УДК 004.896
DOI: https://doi.org/10.53920/ITS-2024-2-7

Максим Богданович ЦІЖ,
кандидат фізико-математичних наук, науковий співробітник,
Астрономічна обсерваторія Львівського національного університету
ім. І. Франка,
Факультет фізики та астрономії, Болонський університет
ORCID ID: 0000-0002-4480-3185

ОГЛЯД ЗАСТОСУВАННЯ СУЧАСНИХ МЕТОДІВ ТОПОЛОГІЧНОГО АНАЛІЗУ ДАНИХ

В цій статті наведено систематичний огляд застосування базових інструментів сучасного топологічного аналізу даних, зокрема, пов'язаних з стійкими (персистентними) гомологіями. Йдеться про діаграми стійкості, числа та криві Бетті. Аналіз та порівняння даних типу хмара точок з допомогою цих інструментів знайшли широке застосування в останнє десятиліття. Зокрема, в таких галузях як, комп'ютерний зір, аналіз біомедичних даних, аналіз часових рядів, матеріалознавство, дослідження великомасштабної структури Всесвіту, аналіз стохастичних та перколяційних процесів, а також в ряді інших фундаментальних та прикладних областях науки. Спираючись на спрощені доступні означення математичних понять симплексів, стійких гомології та інструментів, що на них базуються, ми оглянемо та виділимо найбільш яскраві приклади їхнього застосування в топологічному аналізі даних.

Ключові слова: *топологічний аналіз даних, стійкі гомології, обробка даних, алгоритми.*

Maksym TSIZH,
Philosophy Doctor in field of astrophysics
Astronomical Observatory of Lviv National University named after I. Franko
Dipartimento di Fisica e Astronomia, Universit´a di Bologna

REVIEW OF THE APPLICATION OF MODERN METHODS OF TOPOLOGICAL DATA ANALYSIS

This article presents a systematic review of the use of basic tools of modern topological data analysis, in particular, those related to stable (persistent) homologies. We are talking about stability diagrams, numbers,

and Betty curves. The analysis and comparison of point cloud data using these tools have been widely used in the last decade. In particular, in such fields as computer vision, biomedical data analysis, time series analysis, materials science, studies of the large-scale structure of the Universe, analysis of stochastic and percolation processes, as well as in a number of other fundamental and applied fields of science. Based on the simplified accessible definitions of the mathematical concepts of simplexes, stable homologies, and the tools based on them, we will review and highlight the most striking examples of their application in topological data analysis.

Keywords: *topological data analysis, persistent homologies, data processing, algorithms.*

Постановка проблеми. Топологічні методи аналізу даних, чи просто топологічний аналіз даних (ТАД) — це область галузі обробки даних що зазнала стрімкого розвитку за останні десятиліття. Це пов'язано як із значним розширенням обчислювальних можливостей так і з створенням зручних програмних інтерфейсів та фреймворків для відповідних обчислень. При цьому на даний час відсутній систематичний огляд застосування методів топологічного аналізу даних в різних галузях науки і техніки.

Топологічний аналіз даних базується на концепції сприйняття даних як хмари точок в просторі фізичних координат чи просторі вимірюваних параметрів. Його об'єктом досліджень є в першу чергу геометрична форма та топологія відповідної хмари точок. При цьому для їх дослідження використовується певне наближення — фільтрація симплексного комплексу відповідної хмари точок та альфа-комплексу (Вієторіса-Ріпса, Чеха). Процедура фільтрації полягає в побудові послідовного ряду таких комплексів, де довжина з'єднання (радіус фільтрації) між вершинами комплексу зростає поступово від нуля до (умовно) нескінченності. При цьому, досліджуються топологічні властивості многовиду, що утворюється шляхом побудови сфер радіусу фільтрації навколо кожної вершини під час процесу фільтрації. Утворений ряд многовидів характеризується зокрема своїми порожнинами (групами гомологій нестягуваних циклів). Кожна з утворених порожнин таким чином існує на своєму відрізку радіуса фільтрації. Ті з них, які є найбільш довготривалими (стійкими) в процесі фільтрації як виявилось, грають також і найбільшу роль в характеризації вихідної хмари точок як набору даних. Таким чином, опис стійких гомологій (тобто порожніх, стійких в процесі фільтрації) є одним із найпрості-

ших інструментів ТАД і як ми побачимо далі, знаходить широке застосуваннях в найрізноманітніших ситуаціях.

