Inferring molecular complexity from mass spectrometry data using machine learning

Timothy D. Gebhard[1,2,*], Aaron C. Bell[3,*], Jian Gong[4,*], Jaden J. A. Hastings[5,*], G. Matthew Fricke[6], Nathalie Cabrol[7], Scott Sandford[8], Michael Phillips[9], Kimberley Warren-Rhodes[7], Atılım Güneş Baydin[10]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Institute for Particle Physics & Astrophysics, ETH Zurich, Switzerland
[3]Insight Edge Inc., Chiyoda-ku, Tokyo-to, Japan
[4]Massachusetts Institute of Technology, Cambridge, MA, USA
[5]XO.LABS, Los Angeles, CA, USA
[6]University of New Mexico, Albuquerque, NM, USA
[7]SETI Institute, Mountain View, CA, USA
[8]NASA Ames Research Center
[9]Johns Hopkins Applied Physics Laboratory, Laurel, MD, USA
[10]University of Oxford, United Kingdom

Authors marked with * contributed equally to this work.
Correspondence: tgebhard@tue.mpg.de.

Abstract

Molecular complexity has been proposed as a potential agnostic biosignature — in other words: a way to search for signs of life beyond Earth without relying on "life as we know it." More than one way to compute molecular complexity has been proposed, so comparing their performance in evaluating experimental data collected *in situ*, such as on board a probe or rover exploring another planet, is imperative. Here, we report the results of an attempt to deploy multiple machine learning (ML) techniques to predict molecular complexity scores directly from mass spectrometry data. Our initial results are encouraging and may provide fruitful guidance toward determining which complexity measures are best suited for use with experimental data. Beyond the search for signs of life, this approach is likewise valuable for studying the chemical composition of samples to assist decisions made by the rover or probe, and may thus contribute toward supporting the need for greater autonomy.

1 Introduction

Background: The ability to detect signs of life beyond Earth is a significant frontier in astrobiology, and a core objective in humanity's exploration missions in space [1]. However, finding evidence for extraterrestrial life requires an operational, unambiguous definition of "life", as well as a set of agreements on the kind of measurements and results as acceptable claims [2]. The burden of proof is extraordinary, and a universal agreement on what "life" is does not currently exist [3]. Alternatively, even without a precise definition of life, it has been hypothesized that complex molecules are a corollary to biologically driven activity [4, 5], and intrinsic metrics of molecular complexity (MC) can be used as an agnostic biosignature free from biases from Earth-based life form as we know it [6–8]. The key idea is that, for a suitable definition of molecular complexity, we expect that it is statistically unlikely that a large amount of any given complex molecule would be present in

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

an environmental sample due to random, abiotic processes. Consequently, if complex molecules are found at a higher proportion in a given sample, it follows that there is likely some form of biologically driven activity at work within the local environment [9, 10].

Molecular complexity (MC): Fundamentally, MC measures are numeric features intrinsic to a molecule that represent an abstraction of its structure (or formation process) while also characterizing its information content. Intuitively, one may expect MC to increase with the molecular size, the multiplicity of bonds, or the presence of heteroatoms, while it should decrease with increasing symmetry [11]. MC is usually not considered as an end in itself, but used in a relative fashion to compare molecules or to characterize chemical reactions. Various definitions of MC have been proposed in the literature (see, e.g., the introduction of Böttcher [12] for an overview), typically building on concepts from graph and information theory. In this work, we focus on the following three definitions:

1. *Bertz complexity* C_T [13]: The "first general index of molecular complexity." It combines concepts from graph and information theory and is defined as $C_T = C(\eta) + C(E)$, where $C(\eta)$ describes the bond structure and $C(E)$ the complexity due to heteroatoms. Calculating C_T is fast and scales linearly with the molecule size. We compute C_T using the `BertzCT` method from RDKit [14].
2. *Böttcher complexity* C_m [6, 12]: An information-theoric measure that is based on the information content in the microenvironments of all atoms; it is additive and simple to calculate even for large molecules. Our computation of C_m uses a freely available open source implementation [15].
3. *Molecular Assembly index (MA)* [7]: Also known as pathway complexity, the MA represents the minimum number of steps required to assemble a molecule from fundamental building blocks. MA is particularly well-suited to biosignature detection while being experimentally verifiable [8]. It is at least as hard as NP-complete to compute [16], requiring hundreds of CPU hours even for moderately sized molecules. We use a (currently non-public) implementation kindly provided by the authors of [7].

