nature portfolio

Peer Review File

Engineering microbial consortia for distributed signal processing

Corresponding Author: Professor Lingchong You

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Communications.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

A version of this paper was originally rejected for publication by Nature Communications, however that decision was reconsidered after appeal by the authors.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This reviewer would like to thank the authors for providing a detailed response to their questions and concerns. However, overall, the manuscript has not been significantly altered except for textual clarifications, which does not resolve the central concerns. I will specifically address these concerns in relation to the authors' response. I will reiterate my recommendation that this article is more suitable for a specialty journal, particularly in comparison to prior work in this area.

Specific Comments

1. The authors use hyperbolic language to describe the sensing capabilities of their microbial consortia. The authors write that "our contribution is a generalizable pipeline for converting rich, multiplexed biological responses into quantitative input concentrations." when the consortia contained either 2 or 3 engineered E. coli strains, each using a previously developed cell sensor that expresses a different fluorescent protein reporter. By relying on a fluorescent protein output, the authors are using an approach that would maximally cap the number of cell sensors that could be combined in a consortia to a relatively small number. So even if the authors would like to argue that "3 cell sensors are enough" for a proof-of-principle, the more important point is that this approach can not scale that much beyond 3 cell sensors (e.g. there are 4 or 5 fluorescent reporters with distinct emission spectra). The authors suggest that other output modalities could be used (e.g. barcoding and NGS), but they do not carry out such experiments. The analysis of other output modalities is distinct to analyzing the dynamics of fluorescence levels.

2. The prior state-of-the-art in this area has shown that it is possible to engineer a single E. coli strain to detect 12 molecules using 12 cell sensors [Meyer, A.J., Segall-Shapiro, T.H., Glassey, E., Zhang, J. and Voigt, C.A., 2019. Escherichia coli "Marionette" strains with 12 highly optimized small-molecule sensors. Nature chemical biology, 15(2), pp.196-204.]. The authors might argue that a consortia approach is easier to implement, but they must prove that claim by actually achieving a nominally similar outcome using their consortia approach (e.g. combining 12 engineered E. coli strains utilizing 12 cell sensors). You can not claim that a consortia approach is superior while delivering a much less impactful outcome.

3. The authors have not carried out a real-world demonstration of their sensing approach. Even with the textual modifications, the authors are adding small molecules (vanillin and aTC) to a solution that partly contains a sample from a sink P-trap pipe. Why is detecting vanillin and aTC so important to hospital management (keeping in mind that tetracycline is not a commonly proscribed antibiotic; instead, doxycycline is the preferred antibiotic)? Are there any other chemical components in the P-trap sample that actually interfere with consortia growth/response or is it effectively a "placebo" solution that has no effect (except to somehow mirror a real-world application?). In order for the demonstration to be an effective real-world demonstration, it should in some way mirror the real-world usage.

4. Overall, using phenomenological ODEs to fit dynamic data and then using that parameterization to train a VAE and MLP

model might work well for the relatively simple systems that the authors have developed. But the authors have not proven that this approach will scale towards what would actually be a powerfully useful consortia, e.g. a consortia combines 100+ engineered E. coli strains with different cell sensors. The authors use an approach that is 10+ years old, which would be fine to achieve a desired outcome. But the authors can not claim that this approach is somehow novel or compelling on its own.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors addressed my comments. Thank you. I recommend that some important "assumptions" are added to the discussion so one is aware of the limitations of the work.

In particular, based on the response to one of my questions, the author indicate that the whole approach can work only if the community, composition-wise, is stable over the period of observation, and that each strain gives as a whole a sufficiently large response. This should be clearly stated in the discussion. The authors should comment in the discussion section in what use-cases it is possible to assume that the composition will remain stable, or how they plan to test it so to avoid misinterpreting the results (if some strains go extinct for example).

