Dear Dr. Becker,

We thank you and the anonymous reviewers for the thoughtful and insightful suggestions regarding our manuscript. Below, please find a point-by-point accounting of our responses to the reviews.

Associate Editor comments (Remarks to Author): *I have sent the manuscript out to two referees and received those reports back, and I have also looked at your paper myself. Both referees have identified quite a few items they would like to see addressed, but overall I'm pleased to report that I think the manuscript has merit and impact, and so I am recommending a major revision. In addition to rephrasing, reorganizing and clarifying the paper (per the referees' comments), the two major items to address are (1) details about the real world examples (cf comments from referee 1), and (2) referee 2 raises an issue about different ground truths.*

Referee #2 (Remarks to the Author):

In "Let them have CAKES [...]", the authors describe a new method to perform variants of the nearest neighbor search. The Rust implementation of this new algorithm is relatively (comparable or somewhat slower than the competition) but has a much better recall across many different tasks.

The description of the method is clear, although the description of the run time costs is ambiguous. Overall, CAKES is an interesting contribution to the field of NN-search.

Major comments

The big O notation by default suggest worst-case run time complexity. In this manuscript, it seems that complexity is a mixture of expected run time complexity and under the loose assumption that the data has particular characteristics (e.g., low fractal dimension, the tree is balanced). These assumptions should be made clear in the intro and when giving the run time expressions.

We agree that because the time complexity is highly dependent on the manifold structure of the data, this was insufficiently rigorous. We have attempted to address this in a few ways. First, we added the condition "when the dataset has a manifold structure" to our statement about asymptotic complexity of entropy-scaling search in Section 1.2. Additionally, we added a paragraph to Section 1.2 justifying our approach to complexity analyses, and restated it in Section 2.2.2 and Section 2.2.5 before giving runtime expressions.

For example, line 215, does the cost of partitioning of O(n log(n)) depend on the tree being balanced? It is an important hypothesis because line 197 states that "the imbalance in the tree is a feature, not a bug, as it reflects the underlying structure of the data". That may look like a contradiction.

We have added Section 4.4 to compare the performance of balanced and unbalanced clustering. We show that balanced trees perform strictly worse during search.

The cost on line 260 uses $\hat{r}$, which is a mean radius. Now that cost is in expectation (in addition to the balanced tree hypothesis).

This is correct; we added justification for our approach to complexity analyses in Sections 1.2, 2.2.2, and 2.2.5.

Another example, line 326, comes from the minimum with value 2 on line 284. If the assumption on the LFD is not satisfied, then an extra log factor would appear in the cost.

We have, in Theorem 2.1, explicitly stated the assumption that the query is drawn from the same distribution as the other points in the dataset. This should ensure that, in practical applications with real-world datasets, the LFD around a query is similar to the LFD around its nearest neighbors. In the theoretical worst case where the LFD around a query is extremely low, repeating tree-search with doubling radii might not encounter any additional points, we may introduce an extra factor (logarithmic in the distance between the query and the k-th nearest neighbor) the in contribution of the tree search to the overall complexity. However, we can easily detect and sidestep this issue in our implementation by calculating the distance from the query to the center of the root cluster and using that distance, along with the radius and LFD of the root cluster, to determine an initial radius for tree-search. This approach, once again, brings us to a constant factor for the contribution from tree-search.

The discussion on LFD after its definition line 141 is hard to understand: there are 2 parameters in the definition of LFD (r_1 and r_2), but the discussion seems to imply there is only 1 (a radius r). That discussion only seems to make sense after line 163 and the rewrite of the definition 2.1.

We have revised the LFD definitions in this section for clarity.

Smith-Waterman and Needleman-Wunsch are not distances (unlike Levenshtein and Hamming distances). They are dynamic programming algorithms to compute distances like the Levenshtein distance (and others depending on the exact parameters of the algorithms).

This has been fixed.