We note that, to first order and at low molecular weights, these three measures are strongly correlated; the mass of the molecule acts as a confounder constraining the maximum complexity. When regressing out the mass, however, the correlation becomes less strong, and it becomes apparent that C_T, C_M and MA each capture different aspects of the molecule.

Inferring MC in practice: The above notions of MC are theoretical in that they are defined for a given representation of a molecule (e.g., a graph). However, for MC to be valuable as a biosignature in practice, for example on board a spacecraft deployed to the outer Solar System, we need to be able to make such a determination fully *in situ* for any given sample. One prime candidate to enable this is mass spectrometry (MS; see below): Virtually all upcoming planetary exploration missions, such as the *Dragonfly* mission to Titan [17, 18], or the proposed *Europa Lander* [19] and *Enceladus Orbilander* [20], will carry mass spectrometers to analyze their targets. However, assessing the complexity of a sample in the "classical way"—inverting the mass spectrum to infer the molecule and then computing the MC using an algorithm of choice—may exceed the computational capabilities of a spacecraft. Sending all data back to Earth for analysis may also not be an option, for deep space communications are generally expensive, low-bandwidth, and have large round-trip delays. Fortunately, Marshall *et al.* [8] have already demonstrated that it is possible to infer MC *directly* from MS data by showing a clear correlation between the number of peaks in a high-resolution tandem MS of a molecule and its MA.

Contributions: In this work, we take these studies further and explore different machine learning (ML) methods and their ability to infer MC from MS. Our preliminary results are encouraging and provide evidence that the combination of MS and ML enables us to infer molecular complexity in the field. Even if we do not find extraterrestrial life, this may become a valuable tool to inform decisions about which targets to explore in greater detail, and can help fill the need for more autonomous space missions [1, 21]. Figure 1 illustrates where this capability might fit into a broader, automated analysis pipeline of a rover or probe exploring another celestial body.

2 Dataset generation

Mass spectrometry (MS): MS is an analytical technique used to determine the structure of a molecule or identify unknown molecules in a sample. Molecules are ionized and fragmented (typically using collision-induced dissociation), and the mass-to-charge ratios (m/z) of all fragments is measured. The results are collected as a series of peaks with corresponding m/z values, called a mass spectrum.

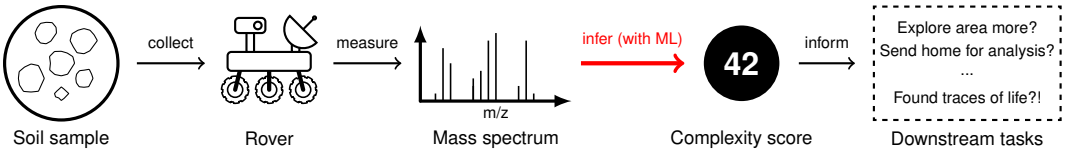
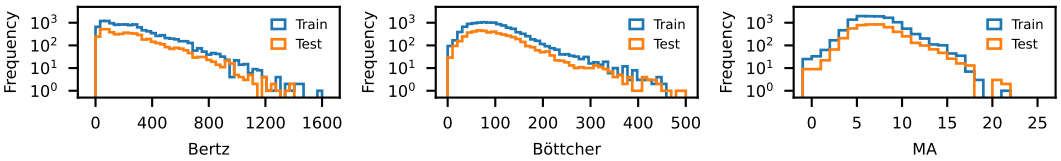


Figure 1: A schematic workflow that illustrates where our work on inferring MC from mass spectral data fits into the "bigger picture" of space exploration and the search for traces of extraterrestrial life.