Also, according to the response to why the carrying capacity is set to 1.5, if it is not set to this value, the estimates go bad. This may be due to a lack of identifiability or to the optimization getting stuck in local minima. This sensitivity to parameter initialization/constraining should me stated so that if someone will use this approach will know about it.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The reviewers have addressed all of my comments. I congratulate the authors on their excellent and exciting work.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Reviewer 1: The authors use hyperbolic language to describe the sensing capabilities of their microbial consortia. The authors write that “our contribution is a generalizable pipeline for converting rich, multiplexed biological responses into quantitative input concentrations.” when the consortia contained either 2 or 3 engineered E. coli strains, each using a previously developed cell sensor that expresses a different fluorescent protein reporter. By relying on a fluorescent protein output, the authors are using an approach that would maximally cap the number of cell sensors that could be combined in a consortia to a relatively small number. So even if the authors would like to argue that “3 cell sensors are enough” for a proof-of-principle, the more important point is that this approach can not scale that much beyond 3 cell sensors (e.g. there are 4 or 5 fluorescent reporters with distinct emission spectra). The authors suggest that other output modalities could be used (e.g. barcoding and NGS), but they do not carry out such experiments. The analysis of other output modalities is distinct to analyzing the dynamics of fluorescence levels.

My take is that this is a fair criticism, definitely it challenges scalability and hence of practical applicability of the approach. I think that if the authors point this out as a limitation of the current study and upfront (best if in the abstract) say that this is just a proof of concept and in no way wants to be a concrete application, then it should be ok since the conceptual advancement that they propose is still there. However, as I pointed out in my comments, the authors have been inflating a bit what they have been doing, putting assumptions and limitations on the side or under the rug. Instead, they may want to spell out the limitations upfront, perhaps even in a specifically dedicated section in the discussion and state in the abstract that this is a proof of concept study.

Regarging the author’s response: “The relevant measure of scalability here is not simply the number of conventional sensors implemented, but the number of input signals that can be resolved from distinguishable output trajectories”. This is I think an OK response but it could have been better. In fact, even if they are correct in saying this, in their approach they do not provide any systematic way to check/determine what the number of input signals is that can be identified by the chosen

outputs. Indeed, this is a standard problem of system identification, which I had also alluded to in my comments, for which they have no general answer. It is mostly “try and see if it works”, and it can work, of course, but it is not clear when and if it does before building the system.

2. Reviewer 1: The prior state-of-the-art in this area has shown that it is possible to engineer a single E. coli strain to detect 12 molecules using 12 cell sensors [Meyer, A.J., Segall-Shapiro, T.H., Glassey, E., Zhang, J. and Voigt, C.A., 2019. Escherichia coli “Marionette” strains with 12 highly optimized small-molecule sensors. Nature chemical biology, 15(2), pp.196-204.]. The authors might argue that a consortia approach is easier to implement, but they must prove that claim by actually achieving a nominally similar outcome using their consortia approach (e.g. combining 12 engineered E. coli strains utilizing 12 cell sensors). You can not claim that a consortia approach is superior while delivering a much less impactful outcome.

Here, I disagree with the reviewer. In fact, the marionette strain can take up to 12 small molecule inputs concurrently, but it has only one output fluorescent reporter. So, it can detect only one of the 12 inputs at the time. In the original marionette paper (Meyer et al, Nat Chem Bio 2019), the authors only demonstrated one fluorescent reporter. Specifically, they had 12 different versions of the output plasmid, all with the same reporter (YFP) but each with a different promoter (one for each of the 12 regulators). They could only demonstrate one reporter at the time. So, the marionette strain is not a whole cell biosensor because it has one output only (YFP). Indeed, we used ourselves the marionette strain by putting more reporters concurrently to have multiplexed output, but issues of ribosome sharing arise, so the outputs interfere with each other. Distributing each sensor in a different cell, as this paper is proposing, can overcome this issue since sensor outputs are not competing for translational resources, but there are other issues. The major one is to ensure that all the strains in the population persist. This, to me is the major drawback of this approach, which will make it potentially poorly applicable. The authors addressed this comment of mine in the discussion stating clearly that they assume the population is stable as this is needed for the approach to work.