Аналіз останніх досліджень і публікацій. Ряд задач з обробці даних часто зводиться до порівняння хмар точок. Ці точки можуть бути виміряні або обчислені положення реальних фізичних об'єктів (галактики, дефекти в твердому тілі, особливості на зображені) чи бути точками в просторі параметрів. Виявляється, немає прямого поняття відстані між двома хмарами точок: розгляд кожної окремої точки призводить до «інформаційного переповнення», неоднозначності, і в більшості випадків буде занадто чутливим до невеликих змін у положенні точок. Вирішити проблему, можна якщо не розглядати окремі точки, а натомість витягти певну узагальнену інформацію про їх розподіл, який міститиме керовану кількість чисел, а потім шукати деяке поняття відстані.

Мета статті – огляд застосування таких інструментів топологічного аналізу даних як діаграми стійкості гомологічних особливостей (дірок) та криві Бетті (в поєднанні і з складнішими інструментами ТАД) у різних галузях наукових досліджень, та проаналізувати застосування таких методів і порівняти їхнє застосування у максимально різноманітних областях, де наука стикається із опрацюванням даних, що мають складну структуру, оскільки дані методи обробки покликані бути застосованими в першу чергу у складних високорозмірних наборах даних. Показати наскільки широким стало можливе застосування методів ТАД і зокрема, тих, що базуються на стійких гомологіях в сучасних галузях наук та мають справу з великою кількістю даних та статистичним висовуванням на їх основі.

Виклад основного матеріалу дослідження. Одним із можливих підходів є наділення кожної точки сферою радіуса r з центром у точці, і розгляд об'єднання всіх цих куль у многовид. Цей многовид матиме певні топологічні властивості, які можуть бути представлені чисельно та використані далі. Таким чином, можна використовувати числа Бетті – ранги відповідних груп гомології даного многовиду. Для простоти можна вважати 0-число Бетті числом зв'язних компонентів, 1-число Бетті як ряд одновимірних або «плоских» порожнин, 2- як кількість двовимірних порожнин. Представлення хмари точок через числа Бетті є досить стабільним щодо переміщення, обертання та невеликої варіації положень точок в хмарі, що є бажаною властивістю.

Однак, вибір радіуса r цих сфер може бути спірним. Щоб вирішити цю проблему, можна розглянути всі можливі значення r і відстежувати

топологічні особливості (порожнини), по мірі того як многовид змінюється зі зміною r. Звідси назва – «стійка гомологія»: кожна топологічна особливість зберігається в межах певного діапазону зміни параметра r (параметра фільтрації). Відповідні числа Бетті змінюються, коли такі особливості (дірки) «народжуються» або «вмирають». Результат можна представити як графік залежності чисел Бетті від параметра r.

Також ми можемо відобразити «народження» і «смерть» кожної окремої топологічної особливості у формі діаграми стійкості. Цікаво, що існує поняття відстані між діаграмами стійкості із строгим математичним сенсом, яке носить назву «відстань Вассерштейна», а його частковий випадок називають «горловинна відстань» (bottleneck distance).

На рис. 1 показано процес фільтрації на прикладі простої хмари точок в двомірному просторі. В даному випадку досліджуваним многовидом буде об'єднання кругів на кожному кроці збільшення радіусів. Видно, як спочатку многовид є незв'язним взагалі, але поступово він перетворюється в однозв'язний. Найцікавіше відбувається посередині — в процесі фільтрації виникають і зникають порожнини. При цьому видно, що найбільш стійкою є порожнина, що відповідає "формі" хмари точок –колу..