The explanation that the augmentation process (starting at line 418) preserves the topological structure of the underlying data is not very convincing. This process adds small clouds of points extending in every dimension around existing points. That is not quite the same as generating new points that belong to a small dimension manifold in the sample space.

We acknowledge this limitation of our data augmentation process; we added discussion about an improved data augmentation process as future work in our Discussion section.

Line 467-473: Why are there multiple linear search implementations used as ground truth? These different implementation are apparently used to evaluate different algorithms (ANNOY, FAISS-IVF vs CAKES). This is troubling: why would the ground truth differ based on which tool is evaluated or which language it is written in?

Do these 2 tools return the same result? If so, why use multiple tools? If not, then the evaluation is flawed.

We have fixed our approach for this and have updated the language in the paper to reflect this. We now perform linear search on the augmented datasets and save that as the ground-truth, using a format that can be loaded in Python for calculating the recall of HNSW, ANNOY and FAISS-IVF. We verified that our linear search in Rust identifies the same neighbors on the original dataset as those provided in the ANN-Benchmarks suite of datasets. Thus, our approach now uses a unified ground truth across all algorithms and both programming languages.

Minor comments

Line 421: The points are created with distance ϵ of x, not $\epsilon||x||$. r is chosen in the sphere of radius ϵ and added to x, and, as stated correctly line 424, x' is within ϵ of x.

We have addressed this, by clarifying the notation, particularly by making all vectors boldfaced.

There are many small LaTeX issues: wrong space after "i.e." and "e.g." (add a backslash), large spaces around formulas (e.g., line 342), em-dashes that look too short (e.g., line 594).

We have addressed much of this, including replacing en-dashes with em-dashes where appropriate. The space around formulas seems to be something that, if accepted, the production team could address.

The figures should not be shrank so much: the labels on the axis and the keys are not readable. Regenerate the pictures at the correct size to avoid scaling.

The figures have all been updated.

Put the keys outside of the figures, not over the data lines.

The figures have all been updated.

Many horizontal and vertical lines in Table~2 should be removed to help with readability.

The table has been broken up and former sub-tables have been enlarged.

Line 585: "we observe performs quite slowly". The sentence seems incomplete.

This has been fixed.

Many of the explicit algorithms do not serve any real purpose: they are not more readable in pseudo-code form and are just fine in explained in the text. (E.g., Algorithm 2.4).

We removed what was formerly Algorithm 2.4.

The Github repository gives very little instructions on how to compile the code, run it or reproduce the results from this manuscript.

We have updated the README to include instructions on how to compile the code, run it, and reproduce the results from the manuscript.

Review of Let them have CAKES: A Cutting-Edge Algorithm for Scalable, Efficient, and Exact Search on Big Data

SIMODS manuscript M166272

MAJOR ISSUES ==========

The CAKES manuscript proposes an entropy scaling approach to the search problems of k Nearest Neighbors (k-NN) and rho Nearest Neighbors ($\rho - NN$) in a data set X. The general approach is to construct a hierarchical data structure of clusters to assist in the searching. While there are interesting ideas in the manuscript, there are major issues both technically and expositionally, as well as some minor issues, that need to be addressed before another round of reviewing.

While there is implicitly a tree that is the key data structure, the tree itself is never named or given a notation; the same is true of the root of the tree. The authors should review the requirements for defining proper terminology and notation for a data structure and apply it consistently to their data structure.

We have addressed this by adjusting the notation.

The authors appear to use the terms similarity and distance interchangeably, though they are certainly different. In particular, a similarity is a measure (nominally, non-negative) that is (ideally) larger the closer two data points are and, in some biological contexts, has no upper bound. Moreover, the use of a distance measure alone is inadequate for the described algorithms. On line 155, there is mention of the geometric median of the points, but there may be no such median depending on the space (set) containing the points of the data set. For example, there is no obvious median for a set of biological sequences. The authors must identify the actual spaces from which data sets can be extracted for CAKES to work.