3. The authors have not carried out a real-world demonstration of their sensing approach. Even with the textual modifications, the authors are adding small molecules (vanillin and aTC) to a solution that partly contains a sample from a sink P-trap pipe. Why is detecting vanillin and aTC so important to hospital management (keeping in mind that tetracycline is not a commonly proscribed antibiotic; instead, doxycycline is the preferred antibiotic)? Are there any other chemical components in the P-trap sample that actually interfere with consortia growth/response or is it effectively a “placebo” solution that has no effect (except to somehow mirror a real-world application?). In order for the demonstration to be an effective real-world demonstration, it should in some way mirror the real-world usage.

This is similar to the comment that I initially made. The authors revised the manuscript by removing any claim that they test it on a real application, which was my request. However, in the response to this reviewer, they are now claiming that this experiment is a “complex-matrix robustness test”. I do not completely agree with statement: to assess robustness of their system, they should probably place it in many different environmental conditions, not just in a specific one. So, I do not think they can claim to have shown robustness of their system. They have shown that it works similarly in two environmental conditions: the lab condition and the Ptrap condition. This suggests that it may perform well in Ptrap applications, although something should be commented on how much the composition of the Ptrap elements changes over time and depending on the sampling location. So, this point warrants a discussion in the discussion section.

4. Overall, using phenomenological ODEs to fit dynamic data and then using that parameterization to train a VAE and MLP model might work well for the relatively simple systems that the authors have developed. But the authors have not proven that this approach will scale towards what would actually be a powerfully useful consortia, e.g. a consortia combines 100+ engineered E. coli strains with different cell sensors. The authors use an approach that is 10+ years old, which would be fine to achieve a desired outcome. But the authors can not claim that this approach is somehow novel or compelling on its own.

I think the authors have proposed a proof of concept approach, which in terms of learning and data generation from the model could theoretically scale to higher dimensional systems. Of course, the curse of dimensionality will hit since they will need to generate a lot more data for learning with 100+ engineered strains. This could be pointed out in the discussion, along with some explanation of the size of the consortium required to sense a practically relevant number of molecules.

(Remarks on code availability)

We thank the reviewers for their careful evaluation of the revised manuscript. We are encouraged that Reviewers 2 and 3 found that the revisions addressed their concerns, and we appreciate Reviewer 1's clarification of the remaining reservations. In response, we have further revised the manuscript to improve rigor and clarity in terms of the central contribution: *a generalizable inference framework for recovering multiplexed quantitative inputs from biological responses that may be non-orthogonal, cross-reactive, or indirect.*

Point-by-point responses

Reviewer #1 (Remarks to the Author)

This reviewer would like to thank the authors for providing a detailed response to their questions and concerns. However, overall, the manuscript has not been significantly altered except for textual clarifications, which does not resolve the central concerns. I will specifically address these concerns in relation to the authors' response. I will reiterate my recommendation that this article is more suitable for a specialty journal, particularly in comparison to prior work in this area.

We thank the reviewer for clarifying the remaining concerns.

We respectfully disagree with the statement that the manuscript was not significantly altered except for textual clarification. In the revised submission, we added substantive analyses and controls in response to the technical concerns raised during review. These include a new benchmarking section on the contextual limits of transfer-function-based inference in the main text, supported by single-time-point transfer-function inversion analyses across datasets (Supplementary Figs. 12 and 13); a new robustness section in the main text, supported by tests of alternative kinetic formulations, parameter non-identifiability, and a conservative split-before-fitting protocol (Supplementary Fig. 7a–c); an additional analysis probing the learned latent space (Supplementary Fig. 8); expanded statistical treatment of crosstalk with replicate-level uncertainty summarized in Supplementary Table 3; and an additional demonstration showing multiplexed decoding of five inputs from two strains with three readouts, without modifying the core pipeline (Supplementary Fig. 17). These additions did not change the central conceptual focus of the manuscript, but they substantially strengthened its rigor, scope, and evidentiary basis.

We agree that the key remaining issues are about scope and positioning. In the revised manuscript, we have further clarified what is directly demonstrated and made the intended contribution more explicit: a generalizable inference framework for recovering multiplexed quantitative inputs from biological responses that may be non-orthogonal, cross-reactive, or indirect. This is distinct from claiming the most highly optimized sensing hardware or the largest demonstrated sensor panel.