(a) Bertz (b) Böttcher (c) MA

Figure 2: Distributions of the three complexity scores in our training and evaluation (test) dataset.

Dataset generation: As no ready-made dataset for our task—inferring MC from MS data—exists, we created our own. For this, we queried a public database (the NIST WebBook [22]), retrieving all molecules below 1000 Dalton for which an MS was available. These molecules were then appended with other basic chemical properties, including our three MC metrics. Empirical MS data on NIST were taken using electron ionization at a m/z-resolution of 1 (standard MS as opposed to higher resolution tandem MS that employs a variety of techniques). This is approximately comparable to the target resolution of 0.4 to 3 of the *DraMS* instrument onboard the *Dragonfly* mission [18]. Our final dataset consists of 17 021 unique molecules with associated MS, randomly split into a training set with 12 000 molecules and an evaluation set with 5021. We are showing the respective distributions of our three complexity metrics in fig. 2. The major limitation of the dataset generation was the computation of MA, taking 65 000+ CPU hours over hundreds of compute-optimized nodes on Google Cloud in parallel. Computing all Böttcher and Bertz scores together took around 20 min of CPU time. To facilitate further research, we plan to release our dataset with an upcoming version of our work.

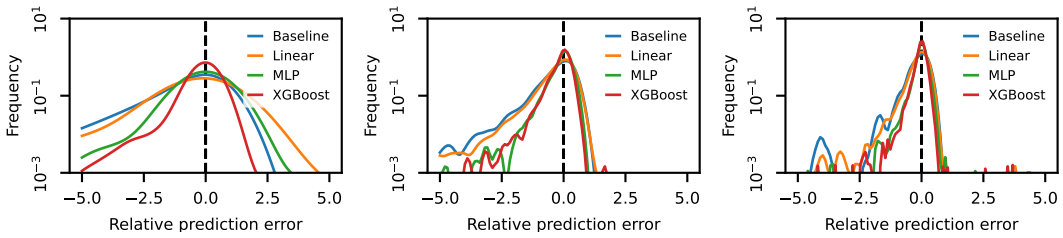
3 Experiments and results

Predicting MC from MS: We compare four different approaches, all of which take a (pre-processed) mass spectrum, and output all three complexity measures at once (i.e., multiple regression):

1. *Baseline:* A linear regression (with L_2 regularization) from the number of peaks in the MS to the complexity measure. Marshall *et al.* [8] have reported a clear correlation ($\rho = 0.89$) for the case of the MA, hence we consider this our baseline. Pre-processing the MS by removing, for example, all peaks smaller than 5% of the highest peak, did not seem to improve the performance.
2. *Linear:* A linear regression (with L_2 regularization) that receives a fixed-length representation of the mass spectra (i.e., a histogram with a 1000 equally spaced bins from 0 Da to 1000 Da) as input.
3. *MLP:* A fully-connected neural network (MLP) that operates on the same binned spectrum as the linear model. All prediction targets were normalized to $[0, 1]$ using a `MinMaxScaler`. The network has 3 `Linear` layers (with 1024 units for the "hidden" layers) and uses LeakyReLU activations, dropout ($p = 0.2$), and batch normalization. Experiments with more layers (or units) did not seem to improve the performance. We found the networks very prone to overfitting, perhaps not least due to the limited training data. Among the various solutions we tried (e.g., increasing dropout), adding random noise to the input spectra during training had the best mitigating effect here.
4. *XGBoost:* Gradient boosted trees as implemented by XGBoost [23], again using the binned mass spectrum. We use default parameters except `n_estimators=1000` and `tree_method="hist"`.

We trained every model five times using different random seeds that control the train / validation split, as well as the model initialization (where applicable). Results are reported as ensemble averages.

