nature portfolio

Peer Review File

HiDDEN: A machine learning method for detection of disease-relevant populations in case-control single-cell transcriptomics data

Editorial Note: This manuscript has been previously reviewed at another journal that is not operating a transparent peer review scheme. This document only contains reviewer comments and rebuttal letters for versions considered at *Nature Communications*.

Reviewer #3 (Remarks to the Author):

The authors have addressed all my concerns in the last round. I do not have any further issues regarding the manuscript.

Comments on Reviewer #1's review:

After carefully reviewing the revision of HiDDEN and their response to Reviewer 1, in my point of view, most of the concerns from Reviewer 1 have been well-addressed or justified.

Specifically in response to the three major concerns:

The novelty and originality of the work are ok. The identification of perturbed cells or phenotype-associated cells is of great significance for both genomic and translation studies and still needs efforts to be more efficient and accurate. New models, rather than optimization of existing ones, do matter. However, indeed, the authors did not elaborate on the challenges and existing efforts in the field in the Introduction section. Adding a paragraph or 2 for clarification would be good.

The authors have implemented new benchmarking analyses in the revision to meet the suggestions from Reviewer 1, including parameter tunings, the addition of comparisons to mixscape, etc. These newly added experiments or justifications convinced me that HiDDEN is robust and has better results than other related tools.

Using a provided sample label in the first place and then refining it is a common strategy, that not only provides an initialized starting point but also increases the tool efficiency. I am not worried about the double-dipping issue. However, I do wonder whether HiDDEN can do it completely without a predefined sample label or if there exist multiple kinds of perturbations.

However, I still feel like some minor issues need more clarification or adjustments.

Although the ECCITE-seq data does not have ground truth perturbation labels, I do agree that the paired CITE-seq data can be a good resource for validation. The authors should at least try to see whether there are included proteins in the paired CITE-seq that can show differences between HiDDEN-predicted perturbed and non-perturbed cell groups.

“However, as a future direction, our method can be applied to additional contexts in which the aim is to focus the latent space on a particular distinction, for example, to explore subtle genotype effects (i.e. eQTLs) and sexual dimorphism.” The expansion statement the authors made in the discussion section is still strong and subjective. It is not trivial to directly apply HiDDEN to genotype effects (sexual dimorphism is fine) as eQTL effects can be much more complicated than simple phenotypic batches/perturbations. Also, the claim of applying HiDDEN to spatial and multi-omics data either needs proof or is completely removed, as the data distribution and structure are different than scRNA-seq.

There are too many ambitious statements using words like “outperform”, “novel”, etc. I suggest the authors can lower the tone by changing to soft words. This is also the requirement of Nature Communications if I remember correctly.

Reviewer #4 (Remarks to the Author):

The authors proposed HiDDEN – a method to define cell-level resolution perturbation labels, and identify potentially rare affected subpopulations. The method is based on dimensionality reduction and logistic regression. The authors validate their methods using synthetic data, a multiple myeloma dataset, and a mouse model of demyelinating disease. Overall, the manuscript is well-written and easy to follow.

The authors have undergone a round of revisions to their original manuscript, and have incorporated substantial improvements to code accessibility and usability. The authors have also added comparisons to Mixscape in addition to the other methods incorporated in the manuscript. Finally, the authors have benchmarked nonlinear dimensionality reduction techniques in addition to PCA for their intermediary step prior to training a logistic regression.

Unfortunately, it remains unclear how the authors contend with the issue of double dipping during subsequent analysis after the HiDDEN treatment label refinement process. Which analyses can be safely and correctly performed after HiDDEN is used to refine treatment labels without concern for increased false positive calls? This is an issue that should be clearly delineated in the text of the manuscript (and ideally also within tutorials and other supplementary user-facing materials) prior to publication, with or without additional experiments.

