Revisiting microbe-metabolite interactions: doing better than random

James T. Morton,[1,2] Daniel McDonald,[1,3] Alexander A. Aksenov,[4,5] Louis Felix Nothias,[4,5] James R. Foulds,[6] Robert A. Quinn,[7] Michelle H. Badri,[8] Tami L. Swenson,[9] Marc W. Van Goethem,[9] Trent R. Northen,[9,10] Yoshiki Vazquez-Baeza,[11,3] Mingxun Wang,[4,5] Nicholas A. Bokulich,[12,13] Aaron Watters,[14] Se Jin Song,[1,3] Richard Bonneau,[8,14,15,16] Pieter C. Dorrestein,[4,5] and Rob Knight[1,2,17,3]

[1] *Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA*

[2] *Department of Computer Science & Engineering,*
University of California, San Diego, La Jolla, CA, USA

[3] *Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA*

[4] *Collaborative Mass Spectrometry Innovation Center,*
University of California San Diego, La Jolla, CA, USA

[5] *Skaggs School of Pharmacy and Pharmaceutical Sciences,*
University of California San Diego, La Jolla, CA, USA

[6] *Department of Information Systems, University of*
Maryland Baltimore County, Baltimore, MD, USA

[7] *Department of Biochemistry and Molecular Biology,*
Michigan State University, East Lansing, MI, USA

[8] *Department of Biology, New York University, New York, 10012 NY, USA*

[9] *Environmental Genomics and Systems Biology Division,*
Lawrence Berkeley National Laboratory,
1 Cyclotron Rd, Berkeley, CA, 94720, USA

[10] *DOE Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, CA, 94598, USA*

[11] *Jacobs School of Engineering, University of California San Diego, La Jolla, CA, USA*

[12] *The Pathogen and Microbiome Institute,*
Northern Arizona University, Flagstaff, AZ, USA

[13] *Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA*

[14] *Flatiron Institute, Simons Foundation, New York, 10010 NY, USA*

[15] *Computer Science Department, Courant Institute, New York, 10012 NY, USA*

[16] *Center For Data Science, NYU, New York, NY 10008, USA*

[17] *Department of Bioengineering University of California, San Diego, La Jolla, CA, USA*

I. ABSTRACT

Recently, Quinn and Erb et al [1] made the case that when used correctly, correlation and proportionality can outperform MMvec when identifying microbe-metabolite interactions. We revisit this comparison and show that the proposed correlation and proportionality are outperformed by MMvec on real data due to their inability to deal with sparsity commonly observed in microbiome and metabolome datasets.

II. RESPONSE

As shown in the original MMvec paper [2] and Quinn and Erb et al [1], scale invariance is key for recovering sensible microbe-metabolite interactions. However, contrary to the scale invariance argument made in the preprint [1], MMvec does not normalize the joint distribution $P(\boldsymbol{u_i}, \boldsymbol{v_i})$ between microbes and metabolites (microbe abundances are represented by $\boldsymbol{u_i}$ and metabolite abundances are given by $\boldsymbol{v_i}$ for sample i). Instead, the MMvec algorithm attempts to model $P(\boldsymbol{v_i}|\boldsymbol{u_i})$ with an inverse alr transform, a known compositionally coherent transform that satisfies scale invariance [3]. This approach is more similar to a conditional version of approach B in the preprint rather than approach A [1]. Because of this, our method does not have the stated problem that microbe and metabolite abundances compete for probability mass in the normalized distribution: "the abundance of microbe 1 is limited by the abundance of microbes 2-to-M, but is in no way limited by the abundance of metabolites 1-to-N".

We relied on simulated data in [2] for the purpose of argument, and to illustrate some of the principles in the paper clearly and without the distractions typically present in real data. However, simulated data will always have limitations because of the inability to model unknown features of the real system, or because of deliberate simplifications that clarify key points in the model system. Therefore, a crucial aspect of the MMvec manuscript was to test performance both on simulations and on real data. Reuse and re-purposing requires a thorough understanding of the simulated data. Performance on real data is the ultimate test of methods, and any simulated data experiment should always be accompanied with evaluation using real data, which was not done in the commentary. Accordingly, we applied the same proportionality-based scripts described in the preprint [1], and evaluated them on one of the real datasets we used in the MMvec paper.