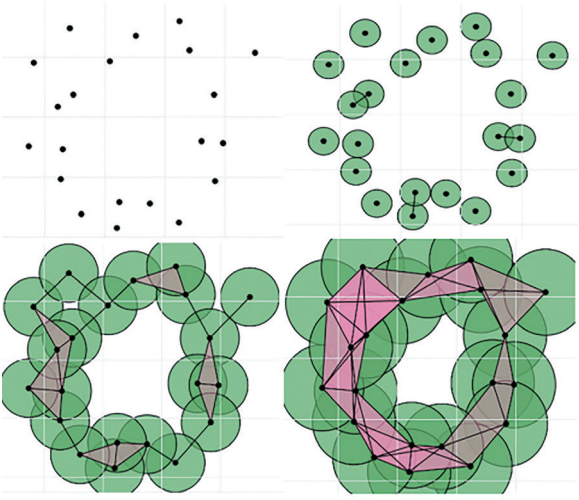

Рис. 1. Процес фільтрації на прикладі простої хмари точок в двомірному просторі

На рис. 2 дано порівняння діаграм стійкості для двох многовидів – сфери та тора, що мають різну топологію. Точки випадково населяють поверхні многовидів. На діаграмах стійкості помітну принципову різницю між многовидами – у сфери лише одна двомірна стійка особливість, а у тора — одна двомірна і дві одномірні стійкі гомологічні особливості (нестягувані дірки).

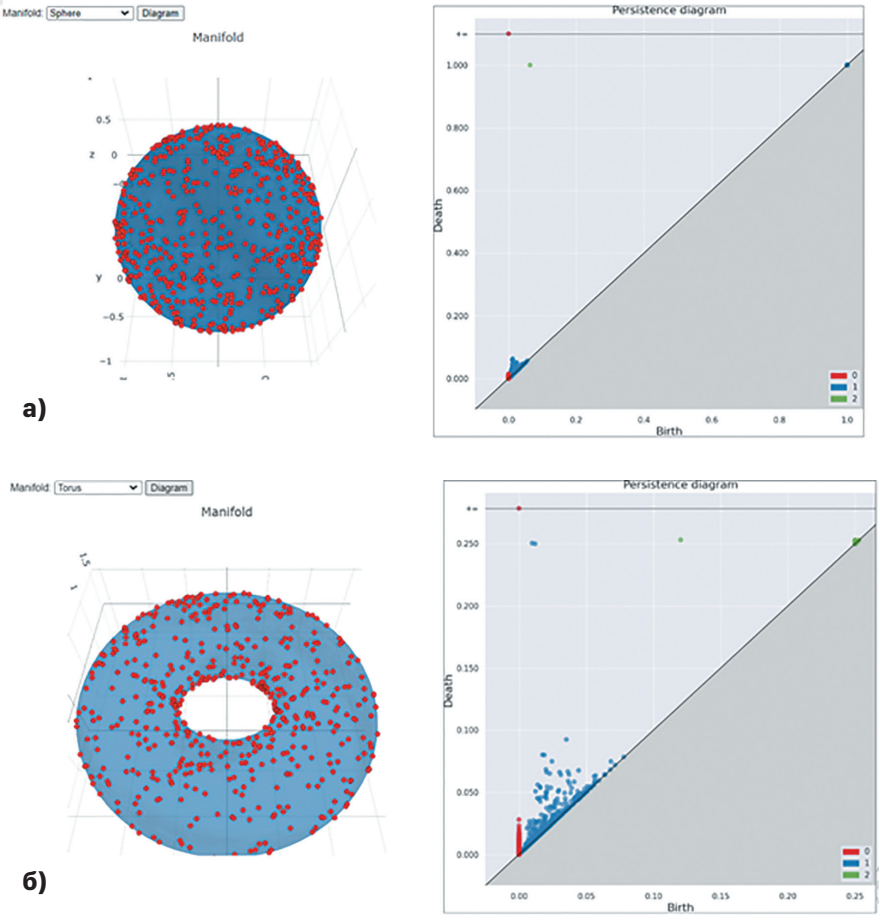

Рис. 2. Порівняння діаграм стійкості для двох многовидів: а) сфери; б) тора

Джерело: виконано автором

Комп'ютерний зір – це одна з перших галузей, куди почали проникати методи ТАД. Знаковою стала рання стаття [1] що задала важливий напрям - кодування зображень через його стійкі гомології. При цьому зображення (після бінаризації) розглядається як граф із 4-звіязністю. Також, методи базовані на стійких гомологіях чудово проявили себе в такій поширеній задачі як класифікація зображень рукописних цифр [2] (на класичному датасеті MNIST). Крім цього, цікавим прикладом є дослідження того, як змінюється картина стійких гомологій зображення при зміні роздільної здатності, яке було зроблено в [3]. В цілому, стійкі топологічні властивості зображення часто служать додатковим дескриптором, що забезпечує розуміння, яке не може бути виявлено традиційними нейронними мережами чи іншими методами машинного навчання застосованого до зображень. Існуючі дослідження в цій галузі зосереджені насамперед на ефективній інтеграції топологічних властивостей даних у процес навчання з метою підвищення продуктивності. В роботі [4] доступний широкий огляд такого застосування ТАД в обробці зображень.

