Flo}_\Sigma(X)$ forms a combinatorial and computable analog of the flow category originally described by R Cohen, J Jones and G Segal for smooth Morse functions on compact Riemannian manifolds [5]. Theorem C provides an extension of discrete Morse theory itself: it only relies on a weak lifting axiom, which is satisfied in more general settings than partial pairings on finite cell complexes. With D Tamaki and K Tanaka [25], I have also obtained a direct homotopy-equivalence $\mathbf{Ent}(X) \to \mathbf{Flo}_\Sigma(X)$ without zig-zagging through the localization.

Geometric Inference for Evolving Systems

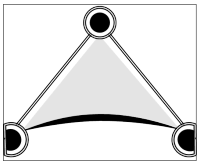


A well-known result of P Niyogi, S Smale and S Weinberger provides explicit bounds on the number of uniformly sampled points required to reconstruct a compact Riemannian manifold $X \subset \mathbb{R}^n$ with high confidence [26]. Given a sufficiently large number of points lying on X, their work produces a simplicial complex Σ_X built around the samples as well as a homotopy-equivalence $\pi_X : \Sigma_X \to X$. From a dynamical perspective, one is also interested in the case where a *function* must be reconstructed using only finitely many evaluations. With S Ferry and K Mischaikow, I recently proved the following result in [11].

THEOREM D (Ferry, Mischaikow and Nanda, 2014). *Given a Lipschitz-continuous function $f : X \to Y$ between compact Riemannian submanifolds of Euclidean space and a probability $\delta > 0$, there exist bounds on the number of points which must be sampled from X and Y so that the following all hold with probability exceeding $(1-\delta)$:*

(1) *the X-samples yield a simplicial reconstruction $\pi_X : \Sigma_X \to X$,*
(2) *the Y-samples yield a simplicial reconstruction $\pi_Y : \Sigma_Y \to Y$, and*
(3) *there is a simplicial map $\Sigma_f : \Sigma_X \to \Sigma_Y$ so that $\pi_Y \circ \Sigma_f$ is homotopic to $f \circ \pi_X$.*

Moreover, Σ_f may be explicitly constructed from knowledge of f restricted to the X-samples.



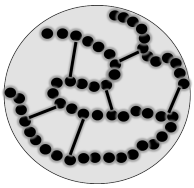
A different, coarser way of analyzing time-series of point samples driven by some unknown underlying transformation involves extracting a sequence of persistence diagrams (one in each dimension). One naturally seeks tools for statistical inference in diagram-space. One might investigate such matters from a geometric perspective by asking what conditions must be satisfied to guarantee the existence of a persistence diagram within some fixed distance of a collection of given diagrams [**10, 2**]. One may view this as a Lipschitz extension problem: given a non-expansive map from some metric space to the space of persistence modules, can we extend that map to a larger metric space while preserving its non-expansiveness? With P Bubenik and V de Silva [**10, 2**], I extracted a categorical *coherence* criterion which imposes higher-order compatibility constraints on non-expansive maps into module-space. With this criterion in place, our Lipschitz extension problem becomes a Kan extension problem, and one obtains the following result.

THEOREM E (Bubenik, de Silva and Nanda, 2016). *Given an arbitrary pair $A \subset M$ of metric spaces, a non-expansive map from A into the space of persistence modules admits a non-expansive extension to all of M if and only if it is coherent.*

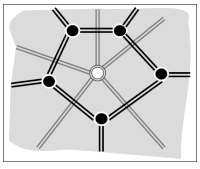
In fact, one can choose the left, right or inner Kan extension to furnish three differently-behaved Lipschitz extensions. The original problem of finding a weighted average for a (coherent) collection of persistence diagrams now reduces to the case where A is a finite metric space and M is obtained from it by adding one point at prescribed distances from all the others.

Topology in Action: Protein Compressibility and Singularity Detection

I've recently been involved in two large interdisciplinary projects where algebraic-topological computations play a major role.



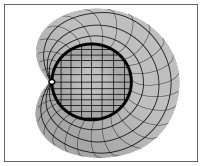
The relationship between physical structure and biological function(s) of protein molecules forms the central focus of contemporary proteomics. An important intermediate step here is the connection between molecular structure and physical properties. One such property is *isothermal compressibility*, which is measured by the change in (log) volume with pressure as temperature is held constant. Compressibility is difficult and expensive to measure experimentally, whereas crystallography data is readily available for most molecules at the protein data bank. In joint work with M Gameiro, Y Hiraoka, S Izumi, M Kramar and K Mischaikow, I devised a topological predictor of protein compressibility (computed as a ratio of certain one and two-dimensional persistence intervals) which requires only crystallography data as input. This topological predictor enjoys a remarkable linear correlation with experimental compressibility values in most cases where those values are available [**13**].