A major obstacle to analyzing real-world microbiome and metabolomics data is sparsity. Traditional compositional methods such as the proposed clr transform cannot automatically deal with zeros, and require imputation as a preprocessing step. This imputation adds bias, and is impractical for the sparse datasets typically encountered [4, 5]. Microbiome and untargeted metabolomics datasets are generally sparse; in large studies, such as the American Gut Project [6], the sparsity for stool samples alone is 99.946%. MMvec was designed to handle very sparse data using bootstrapping and a multinomial likelihood function, without any imputation. With the desert biocrust soils dataset (sparsity of 51%; [7]) that was used in the MMvec publication, we observe that MMvec outperforms the newly proposed linear methods dramatically (Fig. 1).

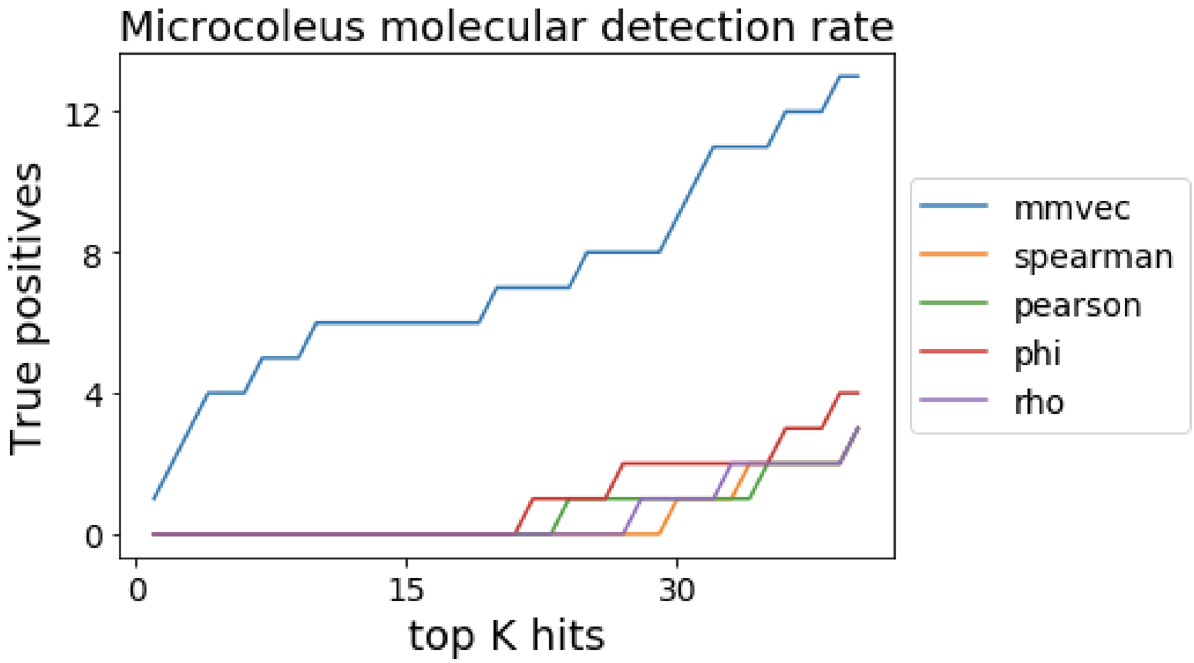

FIG. 1. A comparison of mmvec to metrics proposed by Quinn and Erb [1]. These proposed metrics include Spearman, Pearson, phi and rho applied after a clr transformation [3].

Our findings here are consistent with the key conclusions in a separate preprint [8]. Specifically, they showed that their neural network approach also outperformed linear methods. Coincidentally, their paper evaluated an IBD data set [9] similar to one of the examples in the MMvec paper, and were largely concordant with our MMvec findings. In fact, the abstract states that: "In this paper, we propose a sparse neural encoder-decoder network to predict metabolite abundances from microbe abundances. Using paired data from a cohort of inflammatory bowel disease (IBD) patients, we show that our neural encoder-decoder model outperforms linear univariate and multivariate methods in terms of accuracy, sparsity, and stability." Although their proposed underlying model achieves a different goal, our MMvec findings are consistent with those statements.