Specific Comments

1. The authors use hyperbolic language to describe the sensing capabilities of their microbial consortia. The authors write that "our contribution is a generalizable pipeline for converting rich, multiplexed biological responses into quantitative input concentrations." when the consortia contained either 2 or 3 engineered E. coli strains, each using a previously developed cell sensor that expresses a different fluorescent protein reporter. By relying on a fluorescent protein output, the authors are using an approach that would maximally cap the number of cell sensors that could be combined in a consortia to a relatively small number. So even if the authors would like to argue that "3 cell sensors are enough" for a proof-of-principle, the more important point is that this approach can not scale that much beyond 3 cell sensors (e.g. there are 4 or 5 fluorescent reporters with distinct emission spectra). The authors suggest that other output modalities could

be used (e.g. barcoding and NGS), but they do not carry out such experiments. The analysis of other output modalities is distinct to analyzing the dynamics of fluorescence levels.

We appreciate the reviewer's concern about whether fluorescence-based demonstrations can support broader multiplexed inference. To this end, our manuscript includes a direct experimental demonstration that two strains with three readouts (OD plus two fluorescence channels) can resolve five distinct input chemicals with $R^2 > 0.7$ for all five inputs (**Supplementary Fig. 17**). This result shows that, in our framework, the number of resolvable inputs exceeds the number of fluorescence channels when time-resolved dynamics are used. We have revised the manuscript to make this point more explicit. The relevant measure of scalability here is not simply the number of conventional sensors implemented, but the number of input signals that can be resolved from distinguishable output trajectories. We agree that broader scaling using alternative readout modalities remains a future direction, and we do not present it as an experimental claim of the current study.

2. The prior state-of-the-art in this area has shown that it is possible to engineer a single E. coli strain to detect 12 molecules using 12 cell sensors [Meyer, A.J., Segall-Shapiro, T.H., Glassey, E., Zhang, J. and Voigt, C.A., 2019. Escherichia coli "Marionette" strains with 12 highly optimized small-molecule sensors. Nature chemical biology, 15(2), pp.196-204.]. The authors might argue that a consortia approach is easier to implement, but they must prove that claim by actually achieving a nominally similar outcome using their consortia approach (e.g. combining 12 engineered E. coli strains utilizing 12 cell sensors). You can not claim that a consortia approach is superior while delivering a much less impactful outcome.

We agree that Meyer et al. is a landmark study and that our manuscript should not be interpreted as surpassing it on the axis of sensor engineering. We have revised the manuscript to reflect this distinction. Meyer et al. addresses front-end optimization of sensor orthogonality and performance. Our study addresses a different problem: *how to recover multiplexed quantitative information from responses that are not orthogonal, including cases with strong crosstalk and indirect community-level coupling, without requiring comparable re-engineering of the sensing components*.

We did not claim or intend to claim that our framework is "superior" to other approaches. Rather, our claim is that distributing sensing functions across strains provides a modular route for assembling multiplexed systems, and that the computational framework enables quantitative decoding when the resulting responses are cross-reactive, indirect, or otherwise complex. We have revised the manuscript text to make this complementarity explicit.

3. The authors have not carried out a real-world demonstration of their sensing approach. Even with the textual modifications, the authors are adding small molecules (vanillin and aTC) to a solution that partly contains a sample from a sink P-trap pipe. Why is detecting vanillin and aTC so important to hospital management (keeping in mind that tetracycline is not a commonly proscribed antibiotic; instead, doxycycline is the preferred antibiotic)? Are there any other chemical components in the P-trap sample that actually interfere with consortia growth/response or is it effectively a "placebo" solution that has no effect (except to somehow mirror a real-world application?). In order for the demonstration to be an effective real-world demonstration, it should in some way mirror the real-world usage.

We agree that this experiment should not be framed or interpreted as a turnkey real-world deployment. We have revised the manuscript to state explicitly that the sink-water study is a complex-matrix robustness test, not an *in situ* application demonstration. Its purpose is to test whether inference remains accurate when the background shifts from defined laboratory media to a realistic, uncontrolled environmental matrix.