Major Points:
• The design of the framework is simple and easy to comprehend.
• The authors provide a publicly available GitHub as well as tutorials for use of the tool.
• The authors have provided benchmarks for their tool in several scenarios including simulated and real-world data.
• The authors have pursued biological validation for some findings predicted by their method.
• One major issue with the method remains the concept of “double dipping” into the data, as the HiDDEN framework is designed for subsequent analysis (including differential gene expression analysis) to be performed without modification on the refined labels. While this may result in signal amplification, it may also result in improper downstream hypothesis testing. As reviewer #1 raised in the original review, this was not discussed during the manuscript, and require proper mathematical analysis to identify which analyses can be correctly performed in a “HiDDEN-informed” manner, and which cannot. In addition, I would have liked to see some experiments which assess false-positive calls, or treatment-associated signal when in fact there is none.
• It is not clear how HiDDEN would perform in settings where there are multiple different underlying cell types each with a different response to a perturbation. Both the memory/naïve B-cell experiment and the mouse endothelial cell experiment use a single cell type as the substrate of analysis. How does HiDDEN perform in such scenarios?
• The authors do not provide a comparison to other published single-cell resolution perturbation analysis methods such as CINEMA-OT (for disclosure, I am an author of this method) and CellOT. These methods could also identify “control-like” cells in the perturbation condition and assign a minimal response value.

Minor Points:

• Supplementary figure 6a – typo in figure title should read: “comparing ground truth labels to continuous scores”
• Figure 5B – validation of gene expression with RNA scope for Lgals1 and S100a6. Was any quantitation of the coexpression done? And if so is this shown anywhere? Were these genes ones that were specifically identified using the HiDDEN approach? Or were these genes also identified by a standard sample-level treatment label differential gene analysis?

Best,
Rahul Dhodapkar, MD

Reviewer #4 (Remarks on code availability):

While the code seems to be usable and readable, I would have preferred if it were importable as a PyPI package (hiddensc seems to already be written as such).

Updated manuscript: HiDDEN_Manuscript_20240421

We would like to thank the editor and the reviewers for their thoughtful feedback on our manuscript and our responses to a first round of revisions. Based on this feedback, we have identified the following key aspects of the work that were recommended for improvement:

- Validation using ECCITE-seq data
- Double-dipping concern regarding using all cells
- False positive calls when no perturbation signal is present

A point-by-point response is provided below and an updated version of the manuscript is attached for your consideration.

Reviewer #3 (Remarks to the Author):

- *The authors have addressed all my concerns in the last round. I do not have any further issues regarding the manuscript.*

Comments on Reviewer #1's review:

After carefully reviewing the revision of HiDDEN and their response to Reviewer 1, in my point of view, most of the concerns from Reviewer 1 have been well-addressed or justified.

Specifically in response to the three major concerns:

- *The novelty and originality of the work are ok. The identification of perturbed cells or phenotype-associated cells is of great significance for both genomic and translation studies and still needs efforts to be more efficient and accurate. New models, rather than optimization of existing ones, do matter. However, indeed, the authors did not elaborate on the challenges and existing efforts in the field in the Introduction section. Adding a paragraph or 2 for clarification would be good.*

We agree with the reviewer that a description of existing methods and challenges is better-positioned earlier in the manuscript. Therefore, we have moved the following paragraph from the Discussion section to the Introduction:

“Detecting cell-level transcriptional changes across experimental conditions is one of the big promises of high-resolution single-cell expression data. In recent years, several methods have been proposed to characterize perturbation effects in single-cell data. The standard analysis workflow performs label-agnostic dimensionality reduction and clustering, followed by comparisons of cell attributes across condition labels within clusters. CNA–provides a cluster-free approach identifying regions in the latent space of uneven mixing of condition labels. MELD produces a continuous measure of the perturbation effect by distributing the condition labels among neighbors in the cell state manifold. Milo performs differential abundance testing among experimental conditions in the presence of continuous trajectories. Mixscape removes known confounding sources of variation and dissects successfully from unsuccessfully perturbed cells in gene knockout screens where it is expected that a high proportion of cells in the case sample will be perturbed. These approaches rely on at least one of the following assumptions: 1) the condition labels correctly represent the presence or absence of an effect in individual cells; 2) the perturbation effect is a dominant signal in the latent space; or 3) that any confounding sources of variation are known and can be removed. However, these assumptions might not always be met – often,

perturbation effects are small relative to the biological heterogeneity and technical noise, or the proportion of affected cells is small and therefore the condition labels are mostly incorrect."