Contrary to the argument in the preprint [1] regarding the complexity of neural networks, the MMvec model [2] is not much more complex than the proposed regression techniques; it is a simple one-layer neural network without complex activation functions, which is in effect a two-stage log-bilinear regression. MMvec has even less computational complexity than the proposed proportionality metrics. Proportionality requires the estimation of $O(NM)$ parameters, one parameter for each microbe-metabolite interaction, which can easily result in the need to estimate millions of parameters for systems with only thousands of microbes and thousands of metabolites. This large number of parameters is problematic, but was not discussed in the commentary [1]. In contrast, MMvec assumes a low-dimensional latent representation requiring the estimation of $O(Nk + Mk)$ parameters. In practice, this amounts to estimating only thousands of parameters if the latent dimensionality is small (i.e. k <10).

Methods similar to MMvec have been successful at the task of learning word embeddings. Since Mikolov et al. [10], these models have been designed with an emphasis on practical methods for learning useful embeddings at scale, rather than on perfectly modeling the data distribution. Imperfect modeling assumptions (if they do indeed exist in our case) do not prevent a method's successful use for a particular task. The quote from George Box comes to mind, "all models are wrong. Some models are useful".

MMvec is only one tool in the arsenal of correlative methods. It is not perfect for every correlation type or dataset, and is not a one size fits all "magical" solution [11]. However, we have found that MMvec is a powerful discovery tool, as demonstrated by the other real datasets we evaluated in the original article, and by its wide use; the tool has already been downloaded >1200 times, and several papers based on it have already been submitted [12, 13], with many more in the works. It is critical that we provide accurate guidance to the community so that scenarios where one method works better than others are well understood. Fundamentally, the argument that simple linear methods outperform neural networks was not supported in the commentary, because only dense simulation datasets were evaluated. We appreciate the communication on the topic to the extent that it helps the community better understand the advantages of the different approaches, and prompts the community to continue to innovate in this area.

III. SOFTWARE AVAILABILITY

The analysis can be found under : https://github.com/knightlab-analyses/multiomic-cooccurrences/blob/rebuttal/ipynb/Figure3-rerun.ipynb

IV. REFERENCES

[1] Thomas Quinn and Ionas Erb. Another look at microbe–metabolite interactions: how scale invariant correlations can outperform a neural network. *bioRxiv*, page 847475, 2019.

[2] James T Morton, Alexander A Aksenov, Louis Felix Nothias, James R Foulds, Robert A Quinn, Michelle H Badri, Tami L Swenson, Marc W Van Goethem, Trent R Northen, Yoshiki Vazquez-Baeza, et al. Learning representations of microbe–metabolite interactions. *Nature methods*, pages 1–9, 2019.

[3] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[4] Josep A Martín-Fernández, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278, 2003.

[5] Justin D Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A David. Naught all zeros in sequence count data are the same. *bioRxiv*, page 477794, 2018.

[6] Daniel McDonald, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen,

et al. American gut: an open platform for citizen science microbiome research. *mSystems*, 3(3):e00031–18, 2018.

[7] Tami L Swenson, Ulas Karaoz, Joel M Swenson, Benjamin P Bowen, and Trent R Northen. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. *Nature communications*, 9(1):19, 2018.

[8] Vuong Le, Thomas P Quinn, Truyen Tran, and Svetha Venkatesh. Deep in the bowel: Highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome. *bioRxiv*, page 686394, 2019.

[9] Eric A Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J Haiser, Stefan Reinker, Tommi Vatanen, A Brantley Hall, Himel Mallick, Lauren J McIver, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature microbiology*, 4(2):293, 2019.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[11] David H Wolpert, William G Macready, et al. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

[12] Andre Mu, Glen P Carter, Lucy Li, Nicole S Isles, Alison F Vrbanac, James T Morton, David P De Souza, Vinod K Narayana, Komal Kanojia, Brunda Nijagal, et al. Microbe-metabolite associations linked to the rebounding murine gut microbiome post-colonization with vancomycin resistant enterococcus faecium. *bioRxiv*, page 849539, 2019.

[13] Jonathon L Baker, Jamie T Morton, Marcia Dinis, R Alverez, Nini C Tran, Rob Knight, and Anna Edlund. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *bioRxiv*, page 804443, 2019.