Our main results, in the form of relative prediction errors on the evaluation set, are summarized in table 1 and fig. 3. Unsurprisingly, all models outperform the naïve baseline (reducing the error by more



(a) Bertz (b) Böttcher (c) MA

Figure 3: Distribution of relative predictions errors on the test set (computed as $(t-p)/t$, where t is the true and p the predicted complexity value given by an ensemble average of five models corresponding to five different random seeds) for our three different complexity metrics and four model types.

Table 1: Mean relative predictions errors on the test set (computed as $|t-p|/t$, where t is the true and p the predicted complexity value given by an ensemble average of five models), as well as the 5%, 50% and 95% percentiles, for all model types and complexity scores.

	Bertz	Böttcher	MA
Baseline	1.40 [0.04, 0.47, 4.68]	0.55 [0.03, 0.32, 1.79]	0.29 [0.02, 0.21, 0.84]
Linear	1.22 [0.03, 0.34, 3.43]	0.49 [0.02, 0.28, 1.50]	0.25 [0.02, 0.18, 0.72]
MLP	0.51 [0.01, 0.15, 1.07]	0.29 [0.01, 0.18, 0.81]	0.15 [0.01, 0.10, 0.43]
XGBoost	0.42 [0.00, 0.14, 0.93]	0.27 [0.00, 0.16, 0.79]	0.14 [0.00, 0.09, 0.41]

than 50% in best case), and non-linear models perform better than the linear one. More interestingly, we find that there is a consistent trend across all models that MA is easier to predict than Böttcher complexity, which in turn is easier to predict than Bertz complexity (evidenced by respectively lower predicted errors). We speculate that this may have to do with the definition of the MA, which is, in a way, conceptually similar to the idea of mass spectrometry: The MA counts the number of steps to assemble a molecule from smaller pieces, while MS observers the patterns that emerge when a molecule is fragmented. Finally, fig. 3 shows that all models are slightly biased towards over-estimating MC. Closer inspection reveals that most of this bias is caused by molecules with (relatively) low MC values, and we hypothesize that the bias may be an effect of the fact that our MC metrics are lower-bounded by 0.

Things that did not work: We also briefly tried the following methodologically more sophisticated ideas, but found that both approaches performed worse than the MLP and XGBoost regressor above:

1. *Encode, Aggregate, Predict:* Every peak of a MS—given by a position-intensity pair (p_i, i_i)—is processed separately by an encoder E that produces a representation $z_i = E(p_i, i_i)$. All z_i of one MS are then aggregated as $z = \text{mean}(z_i)$, and a predictor network P estimates the MC from z.
2. *Transfer learning:* The core idea here was to separate the task of inferring a useful representation of a molecule from the estimation of its complexity. Besides inferring MC from MS data, we have also worked on speeding up computation of MC via surrogate models: these can predict MC from a string-based representation of a molecule. One such model consists of an LSTM that takes in the SELFIES representation [24] of a molecule and produces an embedding from which a predictor MLP P then estimates the MC. No MS data is required here, allowing us to train this model on a much larger data set (around 400k molecules). To leverage this bigger data set also for the MS to MC task, we attempted transfer learning: training an encoder network E to take in MS and pass the outputs to a (frozen) version of the pre-trained P to predict the MC.

4 Discussion and outlook

We have demonstrated how ML methods can infer three different MC measures directly from MS—not only the MA (for which we improve over the current baseline) but also the Bertz and Böttcher complexity, albeit with a higher relative prediction error. These differences in performance may provide further insights: Despite the high computational cost of MA, we observed that it is easier to determine from experimental data than C_T and C_m. This may arise from a similarity

between the computation method of MA and the fragmentation processes within a mass spectrometer and could indicate that MA is indeed particularly well suited as a potential biosignature.

Regarding the limitations of this work, we consider the size of our data set the main Achilles' heel. We believe that more data, particularly for high-MC molecules, will be essential to move forward.

From the long-term perspective, we imagine future rovers and probes equipped with MS could use ML to rapidly estimate multiple measures of MC to autonomize decisions and discoveries as they traverse the solar system. While this work targets detecting life beyond Earth, deriving MC metrics efficiently from MS could also benefit applied sciences such as chemical engineering and pharmaceutical discovery, meeting the challenges of characterizing environmental samples collected across multiple disciplines.