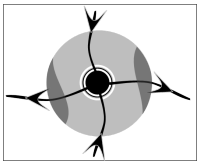
The other project is tied to a research effort coordinated by the US Air Force research lab in Rome, NY and it involves discrete analogues of R Hamilton's *Ricci flow*. Curvature is defined on simplicial complexes embedded in Euclidean space via angle defects in their dual circumcentric complexes [**18**]. As in the smooth case, singularities might develop in the intermediate complexes as one deforms to uniformize curvature; if this occurs, one must manually perform surgery before flow can safely resume. In [**1**], P Alsing, H Blair, M Corne, G Jones, W Miller, K Mischaikow and I implemented the following scheme to automatically detect when human intervention might be necessary. We filtered the intermediate complexes by curvature

values and extracted a sequence of persistence diagrams. Not only does (a simple transform of) persistence interval data detect singularity-formation early, but each singularity also has a characteristic *signature* in persistence-space.

Ongoing Work

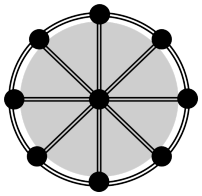


Local cohomology and data stratification. Modern data is often far too heterogeneous to readily succumb to conventional manifold learning techniques: it rarely lies on a single manifold, since its dimensionality can vary drastically from place to place. One therefore seeks principled tools to stratify large datasets optimally into pieces which are locally homogeneous and manifold-like, so that the vast array of existing geometric and statistical methods may be brought to bear on the individual homogeneous pieces. The *local cohomology* of a space is a remarkably efficient detector of dimensionality: it simply uses the fact that sufficiently small neighborhoods around points in n-dimensional manifolds have n-dimensional spheres as boundaries — these only have nontrivial homology in dimensions 0 and n. By clustering data points whenever boundaries of their local neighborhoods exhibit near-isomorphic (persistent) homology, it therefore becomes possible to partition the data into strata organized by dimension. In order to establish that such clustering produces optimal strata, one must construct and projectively resolve a complex of cellular cosheaves on a suitable cell complex built around the underlying data. Fortunately, the practical aspects of algorithmically discovering the various strata are far less intricate: all required computations are local, and can therefore be efficiently distributed across several processors [**22**].



Conley theory and reconfigurable systems. One of the most useful generalizations of smooth Morse theory is furnished by the *Conley index* [**7**], which admits invariant sets far more general than stationary points of a Morse function by enhancing Morse indices to relative (co)homology or homotopy classes. The goal in our discretized case is perhaps best-understood in the setting of Theorem C: partial pairings on regular cell complexes have two basic restrictions. The first is local and removable (and indeed, removed by Theorem C), as it only requires all pairs $(x > y)$ to satisfy $\dim x - \dim y = 1$. A considerably more severe constraint is imposed by the global acyclicity requirement: one must ban partial pairings whose gradient paths $y_0 < x_0 > y_1 < \cdots$ trap the flow either by continuing infinitely or by forming loops. But this is precisely the situation that Conley index theory solves in the smooth case: one must enhance the notion of critical regions from single cells to more general trapping regions of the discrete flow, and suitably broaden the construction of the discrete flow category.

A concrete application of this Conley-based flow category would be towards simplifying control for reconfigurable systems in robotics. Each such system has a complicated configuration space, which is typically discretized into a regular cell complex. One can isolate the frequently-occuring configurations and impose a discrete vector field (using, for instance, the algorithm mentioned in Theorem B) so that all frequent configurations are critical while other cells correspond to transient, gradient-like flow between them.



Symmetry-based compression and inference. A standard principle in random graph theory asserts that most graphs have no non-trivial automorphisms. More precisely, given a graph G on n vertices, if the edge count is more than $O(\log n)$ far from 0 and $\binom{n}{2}$, then its automorphism group $\mathrm{Aut}(G)$ is trivial almost surely as $n \to \infty$. As a corollary, random simplicial complexes also do not admit non-trivial automorphisms away from highly dense and sparse regimes. This basic result provides a wonderful null hypothesis for analyzing filtered simplicial complexes built around large data sets: if any non-trivial automorphisms are detected at intermediate scales, then not only are those scales inherently interesting, but also the underlying data is almost certainly not randomly generated. With L Carbone and Y Naqvi [**3**], I am using a general version of Bass-Serre theory to compress (and reconstruct) a simplicial complex X from a presheaf of stabilizer subgroups over a fundamental domain X/G for any subgroup $G < \mathrm{Aut}(X)$. Simultaneously, I am also investigating an equivariant version of Theorem D with S Ferry: can one learn the

automorphism group of a Riemannian submanifold of Euclidean space from (approximate symmetries of) a sufficiently large point sample lying near that manifold?

References

[1] P. Alsing, H. Blair, M. Corne, G. Jones, W. Miller, K. Mischaikow and V. Nanda. Topological signals of singularities in simplicial Ricci flow. *Axioms*, 6(3), Article 24, 2017.

[2] P. Bubenik, V. de Silva and V. Nanda. Higher interpolation and extension for persistence modules. *SIAM Journal on Applied Algebra and Geometry*, 1(1):272–284, 2017.