Машинне навчання. Зручне представлення даних, яке може як зберегти внутрішню інформацію, так і зменшити їх складність і розмірність, є ключовим для продуктивності моделей машинного навчання. Глибоко вкорінені в алгебраїчній топології, стійкі гомології забезпечують тонкий баланс між спрощенням даних і характеристикою внутрішньої структури, і її успішно застосовують у різних областях машинного навчання. В цілому, основними напрямами поєднання ТАД та машинного навчання є: топологічне представлення особливостей (features) на вхід моделей машинного навчання, відстані та функції подібності між наборами даних, що базуються на стійких гомологіях та зменшення розмірності та редукція даних. Всі ці способи застосування, які і деякі інші наведені в широкому огляді застосування стійких гомологій в машинному навчанні [5]. А в огляді [6] наведено приклади застосування, що стосуються різних фізичних задач, від мезо- до наномасштабу та квантової фізики. Інші прикладні застосування міксу машинного навчання та ТАД що базується на стійких гомологіях, в таких областях як хімія, біологія, геофізика, наведено в [7].

Стохастичні процеси. Картина стійких гомологій для випадкових процесів є цікавою сама по собі, адже випадкові процеси моделюють дуже багато різних явищ в природі, науці та техніці. Отже, вивчення топології випадкових хмар точок дасть змогу, наприклад, зрозуміти в яких випадках ми стикаємось з випадковим шумом, і чи присутній той чи інший важливий сигнал в даних. В роботі [8] автор показує що для

Пуасонівського точкового процесу характерною є певна форма кривих Бетті його гомологій. А в роботі [9] автори вивчають як конкретно залежать від густини точок та крайових умов форма та амплітуда кривих Бетті різних розмірностей для цього ж таки Пуасонівської хмари точок (хмари точок що має рівномірно розподілену ймовірність заповнення об'єму).

Біомедицина. ТАД знайшов широке застосування в аналізі медичних, біологічних та біомедичних даних. Зокрема, стійкі гомології застосовуються, як для аналізу аналізу часово-змінних даних (дослідження ритму биття серця в [10]), так і для аналізу зображень (наприклад, класифікація на основі стійких гомологій зображень пухлин легень та мозку) [11]. Особливо виділяється таке цікаве застосування ТАД в медицині як дослідження інтерактому протеїн-протеїнових взаємодій в тілі людини [12]

З чудовим оглядом робіт на тему застосування ТАД саме в біомедичній галузі можна ознайомитись в [13].

Матеріалознавство. Діаграми стійкості та інші інструменти стійких гомологій в науці про матеріали застосовуються для структурного аналізу скла, кристалізації гранульованих систем і формування дефектів в полімерах, приклади цих застосувань зібрані в [14]. Для застосування ТАД в матеріалознавстві авторами [15] було також розроблено спеціалізоване програмне забезпечення HomCloud. В цій самій роботі автори наводять інші цікаві приклади того, як стійкі гомології допомагають у вивченні матеріалів: структури гранульованих матеріалів (див. також [16]), моделювання полімерів та м'якої матерії, вираження локальної структури двовимірної бінарної колоїдної конфігурації, що обмежена межею розділу газ-рідина тощо.

Великомасштабна структура Всесвіту. Для дослідження великомасштабної структури Всесвіт ТАД є відносно новим, але логічним і природним способом дізнатись більше про еволюцію і властивості цієї структури. Після появи гігантських космологічних симуляцій, а також розростання каталогів спостережних галактик починаючи з кінця минулого століття вчені виділяють окреме поняття Космічної павутини - як окремого явища, способу у який галактики та скупчення галактики розташовуються одне відносно одного на найбільших масштабах спостережного Всесвіту. Таку структуру, яка формується під дією гравітаційної взаємодії і розширення Всесвіту можна сприймати як мережу, вершинами якої будуть галактики або скупчення галактик. Властивості Космічної павутини (мережі) досліджують як

граф-теоретичними методами (обчислюючи мережеві характеристики вершин, такі як кількість сусідів вершини, чи його ступінь посередництва), так і методами ТАД. Так, в роботах [17] та [18] автори показують, що криві Бетті та діаграми стійкості Космічної павутини є чутливими до космологічних параметрів. А в роботі [19] така ж чутливість показано від проміжку мас популяції гало, що розглядається.