Acknowledgments

This work was enabled by and carried out during an eight-week research sprint as part of the *Frontier Development Lab (FDL)*, a public-private partnership between NASA, the U.S. Department of Energy, the SETI Institute, Trillium Technologies, and leaders in commercial AI, space exploration, and Earth sciences, formed with the purpose of advancing the application of machine learning, data science, and high performance computing to problems of material concern to humankind. We thank Google Cloud and the University of New Mexico Center for Advanced Research Computing for providing the compute resources critical to completing this work. GMF was funded by NASA Astrobiology NfoLD grant #80NSSC18K1140. We also thank the Cronin Group at the University of Glasgow for their collaboration, and for providing us with the code for computing MA values.

Broader impact statement

Our work here probably does not have any direct ethical or social implications. However, thinking more broadly, deriving MC metrics efficiently from experimental data could benefit a wide array of applied sciences that require the characterization of molecular mixtures, and may one day even help elucidate the question: Are we alone in this Universe? This discovery alone could change how we perceive ourselves in the vast space-time and arrive at a new understanding of the meaning of life. These fundamental scientific breakthroughs connect people and impact the society in a profound way.

References

[1] National Academies of Sciences, Engineering, and Medicine, *Origins, Worlds, and Life: A Decadal Strategy for Planetary Science and Astrobiology 2023-2032*. Washington, DC: The National Academies Press, 2022, ISBN: 978-0-309-47578-5. DOI: 10.17226/26522.

[2] M. Neveu, L. E. Hays, M. A. Voytek, M. H. New, and M. D. Schulte, "The Ladder of Life Detection," *Astrobiology*, volume 18(11): 1375–1402, 2018, ISSN: 1531-1074. DOI: 10.1089/ast.2017.1773.

[3] E. N. Trifonov, "Vocabulary of Definitions of Life Suggests a Definition," *Journal of Biomolecular Structure and Dynamics*, volume 29(2): 259–266, 2011. DOI: 10.1080/073911011010524992.

[4] D. E. Koshland, "The Seven Pillars of Life," *Science*, volume 295(5563): 2215–2216, 2002. DOI: 10.1126/science.1068489.

[5] P. L. Luisi, *Origins of Life and Evolution of the Biosphere*, volume 28(4/6): 613–622, 1998. DOI: 10.1023/a:1006517315105.

[6] T. Böttcher, "From Molecules to Life: Quantifying the Complexity of Chemical and Biological Systems in the Universe," *Journal of Molecular Evolution*, volume 86(1): 1–10, 2017. DOI: 10.1007/s00239-017-9824-6.

[7] S. M. Marshall, A. R. G. Murray, and L. Cronin, "A probabilistic framework for identifying biosignatures using Pathway Complexity," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, volume 375(2109): 20160342, 2017. DOI: 10.1098/rsta.2016.0342.

[8] S. M. Marshall, C. Mathis, E. Carrick, G. Keenan, G. J. Cooper, H. Graham, M. Craven, P. S. Gromski, D. G. Moore, S. Walker, and L. Cronin, "Identifying molecules as biosignatures with assembly theory and mass spectrometry," *Nature Communications*, volume 12(1), 2021. DOI: 10.1038/s41467-021-23258-x.

[9] S. M. Marshall, D. G. Moore, A. R. G. Murray, S. I. Walker, and L. Cronin, "Formalising the Pathways to Life Using Assembly Spaces," *Entropy*, volume 24(7): 884, 2022. DOI: 10.3390/e24070884.

[10] A. Sharma, D. Czégel, M. Lachmann, C. P. Kempes, *et al.* "Assembly Theory Explains and Quantifies the Emergence of Selection and Evolution." arXiv:2206.02279. (2022).