We have also clarified that vanillic acid and aTc were chosen because well-characterized sensor strains were available, thus providing a tractable test case without requiring additional

component engineering (which is not our central goal). The significance of the experiment is that hospital sink P-trap samples are compositionally complex and contain diverse unknown components and potential interferents not present in standard laboratory media. The result therefore supports robustness to matrix complexity, which is the appropriate practical benchmark for a framework study at this stage.

4. Overall, using phenomenological ODEs to fit dynamic data and then using that parameterization to train a VAE and MLP model might work well for the relatively simple systems that the authors have developed. But the authors have not proven that this approach will scale towards what would actually be a powerfully useful consortia, e.g. a consortia combines 100+ engineered E. coli strains with different cell sensors. The authors use an approach that is 10+ years old, which would be fine to achieve a desired outcome. But the authors can not claim that this approach is somehow novel or compelling on its own.

We agree that the novelty does not lie in any one component of the pipeline in isolation. We have revised the manuscript to clarify that the contribution is the integration of structured ODE-based data augmentation, latent representation learning, and downstream inference into a framework that resolves multiplexed inputs from cross-reactive and indirect biological responses. The ODE model is not used as the inference model at prediction time, but as a structured generator to augment sparse experimental datasets.

We also agree that our study does not demonstrate performance for arbitrarily large consortia. Our demonstrated claim is that the framework works across multiple qualitatively distinct settings: low-crosstalk sensor communities, high-crosstalk sensor communities, indirect antibiotic-response communities, and complex sample matrices. We believe this establishes the conceptual and practical value of the framework without implying that large-scale deployment has already been demonstrated.

Reviewer #2 (Remarks to the Author)

The authors addressed my comments. Thank you. I recommend that some important "assumptions" are added to the discussion so one is aware of the limitations of the work.

In particular, based on the response to one of my questions, the author indicate that the whole approach can work only if the community, composition-wise, is stable over the period of observation, and that each strain gives as a whole a sufficiently large response. This should be clearly stated in the discussion. The authors should comment in the discussion section in what use-cases it is possible to assume that the composition will remain stable, or how they plan to test it so to avoid misinterpreting the results (if some strains go extinct for example).

We agree and have stated this assumption explicitly in the Discussion. Our framework does not require perfect long-term ecological stability. Rather, it requires that during the assay window the community remains sufficiently stable to produce reproducible, information-rich readouts, and that each relevant strain remains present at sufficient abundance to contribute a detectable response. We have clarified that this assumption is most reasonable in short-term assay settings, batch measurements, or other contexts where composition drift is limited over the observation period. We have also noted that, in applications where strain dropout or strong composition drift is possible, these effects should be monitored experimentally as part of assay validation.

Also, according to the response to why the carrying capacity is set to 1.5, if it is not set to this value, the estimates go bad. This may be due to a lack of identifiability or to the optimization getting stuck in local minima. This sensitivity to parameter initialization/constraining should me stated so that if someone will use this approach will know about it.

We agree and have added this point explicitly. In our implementation, the carrying capacity was fixed as a stabilizing constraint to prevent unrealistic trajectories and erratic optimization behavior. This is consistent with the phenomenological role of the ODE model in our framework: it is used to generate realistic synthetic trajectories for augmentation, not to uniquely identify biological parameters. We have therefore clarified that parameter estimation can be sensitive to initialization and constraints, and that the practical objective is robust trajectory generation for downstream inference rather than unique mechanistic parameter recovery.

We thank Reviewer #2 for the careful and constructive assessment of our revised manuscript. We are encouraged that the reviewer did not identify major technical concerns and recognized the conceptual advancement of the work. In response to the remaining comments, we have further revised the manuscript to present the study explicitly as a proof of concept; provide a dedicated discussion of requirements, assumptions, and limitations; define a system-specific operational procedure for evaluating input resolvability; strengthen the discussion of consortium persistence; narrow the interpretation of the hospital sink-water experiment; and acknowledge the increasing calibration and data requirements associated with scaling. The corresponding changes are marked in blue in the revised manuscript.