- *The authors have implemented new benchmarking analyses in the revision to meet the suggestions from Reviewer 1, including parameter tunings, the addition of comparisons to mixscape, etc. These newly added experiments or justifications convinced me that HiDDEN is robust and has better results than other related tools.*

- *Using a provided sample label in the first place and then refining it is a common strategy, that not only provides an initialized starting point but also increases the tool efficiency. I am not worried about the double-dipping issue. However, I do wonder whether HiDDEN can do it completely without a predefined sample label or if there exist multiple kinds of perturbations.*

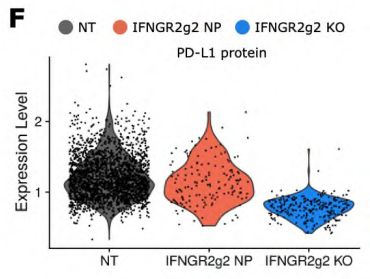
In the absence of a predefined sample-level label, we would not be able to leverage the HiDDEN framework, since without any label there would be no data to train the prediction model on. To deploy HiDDEN, the user could provide a global-level label, for example by clustering the data into two groups. One approach in that vein is the standard analysis pipeline of pulling all cells together and clustering them to identify perturbed subpopulations agnostic to the origin of the samples. However, as demonstrated in our Naive B/Memory B ground truth mixtures, this approach is underpowered in distinguishing the perturbed subset especially when the size or the strength of the perturbation is small (Figure 2G, Supplementary figure 3).

Regarding the question of whether HiDDEN can be applied if there exist multiple kinds of perturbations – as mentioned in the discussion, we foresee HiDDEN's potential extension to the setting of multiple kinds of perturbations by replacing logistic regression with a multi-class classification method https://scikit-learn.org/stable/modules/classes.html#module-sklearn.multiclass.

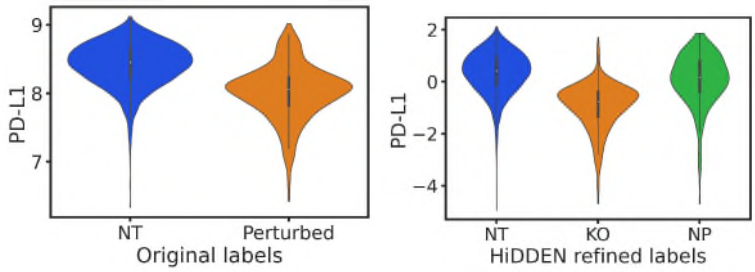
However, I still feel like some minor issues need more clarification or adjustments.

- *Although the ECCITE-seq data does not have ground truth perturbation labels, I do agree that the paired CITE-seq data can be a good resource for validation. The authors should at least try to see whether there are included proteins in the paired CITE-seq that can show differences between HiDDEN-predicted perturbed and non-perturbed cell groups.*

We appreciate the reviewer's suggestion that we utilize the published ECCITE-seq dataset generated from stimulated THP-1 cells by Papalexi et al. 2021 as a source of additional validation for HiDDEN. We fit our model to the RNA-seq expression profiles of 3599 cells (1213 cells targeted with IFNGR2 and 1386 cells expressing non-targeting (NT) gRNAs). HiDDEN outputs three groups of cells – NT cells retain their status and targeted cells are divided into knockout (KO) and non-perturbed (NP) cells. The authors of the original study validate the refinement of Perturbed cells into KO and NP for the cells targeted with IFNGR2 by plotting the expression of the PD-L1 protein in these cells (Figure 2G reproduced below).



Similar to the original study, we also confirm that the cells identified by HiDDEN as KO have lower PD-L1 expression compared to the non-perturbed targeted cells.