[11] M. Randić, X. Guo, Plavšić, and A. T. Balaban, "On the Complexity of Fullerenes and Nanotubes," in *Complexity in Chemistry, Biology, and Ecology*, New York, NY, USA: Springer, pages 1–48. DOI: 10.1007/0-387-25871-x_1.

[12] T. Böttcher, "An Additive Definition of Molecular Complexity," *Journal of Chemical Information and Modeling*, volume 56(3): 462–470, 2016. DOI: 10.1021/acs.jcim.5b00723.

[13] S. H. Bertz, "The first general index of molecular complexity," *Journal of the American Chemical Society*, volume 103(12): 3599–3601, 1981. DOI: 10.1021/ja00402a071.

[14] The RDKit Team (G. Landrum *et al.*), *RDKit: Open-source cheminformatics*. [Online]. Available: https://www.rdkit.org.

[15] Boskovic Research Group, *bottchercomplexity*, 2020. [Online]. Available: https://github.com/boskovicgroup/bottchercomplexity, Commit: a212f96.

[16] Y. Liu, C. Mathis, M. D. Bajczyk, S. M. Marshall, L. Wilbraham, *et al.*, "Exploring and mapping chemical space with molecular assembly trees," *Science Advances*, volume 7(39), 2021. DOI: 10.1126/sciadv.abj2465.

[17] R. Lorenz, E. Turtle, J. Barnes, M. Trainer, D. Adams, *et al.*, "Dragonfly: A rotorcraft lander concept for scientific exploration at Titan," *Johns Hopkins APL Technical Digest (Applied Physics Laboratory)*, volume 34(3): 374–387, 2018, ISSN: 0270-5214.

[18] A. Grubisic, M. G. Trainer, X. Li, W. B. Brinckerhoff, F. H. van Amerom, *et al.*, "Laser Desorption Mass Spectrometry at Saturn's moon Titan," *International Journal of Mass Spectrometry*, volume 470: 116707, 2021. DOI: 10.1016/j.ijms.2021.116707.

[19] K. P. Hand, C. B. Phillips, A. Murray, J. B. Garvin, E. H. Maize, *et al.*, "Science Goals and Mission Architecture of the Europa Lander Mission Concept," *The Planetary Science Journal*, volume 3(1): 22, 2022. DOI: 10.3847/psj/ac4493.

[20] S. M. MacKenzie, M. Neveu, A. F. Davila, J. I. Lunine, K. L. Craft, *et al.*, "The Enceladus Orbilander Mission Concept: Balancing Return and Resources in the Search for Life," *The Planetary Science Journal*, volume 2(2): 77, 2021. DOI: 10.3847/psj/abe4da.

[21] B. P. Theiling, L. Chou, V. D. Poian, M. Battler, K. Raimalwala, *et al.*, "Science Autonomy for Ocean Worlds Astrobiology: A Perspective," *Astrobiology*, volume 22(8): 901–913, 2022. DOI: 10.1089/ast.2021.0062.

[22] W. E. Wallace, "NIST Mass Spectrometry Data Center," in *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, P. Linstrom and W. Mallard, Eds., Gaithersburg MD, 20899, USA: National Institute of Standards and Technology, 2022. DOI: 10.18434/t4d303.

[23] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016. DOI: 10/gdp84q.

[24] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, volume 1(4): 045024, 2020. DOI: 10.1088/2632-2153/aba947.

NeurIPS paper checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Yes
   (b) Have you read the ethics review guidelines and ensured that your paper conforms to them?
   Yes
   (c) Did you discuss any potential negative societal impacts of your work?
   Yes
   (d) Did you describe the limitations of your work?
   Yes
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results?
   Not applicable
   (b) Did you include complete proofs of all theoretical results?
   Not applicable
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   No
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
   Yes
   (d) Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
   Yes
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators?
   Yes.
   (b) Did you mention the license of the assets?
   No.
   (c) Did you include any new assets either in the supplemental material or as a URL?
   No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
   Not applicable
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
   Not applicable
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable?
   Not applicable
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable?
   Not applicable
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?
   Not applicable