Point-by-point responses

Reviewer #2 (Remarks to the Author)

Specific Comments

1. Reviewer 1: The authors use hyperbolic language to describe the sensing capabilities of their microbial consortia. The authors write that "our contribution is a generalizable pipeline for converting rich, multiplexed biological responses into quantitative input concentrations." when the consortia contained either 2 or 3 engineered E. coli strains, each using a previously developed cell sensor that expresses a different fluorescent protein reporter. By relying on a fluorescent protein output, the authors are using an approach that would maximally cap the number of cell sensors that could be combined in a consortia to a relatively small number. So even if the authors would like to argue that "3 cell sensors are enough" for a proof-of-principle, the more important point is that this approach can not scale that much beyond 3 cell sensors (e.g. there are 4 or 5 fluorescent reporters with distinct emission spectra). The authors suggest that other output modalities could be used (e.g. barcoding and NGS), but they do not carry out such experiments. The analysis of other output modalities is distinct to analyzing the dynamics of fluorescence levels.

My take is that this is a fair criticism, definitely it challenges scalability and hence of practical applicability of the approach. I think that if the authors point this out as a limitation of the current study and upfront (best if in the abstract) say that this is just a proof of concept and in no way wants to be a concrete application, then it should be ok since the conceptual advancement that they propose is still there. However, as I pointed out in my comments, the authors have been inflating a bit what they have been doing, putting assumptions and limitations on the side or under the rug. Instead, they may want to spell out the limitations upfront, perhaps even in a specifically dedicated section in the discussion and state in the abstract that this is a proof of concept study.

Regarging the author's response: "The relevant measure of scalability here is not simply the number of conventional sensors implemented, but the number of input signals that can be resolved from distinguishable output trajectories". This is I think an OK response but it could have been better. In fact, even if they are correct in saying this, in their approach they do not provide any systematic way to check/determine what the number of input signals is that can be identified by the chosen outputs. Indeed, this is a standard problem of system identification, which I had also alluded to in my comments, for which they have no general answer. It is mostly "try and see if it works", and it can work, of course, but it is not clear when and if it does before building the system.

Reviewer 2's observation identifies the issue precisely. Whether a chosen set of biological outputs can resolve a given panel of inputs is fundamentally a question of system identifiability. We fully agree with the reviewer's broader point that this is a standard and, in general, open problem in system identification: for nonlinear, cross-reactive systems of this kind, no universal a priori criterion returns the number of resolvable inputs before the system is characterized.

We now state this constraint directly and provide a systematic operational procedure, supported by empirical evidence that brackets where resolvability succeeds and where it fails. We have made this framing explicit in three places.

First, we have revised the Abstract and Introduction to describe the work as a proof-of-concept framework rather than a concrete application or a demonstration of unrestricted scalability, and we have removed hyperbolic framing throughout the manuscript.

Second, we have added a dedicated Discussion subsection, *"Requirements, assumptions, and limitations of the inference framework,"* which consolidates the conditions required for successful inference and makes explicit that the workflow can be applied broadly while high-performance inference is not guaranteed for an arbitrary sensing system.

Third, the revised text in the added subsection gives a concrete, system-specific decision procedure for each implementation and substantiates it with two contrasting cases: recovery of five inputs from three readouts (Supplementary Fig. 17) and the failure to resolve TAZ concentrations in the HSGBlaM community (Supplementary Fig. 16H).

> *"Whether a chosen set of biological outputs can resolve a given panel of inputs is a question of system identifiability: whether the input concentrations can be recovered from the measured output trajectories. We determine this resolvability operationally for each implementation. For a given system, the intended input range and the prediction error acceptable for the application are defined, the system is calibrated on representative data, and performance is evaluated on independent data spanning that range; the input panel is supported when the predefined error criterion is met across validation, and failure indicates that the chosen outputs, temporal sampling, or calibration dataset are insufficient for that panel. This assessment is system-specific because identifiability here is a property of the calibrated input-output map and the measurement noise, and it depends on response dynamic ranges, cross-reactivity, input ranges, temporal sampling, and biological and measurement variability, which are not known before the system is characterized. The number of resolvable inputs is not fixed by the number of readout channels. Time-resolved trajectories can resolve more inputs than channels, as illustrated by our example of inferring five inputs from three readouts (Supplementary Fig. 17). In contrast, limited dynamic range and responsiveness of readouts to inputs will compromise the resolvability, as reflected by the poor model performance in inferring TAZ concentrations by the HSGBlaM community (Supplementary Fig. 16H). As a result, no general a priori criterion can return the maximum number of resolvable inputs from a chosen set of outputs in advance for nonlinear, cross-reactive systems of this kind. Our workflow (Fig. 1) can therefore be applied broadly, while successful high-performance inference must be established for each implementation rather than guaranteed a priori."*