Висновки та пропозиції. В цій статті було зроблено систематичний огляд застосування елементарних інструментів топологічного аналізу даних - гілки науки про данні, що динамічно розвивається в останні десятиліття - у таких галузях як комп'ютерний зір, машинне навчання, стохастичні процеси та їх використанні у біомедицині, матеріалознавстві та вивчені багатомасштабної структури Всесвіту.

Було показано, що такі інструменті як: діаграми стійкості, числа та криві Бетті володіють рядом переваг при розгляді даних, що можуть бути представлені як хмари точок. До цих переваг належать: стійкість до шуму в даних, безмасштабність (непараметричність) підходу, універсальність. Як наслідок, ці інструменти застосовуються в дуже широкому ряді областей науки та обробки даних, від космології до нанофізики, від біомедицини до вивчення стохастичних процесів.

Отже, серед пропозицій можна виділити: ширше застосування методів топологічного аналізу даних, зокрема, заснованих на стійких гомологіях у комп'ютерних науках та науках про обчислення та аналіз даних у вітчизняному науковому середовищі, впровадження цих методів у практиці профільних наукових інститутів України.

Ця робота була виконана за фінансової підтримки Національного Фонду Досліджень України, номер гранту No. 2023.03/0098 (державний реєстраційний номер 0124U004029)

ЛІТЕРАТУРА

1. Li, C., Ovsjanikov, M., & Chazal, F. (2014). Persistence-Based Structural Recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*.

2. Garin, A., & Tauzin, G. (2019). A Topological "Reading" Lesson: Classification of MNIST using TDA.

3. Heiss, T., Tymochko, S., Story, B., Garin, A., Bui, H., Bleile, B., & Robins, V. (2021). The Impact of Changes in Resolution on the Persistent Homology of Images.

4. Khramtsova, E., Zuccon, G., Wang, X., & Baktashmotlagh, M. (2022). Rethinking Persistent Homology for Visual Recognition.

5. Pun, C.S., Lee, S.X., & Xia, K. (2022). Persistent-homology-based machine learning: a survey and a comparative study. *Artificial Intelligence Review*, 55(7), 5169–5213.

6. Leykam, D., & Angelakis, D.G. (2023). Topological data analysis and machine learning. *Advances in Physics: X*, 8(1).

7. Yavorskyi, O., Asseko-Nkili, A., & Kussul, N. (2023). Persistent Homology in Machine Learning: Applied Sciences Review.

8. Robins, V. (2006). Betti number signatures of homogeneous Poisson point processes. *Physical Review E*, 74(6).

9. Gluzberg, V.E., & Katz, Y.A. (2023). Topological data analysis of noise: Uniform unimodal distributions. *Communications in Nonlinear Science and Numerical Simulation*, 121, 107216.

10. Chung, Y.-M., Hu, C.-S., Lo, Y.-L., & Wu, H.-T. (2021). A Persistent Homology Approach to Heart Rate Variability Analysis With an Application to Sleep-Wake Classification. *Frontiers in Physiology*, 12.

11. Moon, C., Li, Q., & Xiao, G. (2023). Using persistent homology topological features to characterize medical images: Case studies on lung and brain cancers. *The Annals of Applied Statistics,* 17(3).

12. Song, E. (2023). Persistent homology analysis of type 2 diabetes genome-wide association studies in protein-protein interaction networks. *arXiv*.

13. Skaf, Y., & Laubenbacher, R. (2022). Topological data analysis in biomedicine: A review. *Journal of Biomedical Informatics*, 130, 104082.

14 Buchet, M., Hiraoka, Y., & Obayashi, I. (2018). Persistent Homology and Materials Informatics. *Nanoinformatics*, 75–95.

15. Obayashi, I., Nakamura, T., & Hiraoka, Y. (2022). Persistent Homology Analysis for Materials Research and Persistent Homology Software: HomCloud. *Journal of the Physical Society of Japan*, 91(9).

16. Mei, J., Ma, G., Liu, J., Nicot, F., & Zhou, W. (2023). Modeling shear-induced solid-liquid transition of granular materials using persistent homology. *Journal of the Mechanics and Physics of Solids*, 176, 105307.