- *"However, as a future direction, our method can be applied to additional contexts in which the aim is to focus the latent space on a particular distinction, for example, to explore subtle genotype effects (i.e. eQTLs) and sexual dimorphism." The expansion statement the authors made in the discussion section is still strong and subjective. It is not trivial to directly apply HiDDEN to genotype effects (sexual dimorphism is fine) as eQTL effects can be much more complicated than simple phenotypic batches/perturbations. Also, the claim of applying HiDDEN to spatial and multi-omics data either needs proof or is completely removed, as the data distribution and structure are different than scRNA-seq.*

Following the reviewer's suggestion we have adjusted the language we use to describe potential future directions for expansion of HiDDEN in the Discussion section.

"However, as a future direction, the general idea of the HiDDEN framework has the potential to serve as a blueprint that can be modified and build upon to accommodate the challenges in additional contexts in which the aim is to focus the latent space on a particular distinction, for example to explore subtle genotype effects (i.e. eQTLs) and sexual dimorphism. The HiDDEN framework is amenable to extensions to spatial and multi-omics data, as well as applications beyond a binary output, such as multi-stage disease progressions or time-course experiments, with appropriate modification to the dimensionality reduction and prediction modules of the framework."

- *There are too many ambitious statements using words like "outperform", "novel", etc. I suggest the authors can lower the tone by changing to soft words. This is also the requirement of Nature Communications if I remember correctly.*

We have gone through the manuscript and removed claims of novelty, and further focused on simply and factually reporting our findings.

Reviewer #4 (Remarks to the Author):

The authors proposed HiDDEN – a method to define cell-level resolution perturbation labels, and identify potentially rare affected subpopulations. The method is based on dimensionality reduction and logistic regression. The authors validate their methods using synthetic data, a multiple myeloma dataset, and a mouse model of demyelinating disease. Overall, the manuscript is well-written and easy to follow.

The authors have undergone a round of revisions to their original manuscript, and have incorporated substantial improvements to code accessibility and usability. The authors have also added comparisons to Mixscape in addition to the other methods incorporated in the manuscript. Finally, the authors have benchmarked nonlinear dimensionality reduction techniques in addition to PCA for their intermediary step prior to training a logistic regression.

- *Unfortunately, it remains unclear how the authors contend with the issue of double dipping during subsequent analysis after the HiDDEN treatment label refinement process. Which analyses can be safely and correctly performed after HiDDEN is used to refine treatment labels without concern for increased false positive calls? This is an issue that should be clearly delineated in the text of the manuscript (and ideally also within tutorials and other supplementary user-facing materials) prior to publication, with or without additional experiments.*

We understand the reviewer's concern that HiDDEN uses expression profiles and sample-level labels of all cells in a dataset to define refined labels which are later used along with the expression profiles in downstream tasks, such as DE testing. While we agree with Reviewer 1, that *"Using a provided sample label in the first place and then refining it is a common strategy, that not only provides an initialized starting point but also increases the tool efficiency. I am not worried about the double-dipping issue.",* we did include additional experiments to investigate the potential effect of double-dipping through a leave-one-out cross-validation strategy, detailed below.

We would like to further note that the standard practice routinely performed in nearly all single-cell analyses involves generating cluster labels using all cells and then performing differential expression analysis on those cells stratified by these cluster labels. According to this reviewer's concern, this practice would also qualify as 'double dipping'. Nevertheless, this strategy has revealed many interesting facets of biology, and we believe that HiDDEN will too.

Major Points:
- *The design of the framework is simple and easy to comprehend.*
- *The authors provide a publicly available GitHub as well as tutorials for use of the tool.*
- *The authors have provided benchmarks for their tool in several scenarios including simulated and real-world data.*
- *The authors have pursued biological validation for some findings predicted by their method.*

- *One major issue with the method remains the concept of "double dipping" into the data, as the HiDDEN framework is designed for subsequent analysis (including differential gene expression analysis) to be performed without modification on the refined labels. While this may result in signal amplification, it may also result in improper downstream hypothesis testing. As reviewer #1 raised in the original review, this was not discussed during the manuscript, and require proper mathematical analysis to identify which analyses can be correctly performed in a "HiDDEN-informed" manner, and which cannot.*

We thank the reviewer for inviting an additional analysis regarding the 'double dipping' concern, i.e. that we are using a cell's expression profile and condition label (Case/Control) for refinement of the cell's perturbation label. To isolate the effect of a cell's expression vector and metadata on the binary label HiDDEN assigns to that cell, we adopt a Leave-One-Out (LOO) cross-validation strategy. LOO is a commonly used approach to cross-validation in machine learning when the aim is to evaluate how well the model would do when asked to make predictions on a data point it has not already seen.