2. Reviewer 1: The prior state-of-the-art in this area has shown that it is possible to engineer a single E. coli strain to detect 12 molecules using 12 cell sensors [Meyer, A.J., Segall-Shapiro, T.H., Glassey, E., Zhang, J. and Voigt, C.A., 2019. Escherichia coli "Marionette" strains with 12 highly optimized small-molecule sensors. Nature chemical biology, 15(2), pp.196-204.]. The authors might argue that a consortia approach is easier to implement, but they must prove that claim by actually achieving a nominally similar outcome using their consortia approach (e.g. combining 12 engineered E. coli strains utilizing 12 cell sensors). You can not claim that a consortia approach is superior while delivering a much less impactful outcome.

Here, I disagree with the reviewer. In fact, the marionette strain can take up to 12 small molecule inputs concurrently, but it has only one output fluorescent reporter. So, it can detect only one of the 12 inputs at the time. In the original marionette paper (Meyer et al, Nat Chem Bio 2019), the authors only demonstrated one fluorescent reporter. Specifically, they had 12 different versions

of the output plasmid, all with the same reporter (YFP) but each with a different promoter (one for each of the 12 regulators). They could only demonstrate one reporter at the time. So, the marionette strain is not a whole cell biosensor because it has one output only (YFP). Indeed, we used ourselves the marionette strain by putting more reporters concurrently to have multiplexed output, but issues of ribosome sharing arise, so the outputs interfere with each other. Distributing each sensor in a different cell, as this paper is proposing, can overcome this issue since sensor outputs are not competing for translational resources, but there are other issues. The major one is to ensure that all the strains in the population persist. This, to me is the major drawback of this approach, which will make it potentially poorly applicable. The authors addressed this comment of mine in the discussion stating clearly that they assume the population is stable as this is needed for the approach to work.

We thank the reviewer for this clarification and agree with this assessment. As the reviewer notes, the Marionette study established 12 optimized small-molecule sensing systems, but the original demonstration used a single fluorescent reporter at a time and therefore did not demonstrate the simultaneous quantification of 12 inputs from 12 distinguishable outputs. Requiring our consortium to contain 12 sensor strains would therefore not constitute a like-for-like benchmark for the capability examined here, which is the simultaneous inference of multiple input concentrations from shared time-resolved biological responses.

We also appreciate the reviewer's firsthand observation that implementing multiple reporter channels within a single strain can introduce crosstalk through competition for shared cellular resources. Distributing sensing functions among strains can reduce such intracellular interactions and provide a modular route for multiplexed sensing. At the same time, as the reviewer emphasizes, this strategy introduces a different and important constraint: the populations contributing informative responses must persist over the assay period.

We have therefore strengthened the Discussion to state explicitly that the framework requires the community to remain sufficiently stable over the measurement window for all relevant strains to contribute reproducible, information-rich readouts. The dedicated limitations subsection further states:

> *"Consortium-based inference further requires that populations contributing informative responses remain sufficiently stable over the assay period. In our engineered systems, we promoted this stability by choosing strains with similar growth rates and by inoculating them at equal fractions. For systems in which substantial compositional drift or strain loss is plausible, strain abundance or strain-specific output baselines should be monitored during calibration and validation, and measurements outside the calibrated range of community states should not be interpreted using the trained model."*

This revision makes clear that distributed and single-strain sensing strategies involve different tradeoffs. The consortium approach may reduce intracellular resource competition, but its successful application requires functional persistence of the informative populations during the assay.