17. Cisewski-Kehe, J., Fasy, B.T., Hellwing, W., Lovell, M.R., Drozda, P., & Wu, M. (2022). Differentiating small-scale subhalo distributions in CDM and WDM models using persistent homology. *Physical Review D*, 106(2).

18. Tsizh, M., Tymchyshyn, V., & Vazza, F. (2023). Wasserstein distance as a new tool for discriminating cosmologies through the topology of large-scale structure. *Monthly Notices of the Royal Astronomical Society*, 522(2), 2697–2706.

19. Bermejo, R., Wilding, G., van de Weygaert, R., Jones, B.J.T., Vegter, G., & Efstathiou, K. (2024). Topological bias: how haloes trace structural patterns in the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 529(4), 4325–4353.

REFERENCES

1. Li, C., Ovsjanikov, M., & Chazal, F. (2014). Persistence-Based Structural Recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*.

2. Garin, A., & Tauzin, G. (2019). A Topological "Reading" Lesson: Classification of MNIST using TDA.

3. Heiss, T., Tymochko, S., Story, B., Garin, A., Bui, H., Bleile, B., & Robins, V. (2021). The Impact of Changes in Resolution on the Persistent Homology of Images.

4. Khramtsova, E., Zuccon, G., Wang, X., & Baktashmotlagh, M. (2022). Rethinking Persistent Homology for Visual Recognition.

5. Pun, C.S., Lee, S.X., & Xia, K. (2022). Persistent-homology-based machine learning: a survey and a comparative study. *Artificial Intelligence Review*, 55(7), 5169–5213.

6. Leykam, D., & Angelakis, D.G. (2023). Topological data analysis and machine learning. *Advances in Physics: X*, 8(1).

7. Yavorskyi, O., Asseko-Nkili, A., & Kussul, N. (2023). Persistent Homology in Machine Learning: Applied Sciences Review.

8. Robins, V. (2006). Betti number signatures of homogeneous Poisson point processes. *Physical Review E*, 74(6).

9. Gluzberg, V.E., & Katz, Y.A. (2023). Topological data analysis of noise: Uniform unimodal distributions. *Communications in Nonlinear Science and Numerical Simulation*, 121, 107216.

10. Chung, Y.-M., Hu, C.-S., Lo, Y.-L., & Wu, H.-T. (2021). A Persistent Homology Approach to Heart Rate Variability Analysis With an Application to Sleep-Wake Classification. *Frontiers in Physiology*, 12.

11. Moon, C., Li, Q., & Xiao, G. (2023). Using persistent homology topological features to characterize medical images: Case studies on lung and brain cancers. *The Annals of Applied Statistics*, 17(3).

12. Song, E. (2023). Persistent homology analysis of type 2 diabetes genome-wide association studies in protein-protein interaction networks. *arXiv*.

13. Skaf, Y., & Laubenbacher, R. (2022). Topological data analysis in biomedicine: A review. *Journal of Biomedical Informatics*, 130, 104082.

14 Buchet, M., Hiraoka, Y., & Obayashi, I. (2018). Persistent Homology and Materials Informatics. *Nanoinformatics*, 75–95.

15. Obayashi, I., Nakamura, T., & Hiraoka, Y. (2022). Persistent Homology Analysis for Materials Research and Persistent Homology Software: HomCloud. *Journal of the Physical Society of Japan*, 91(9).

16. Mei, J., Ma, G., Liu, J., Nicot, F., & Zhou, W. (2023). Modeling shear-induced solid-liquid transition of granular materials using persistent homology. *Journal of the Mechanics and Physics of Solids,* 176, 105307.

17. Cisewski-Kehe, J., Fasy, B.T., Hellwing, W., Lovell, M.R., Drozda, P., & Wu, M. (2022). Differentiating small-scale subhalo distributions in CDM and WDM models using persistent homology. *Physical Review D*, 106(2).

18. Tsizh, M., Tymchyshyn, V., & Vazza, F. (2023). Wasserstein distance as a new tool for discriminating cosmologies through the topology of large-scale structure. *Monthly Notices of the Royal Astronomical Society*, 522(2), 2697–2706.

19. Bermejo, R., Wilding, G., van de Weygaert, R., Jones, B.J.T., Vegter, G., & Efstathiou, K. (2024). Topological bias: how haloes trace structural patterns in the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 529(4), 4325–4353.

СТАТТЯ НАДІЙШЛА ДО РЕДАКЦІЇ 01.10.2024