In particular, we took our ground truth Naive B/Memory B cell mixture dataset with 50% Memory B cells in the Case condition. We randomly drew the indices of a tuple of cells from the Case condition denoted (MemoryB_Case, NaiveB_Case) – one with ground truth identity Memory B and the other, Naive B. We then trained two HiDDEN models – one that utilizes all the data, and a second one that does not use the expression and metadata of the LOO tuple. The first model

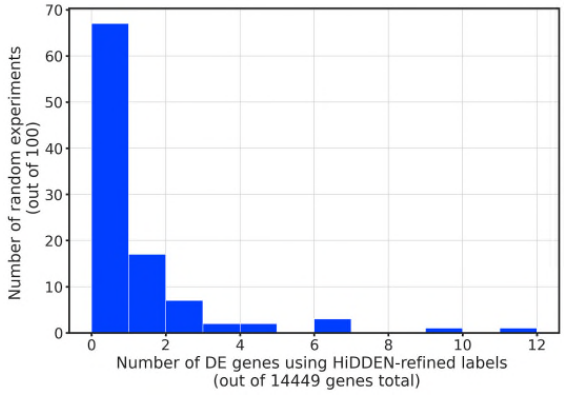
is the usual way of applying the HiDDEN framework to output binarized refined labels. For the second model, all parameters of the dimensionality reduction, prediction model, and clustering procedure are trained without the LOO tuple. Then, the trained model is applied to the expression profiles of the LOO tuple of cells to derive their LOO-refined-labels.

We repeated this random experiment 200 times and evaluated the agreement between the two training regimes by directly comparing the values of the binary refined labels. We found that they agreed 95.5% of the time for the MemoryB_Case condition, and 92.5% for the NaiveB_Case condition. Furthemore, in the instances where the labels did not match, the LOO label was correct 55.5% of the time for the MemoryB_Case condition, and 26.7% for the NaiveB_Case condition, suggesting that the differences are due to stochastic effects.

In conclusion, we find that training the components of the HiDDEN framework with all cells does not introduce substantial differences on the output of the method, compared to a leave-one-out approach. Furthermore, training the model once with all the data is much more computationally efficient than training the model N times, where N is the number of cells in the dataset.

- *In addition, I would have liked to see some experiments which assess false-positive calls, or treatment-associated signal when in fact there is none.*

We appreciate the reviewer's request for additional characterization of HiDDEN's false-positive rate when no treatment-associated signal is present. Towards this end, we created a ground truth dataset where both the Case and Control conditions are composed of Naive B cells only. We took the 1900 Naive B cells we used in our ground truth experiments and assigned each cell a random Case or Control label. We repeated this random experiment 100 times. Each time, we applied the HiDDEN framework to derive binary refined labels. Using these labels, we performed DE testing between the cells labeled unperturbed vs the cells labeled perturbed after the application of HiDDEN. Note that there are 14 DE genes between Naive B and Memory B cells in our ground truth dataset with the smallest percent Memory B cells in the Case condition (5%). The distribution of the number of DE genes across the 100 random experiments is plotted below. In conclusion, in all of the random experiments where no perturbation effect is present, the number of false positive DE genes was smaller that the number of DE genes in our ground truth dataset with the smallest perturbation effect size.



- *It is not clear how HiDDEN would perform in settings where there are multiple different underlying cell types each with a different response to a perturbation. Both the memory/naïve B-cell experiment and the mouse endothelial cell experiment use a single cell type as the substrate of analysis. How does HiDDEN perform in such scenarios?*

As we describe in the Discussion section "HiDDEN has several limitations. First, given that the perturbation effect would likely differ across cell types, the method needs to be applied one cell type at a time." The field has settled on a very well-established workflow to clustering single-cell data into cell types. Furthermore, one can transfer cell type labels from well-annotated atlases. Therefore, clustering single-cell transcriptional profiles into cell types is a relatively stably solved problem. Overall, we do not see this as a major limitation, but nevertheless, we already address it in the Discussion section.