3. The authors have not carried out a real-world demonstration of their sensing approach. Even with the textual modifications, the authors are adding small molecules (vanillin and aTC) to a solution that partly contains a sample from a sink P-trap pipe. Why is detecting vanillin and aTC so important to hospital management (keeping in mind that tetracycline is not a commonly proscribed antibiotic; instead, doxycycline is the preferred antibiotic)? Are there any other chemical components in the P-trap sample that actually interfere with consortia growth/response or is it effectively a "placebo" solution that has no effect (except to somehow mirror a real-world application?). In order for the demonstration to be an effective real-world demonstration, it should in some way mirror the real-world usage.

This is similar to the comment that I initially made. The authors revised the manuscript by removing any claim that they test it on a real application, which was my request. However, in the response to this reviewer, they are now claiming that this experiment is a "complex-matrix robustness test". I do not completely agree with statement: to assess robustness of their system, they should probably place it in many different environmental conditions, not just in a specific one. So, I do not think they can claim to have shown robustness of their system. They have shown that it works similarly in two environmental conditions: the lab condition and the Ptrap condition. This suggests that it may perform well in Ptrap applications, although something should be commented on how much the composition of the Ptrap elements changes over time and depending on the sampling location. So, this point warrants a discussion in the discussion section.

We agree. We have removed the characterization of the sink-water experiment as a "complex-matrix robustness test" and no longer present it as evidence of robustness across environmental conditions or as an in situ application demonstration. The Results now frame the experiment as a calibrated feasibility test in one pooled hospital sink-water matrix. Specifically, we state that we asked whether the framework could infer spiked analytes after calibration in hospital-derived sink-water samples, rather than attempting an in situ deployment. We further conclude:

> *"This experiment establishes feasibility in a sink-water background rather than robustness across environmental matrices or an in situ deployment."*

We have also addressed the reviewer's concern about temporal and spatial variation directly in the limitations subsection:

> *"Finally, the hospital sink-water experiment evaluates one pooled P-trap-derived matrix and does not establish robustness across different sites, sampling times, or environmental matrices."*

The manuscript therefore limits the conclusion to successful inference after matrix-specific calibration in the particular pooled sample tested. It does not infer broader performance across P-traps, sampling locations, time points, or hospital-monitoring applications.

4. Overall, using phenomenological ODEs to fit dynamic data and then using that parameterization to train a VAE and MLP model might work well for the relatively simple systems that the authors have developed. But the authors have not proven that this approach will scale towards what would actually be a powerfully useful consortia, e.g. a consortia combines 100+ engineered E. coli strains with different cell sensors. The authors use an approach that is 10+ years old, which would be fine to achieve a desired outcome. But the authors can not claim that this approach is somehow novel or compelling on its own.

I think the authors have proposed a proof of concept approach, which in terms of learning and data generation from the model could theoretically scale to higher dimensional systems. Of course, the curse of dimensionality will hit since they will need to generate a lot more data for learning with 100+ engineered strains. This could be pointed out in the discussion, along with some explanation of the size of the consortium required to sense a practically relevant number of molecules.

We agree with the reviewer that this study is a proof of concept and does not demonstrate performance for a consortium containing 100 or more engineered strains. We now state the proof-of-concept scope explicitly in the Abstract and Introduction and avoid suggesting that the present experiments establish unrestricted scalability or practical deployment at that scale.

We have also added the requested discussion of the increasing data burden. The dedicated limitations subsection states that

“Scaling to larger input panels will also increase calibration and data requirements, despite the additional information contained in time-resolved measurements.”

In addition, this subsection explains that resolvability depends on system-specific response ranges, cross-reactivity, temporal sampling, biological and measurement variability, and calibration data. Consequently, there is no universal number of strains required to resolve a “practically relevant” number of molecules. The number of resolvable inputs is not necessarily equal to the number of strains or reporters, as illustrated by the five-input example, but it must be established empirically for each implementation using representative calibration and independent validation.

Finally, we have clarified that the contribution does not reside in the novelty of the ODE, VAE, or MLP components individually. Rather, the study evaluates their integration into a new workflow for quantitative inference from time-resolved biological responses that may be cross-reactive or indirect. The ODE model is described as phenomenological and is used for structured data augmentation rather than as a uniquely identifiable mechanistic model. The revised manuscript therefore presents the conceptual and experimental contribution at the level supported by the demonstrated systems, while explicitly identifying larger-scale consortia and alternative readout modalities as future directions.