[3] L. Carbone, V. Nanda and Y. Naqvi. Reconstructing group actions on graphs. *In preparation*, 2017

[4] G. Carlsson. Topology and data. *Bulletin of the AMS*, 46(2):255–308, 2009.

[5] R. Cohen, J. Jones, and G. Segal. Morse theory and classifying spaces. *Warwick University Preprint*, http://math.stanford.edu/~ralph/morse.ps, 1995.

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

[7] C. Conley. *Isolated Invariant Sets and the Morse Index*. Volume 38 of the CBMS Regional Conference Series in Mathematics, AMS, 1978.

[8] J. Curry. Sheaves, cosheaves and applications. arXiv:1303.3255 [math.AT], 2013.

[9] J. Curry, R. Ghrist, and V. Nanda. Discrete Morse theory for computing cellular sheaf cohomology. *Foundations of Computational Mathematics*, 16(4):875–897, 2016.

[10] V. de Silva and V. Nanda. Geometry in the space of persistence modules. In *Proc. 29-th Annual Symposium on Computational Geometry*, 397–404, 2013.

[11] S. Ferry, K. Mischaikow, and V. Nanda. Reconstructing functions from random samples. *Journal of Computational Dynamics*, 1(2):233–248, 2014.

[12] R. Forman. Morse theory for cell complexes. *Advances in Mathematics*, 134:90–145, 1998.

[13] M. Gameiro, Y. Hiraoka, S. Izumi, M. Kramar, K. Mischaikow, and V. Nanda. A topological measurement of protein compressibility. *Japan Journal of Industrial and Applied Mathematics*, 32(1):1–17, 2015.

[14] R. Ghrist. Barcodes: the persistent topology of data. *Bulletin of the AMS*, 45(1):61–75, 2008.

[15] S. Harker et al., The Computational HOMology Project www.chomp.rutgers.edu, 2009.

[16] S. Harker, K. Mischaikow, M. Mrozek, and V. Nanda. Discrete Morse theoretic algorithms for computing homology of complexes and maps. *Foundations of Computational Mathematics*, 14(1):151 – 184, 2014.

[17] M. Kramar, A. Goullet, L. Kondic and K. Mischaikow Quantifying force networks in particulate systems. *Physica D: Nonlinear Phenomena* 283:37–55, 2014.

[18] W. Miller, J. McDonald, P. Alsing, D. Gu, and S. Tung Yau. Simplicial Ricci flow. *Communications in Mathematical Physics*, 329(2):579–608, 2014.

[19] K. Mischaikow and V. Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete and Computational Geometry*, 50(2):330–353, 2013.

[20] V. Nanda. Discrete Morse theory and localization. *Under review*, arXiv:1510.01907 [math.AT], 2017.

[21] V. Nanda. Discrete Morse theory for filtrations. *PhD Thesis*, Rutgers University, 2012.

[22] V. Nanda. Local cohomology and stratification. *Under review*, arXiv:1707.00354 [math.AT], 2017.

[23] V. Nanda. Perseus: the persistent homology software, people.maths.ox.ac.uk/nanda/perseus, 2014.

[24] V. Nanda and R. Sazdanovic. Simplicial models and topological inference in biological systems. Chapter 6 of *Discrete and Topological Models in Molecular Biology*, (eds. N Jonoska and M Saito). Volume 13 of Springer's *Natural Computing Series*, 2014.

[25] V. Nanda, D. Tamaki, and K. Tanaka. Discrete Morse theory and classifying spaces. *Under review*, arXiv:1612.08429 [math.AT], 2016.

[26] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39:419–441, 2008.

[27] D. Quillen. Higher algebraic K-theory I. *Lecture Notes in Mathematics*, 341:85–147, 1973.

[28] E. Riehl. *Categorical Homotopy Theory*. Cambridge University Press, 2014.

[29] M. Robinson. *Topological Signal Processing*. Springer, ISBN: 978-3-642-36103-6, 2015.

[30] N. Singh, H. Couture, J. Marron, C. Perou and M. Niethammer. Topological descriptors of histology images. In *Breast Cancer: Machine Learning in Medical Imaging*, 8679:231–239, 2014.

[31] E. Sköldberg. Morse theory from an algebraic viewpoint. *Transactions of the AMS*, 358(1):115–129, 2006.

[32] D. Taylor, F. Klimm, H. Harrington, M. Kramar, K. Mischaikow, M. Porter and P. Mucha. Topological data analysis of contagion maps for examining spreading processes on networks. *Nature Communications 6*, Article number 7723, 2015.

[33] J. H. C. Whitehead. Combinatorial homotopy I. *Bulletin of the American Mathematical Society*, 55(5):453–496, 1949.

[34] K. Xia, X. Feng, Y. Tong and G. Wei. Persistent homology for the quantitative prediction of fullerene stability. *Journal of Computational Chemistry* 36(6):408–422, 2015.

[35] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.