We have added a definition of the geometric median in the paper, and have clarified how we interpret it for any metric space. We have also fixed the language around similarity and distance, and have made sure to use the correct term in each context.

On line 141, there is a function defined for the local fractal dimension (LFD), but it is not expressed in functional notation. In other words, it should start with $LFD(X, r_1, r_2)$, since it is clearly a function of all three. This functional notation should propagate throughout the rest of the manuscript. Much the same is true on line 142, where it should read $B(X, q, r)$, not $B_X(q, r)$. If makes little sense to put X as a subscript in this context.

We have dropped the X since we are only dealing with one dataset.

In the algorithms, starting with Algorithm 2.1, the pseudocode style is ugly. Much better would be the pseudocode style from the fourth edition of Introduction to Algorithms, CLRS. Also, the description of criteria is as a "stopping criteria", while the logic of the code has it as a continuing criteria. Please fix the pseudocode throughout.

We evaluated the CLRS package, but did not see any aesthetic differences. We have, however, increased the font size for the algorithm environments, so that it should be more legible. We have also fixed the name for the criteria.

Section 2.2 is meant to be the definition of the search problem(s), but that is really only contained in the paragraph from lines 218 to 223. In any case, it is essential to put the problem(s) to be solved earlier, that is, first, before any description of the solutions with ideas such as clusters. This movement of text will require a bit of text refactoring at a minimum. Also, the sentence at lines 221 to 223 should be expanded in full detail into a separate paragraph. Finally, one assumes that the integer k or the number ρ is a parameter to the corresponding problem and should be mentioned as such.

We have refactored the methods section to have a better flow, and have defined several terms in their proper place. We have also removed ρ-NN search from the main body of the paper in order to focus exclusively on k-NN search. The supplement now contains a section on the improvements to ρ-NN search.

There are no theorems in the manuscript, yet there are arguments for quantities like the quality of the results or the time complexity that are given without much precision. It is better to give precise theorems or propositions for each thing to be proved, with appropriate assumptions, and to give precise proofs of same.

We have expressed claimed time complexities as precise theorems with assumptions and provided proofs.

The manuscript makes too much use of unfamiliar terminology taken from earlier papers without defining terms. For example, see line 233 and Reference [18]. See also the paragraph at lines 306 to 310. All terms used must be precisely defined in the manuscript to allow proper assessment of the manuscript.

We have addressed this by making the paper more self-contained.

The Results section is a slog to read and parts of the Discussion seem redundant. Take lines 651 to 668 as an example. Tightening up the Results and the Discussion can only help the exposition.

We attempted to address this by omitting unnecessary details, offloading observations to figure captions, and eliminating redundant sections of the discussion.

I led a bioinformatics project that used an entropy scaling approach to search a database of protein sequences. See Reference [1]. As a result, I am confused about some of the real world examples. I know that protein sequences are challenging to search for reasons such as their having widely varying lengths and there being no obvious notion of a center of a set of protein sequences. In any case, I would like the authors to reference [1] and discuss in more depth how they address the challenges of the various real world examples.

The fact that protein sequences have widely varying lengths does pose problems in terms of the potential sparsity of the manifold; it suggests that many distances would be maximized. While this makes the notion of a 'center' somewhat arbitrary and meaningless in domain terms, there would still be a computable cluster center (as demonstrated in work we just published in IEEE Big Data, largely done after the CAKES work). However, the approach in [1] is quite justified. We have added a discussion of protein sequence search as a specific example of a difficult domain, and cited [1] among others.

MINOR ISSUES ==========

"Data set" is pronounced as two words and should be written as such throughout the manuscript.

We respectfully disagree; this appears as a term of art in existing literature, and is in the Merriam-Webster Dictionary. "Dataset" is listed by Merriam-Webster as the primary form, with "data set" being the less common form (https://www.merriam-webster.com/dictionary/dataset). Unless this is an editorial policy, we would rather keep it as "dataset."