- *The authors do not provide a comparison to other published single-cell resolution perturbation analysis methods such as CINEMA-OT (for disclosure, I am an author of this method) and CellOT. These methods could also identify "control-like" cells in the perturbation condition and assign a minimal response value.*

We attempted to benchmark HiDDEN against CINEMA-OT on our Naive B/Memory B ground truth mixtures. We used the CINEMA-OT implementation provided in the pertpy library found at https://github.com/theislab/pertpy/blob/main/pertpy/tools/_cinemaot.py. We found that CINEMA-OT does not return a single-cell level treatment effect for all cells in the dataset. By default, it returns a matrix, de.X in original dim or de.obsm['X_embedding'] in reduced dim (by default 20), only for the cells in the Case condition. It is possible to force the dimension of de.obsm['X_embedding'] to be equal to 1 (by setting the dim parameter in the Cinemaot.causaleffect function to 1) to get a scalar single-cell treatment effect score. However, this score is only computed for the cells in the Case condition. Since there are no comparable scores produced for the cells in the Control condition, the CINEMA-OT treatment effect scores cannot be utilized to distinguish between perturbed and unperturbed cells in the Case condition, and cannot be directly compared with HiDDEN.

Likewise, we considered the suggestion to benchmark HiDDEN against CellOT. However, CellOT also does not provide a perturbation score that reflects the extent to which each cell is perturbed. The method outputs a learnt mapping that tells us how to map control cells to their predicted perturbation state. To achieve this, CellOT first assumes accurate Case/Control labels, then learns an optimal transport map modeled by a function that maps untreated (control) cells to their treated (case) counterparts. As a result, all case cells are mapped into the perturbation space without the possibility to augment their status as unperturbed and hence a direct comparison with HiDDEN is not feasible.

We wish to emphasize that we have made extensive comparisons with Mixscape, CNA, MELD, and Milo. We would have been happy to include a comparison with CINEMA-OT and CellOT but we were unable to identify a strategy for direct comparison with HiDDEN.

Minor Points:

- *Supplementary figure 6a – typo in figure title should read: "comparing ground truth labels to continuous scores"*

Thank you. The figure has been adjusted accordingly.

- *Figure 5B – validation of gene expression with RNA scope for Lgals1 and S100a6. Was any quantitation of the coexpression done? And if so is this shown anywhere? Were these genes ones that were specifically identified using the HiDDEN approach? Or were these genes also identified by a standard sample-level treatment label differential gene analysis?*

The purpose of this RNA scope validation was to confirm the existence of triply positive cells–those that express *Flt1, Lgals1*, and *S100a6* as predicted by the HiDDEN analysis (and were not identified using ordinary differential gene expression)--rather than to quantitatively assess the proportions of different endothelial populations. As can be appreciated in the images, these cells are quite sparse, owing in large part to both the sparsity of endothelial cells–as judged by the sparse expression of *Flt1*. Indeed, as shown in the figure, these triply positive cells were detected within the demyelinating lesion (and could not be found in the vehicle-injected mice).

Reviewer #4 (Remarks on code availability):

- *While the code seems to be usable and readable, I would have preferred if it were importable as a PyPI package (hiddensc seems to already be written as such).*

We have deliberately chosen to publish HiDDEN's code in this format. We find that it installs very easily and efficiently without being a PyPI package. Furthermore, it keeps the underlying codebase more flexible. We found that users of HiDDEN particularly value being able to combine our framework with existing dimensionality reduction techniques which evolve and continue being updated, and we do not see it feasible to keep track of updates on all of the possible dimensionality reduction modules, eg, scvi.

Reviewer #3 (Remarks to the Author):

The authors have addressed all my concerns.

Reviewer #4 (Remarks to the Author):

The authors have substantially improved the manuscript and responded thoughtfully to all reviewer comments. I have no further concerns.