In line 22, use "rate" instead of "scale".

This has been fixed.

Throughout, the word "which" is used when "that" is the correct word. If "which" is used, then it is not defining and is preceded by a comma. Change all "which" 's to "that" 's.

We somewhat agree, but disagree that *all* instances should be changed. We have made selected changes where it seems to fit the pattern suggested by this reviewer.

On line 75, there is a Subsection 1.1, but there is no Subsection 1.2. This is not a good practice. Just refactor the text without Section 1.1 or add a Subsection 1.2.

We added Subsection 1.2 (Entropy-Scaling Search).

The use of $\forall$, $\exists$, $\Leftrightarrow$, and other similar logical operators is not appropriate in a manuscript that is not about logic. Spell out the meaning in English, and be creative.

We respectfully disagree. These operators are also standard in set theory and functional analysis, and it is in this context that we use them. They are also standard in the algorithms literature, including CLRS. Unless this is editorial policy, we would prefer to keep them. We believe they lend precision.

Near line 154, it is important to give an overview of your solution strategy and, in particular, why clustering is used first. Clustering is a heavily overloaded word. I would have chosen a more neutral term instead of "cluster".

We clearly define our clustering approach, and moreover, we build on prior work (some ours, some others) that relies on clustering. We respectfully would prefer to stay with the language we have here.

On line 155, there is the parenthetical comment about "a smaller sample of the points". What does that mean precisely and where does it occur later in the manuscript?

We clarified this by changing the language about "a smaller sample of the points" to "a random subsample of $\sqrt{|C|}$ points."

On line 156, there is the claim "so it is a real data point". What is a real data point? Please explain.

We removed the language "real data point" and clarified that the center of a cluster is the point which is its geometric median.

On line 161 and elsewhere, "i.e." is always preceded and followed by commas.

Fixed– "i.e." is now always followed by a comma.

On line 178, there is the phrase "find the point ... in the sample". Can you make the entire paragraph more precise by giving notation for the sample?

We now refer to the sample as S.

Section 2.1.1 overall is lacking in mathematical rigor.

We have updated some language in this section. This, combined with Algorithm 2.1, should provide a reasonable level of rigor.

In Section 2.1.3, the cost of assessing "criteria' is not counted in the time complexity, but it must be.

We have added a discussion of this in the corresponding complexity analysis.

In line 252, there is again mention of "real points" without any indication of the meaning of "real".

We have fixed language around "real points" and have clarified that the points are actual data points.

The fonts in Algorithms 2.2 through 2.7 are simply too small. Make them the font size of the remainder of the text.

We have made this larger to improve readability.

The numbers in Table 2 are too small. Refactor into multiple tables perhaps.

We have broken this table up into four tables (one per dataset) which we believe helps readability.

In line 641, what is "radial increase 2.3"?

We added that 2.3 is an equation.

In line 642, why is "harmonic" italicized?

This has been fixed.

The References section needs a through going over for proper formatting and capitalization. For example, check the title capitalization in [2], [4], and [8]. Check the journal name capitalization in [9] and [21], among others.

We have capitalized titles as published (confirmed for all of the references mentioned). We have corrected the journal capitalization where appropriate.

Several references have repeated URLs. Trim these up.

We believe these are now fixed.

References [2] and [3] appear to be duplicates.

This has been fixed.

Reference [39] is incomplete and inadequate.

This has been fixed.

REFERENCE ==========

*[1] Yoonjin Kim, Zhen Guo, Jeffrey A. Robertson, Benjamin Reidys, Ziyan Zhang, Lenwood S. Heath, EnTrance: Exploration of Entropy Scaling Ball Cover Search in Protein Sequences, bioRxiv https://urldefense.com/v3/**https://www.biorxiv.org/content/10.1101/2021.**q3TzzRIBjJpAmGsw1kVTB7zosA7khu71pJpSxxrEqcYu6KceYMTg_VZ_MJ8B-PoPqf$, 2021.*