1 Joeri Ruyssinck

1.1 Strengths of the thesis

▷ *The research domain of this thesis is young and technical evolutions in the encompassing field move fast. This makes it especially challenging to do research in the field, as often the original PhD thesis subject is obsolete at the end. The candidate has done exceptionally well to not only suggest and create novel algorithms which are at the state-of-the-art but also to grasp and tackle the greater high level challenges in the field, resulting in a potential seminal benchmarking paper.*

▷ *The thesis as a whole contains many relevant contributions, of which several have been published in high quality journals.*

▷ *The figures in the thesis are of exceptional quality. They are creative and able to convey large volumes of information to the reader in a visually pleasing format.*

▷ *The introduction chapter of the thesis demonstrates that the candidate understands and is able to convey the background of both (computational) biology and computer science.*

▷ *The passion of the candidate to bring modern good software development practices to the field is present throughout the work and might contribute to a more mature algorithm development in computational biology.*

1.2 Weaknesses of the thesis

▷ *The quality of writing, layout and graphical presentation is of a lower quality in the non-published parts of the thesis. Certain unpublished chapters feel rushed and unfinished. Furthermore, for some chapters it feels that they could be merged into a broader discussion chapter.* ⚠⧖

▷ *The Introduction and Discussion chapter is missing a more broad context description. After reading the work, I gained very little insight how other researchers are using these tools to advance science.* ⚠⧖

▷ *The main contribution of this work is the excellent benchmarking study published in Nature Biotechnology. However, this also effects the overall cohesiveness of the work, as information is repeated throughout the book at several levels of detail. This however a common side-effect of grouping published manuscripts into a thesis and I'm aware there is no (quick) solution to this.*

▷ *It is very rare to see that research groups (or individual researchers) invest this much effort and time in creating qualitative software. I agree with the ideas of the candidates that these practices to adopt software to accompany research ideas should be encouraged. Yet I wonder how the current academic system can support such concepts. Academics often hold different positions in different research groups around the world, who should be the maintainer in these cases? The original PhD students that developed the work? The original research group? Will such higher standards not lead to barriers which might be harder to overcome for PhD students which might be constrained by limited support or funding?* ⚠

1.3 Summary and chapter 1

▷ *As an interdisciplinary thesis, it is very much enjoyable that the introduction chapter contains a concise, yet precise overview of the concepts which are needed to understand the work of the candidate. I believe the chapter is nicely written and especially the flow of information is well chosen. However, the candidate only briefly touches upon the societal value of the technological innovations he describes. For a PhD thesis, I think it would be an added value if the candidate could briefly describe the practical implications of the technological advances (e.g. precision medicine, new treatments) and motivate the research from a non-technical view point (How is/can this PhD change the world?)* ⚠⧖

▷ *Small details / typo's: [...]*
Answer: These issues have been fixed, unless noted below.

▷ *Figure 1.9 does not contain SCORPIUS (probably published before, but still a bit weird for a PhD*

containing the manuscript)
Answer: Indeed, this figure was adapted from Cannoodt et al. 2016 and serves as an illustration that TI methods often have common building blocks. Extending this figure to >75 TI methods would decrease the interpretability of the figure.

▷ *Bold is only used in 1.1.2 and research challenges*
Answer: Usage of the bold font style in 1.1.2 and 1.3 serve clear, but different, purposes. In 1.1.2, three words are highlighted to establish a link between each term's introduction in the first paragraph, and their explanations in the second paragraph. In 1.3, important sections of each research objective are highlighted in order to highlight the structure of this section.

▷ *Use of we/I in summary*
Answer: Usage of first person pronouns in this dissertation is very limited. In the summary of my contributions to the field, I believe usage of the first person improves legibility in comparison to when the phrases in question are converted to a passive form.

1.4 Chapter 2

▷ *It is not clear to me what is meant with 'such as sampling a cell at a certain time point and once more at a later point' and why this is currently not possible?*

▷ *Figure 2.1.A. What does the figure in 'Combine simulations' represent? Are these single simulations which are mapped continuously to the backbone?*

▷ *Figure 2.5. It is not explained what Subfigure F represents.* *Only after the reader has completed the chapter, it becomes clear that the red lines probably represent snapshots and the lines, curves of gene expression.*

▷ *Figure 2.6: Consider changing the dashed lines*, *as they are not (clearly) visible printed*

▷ *Figure 2.6 and the example text for the cyclical example is unclear.*

▷ *While I understand that the instructions result in the desired state change, I also see many other configurations which would result in a cyclical example.*

▷ *Why is it currently needed to explicitly keep certain module expressions constant to determine the backbone? Would the same result not be achieved by simply simulating X times and averaging expression levels?*

▷ *Figure 2.7. The figure legend could be extended*

▷ *The introduction of FANTOM5 is abrupt and it would be better if the context would be briefly explained to the reader.*

▷ *The paragraph describing Step 3 is unclear: -> Target genes ARE? Regulated ... but is-> are ?*

▷ *Figure 2.8 can incorrectly suggest that target genes can only have a single regulator assigned*

▷ *Table 2.2. contains many default 'magic' numbers?* *It would be nice to know how in practice you ended up selecting them and which problems you encountered during design.*

▷ *Since these are default values, it would suggest that the user can modify them? Why would a user do that and what would be the impact?*

▷ *Am I correct that currently splicing offers little extra value as the formula's would imply that alternative splicing is not supported?*

▷ *During a first read, it was confusing at times to read specifically about the strategy to determine the backbone while it had not been described how the simulation was performed.* *At a certain time I was under the impression that dyngen did not simulate cell evolution end-to-end but only between states. Would the readability be improved if the order was changed in which the concepts are explained? The backbone is only needed for the visualisation.* ⚠⧖

▷ *2.4.6. is a bit worrying. Could I reproduce dyngen without knowing these details? Does the sampling process not deserve more explanation?*

▷ *Small details / typo's: [...]*
Answer: These issues have been fixed.

1.5 Chapter 3

▷ *There are some artefacts of converting the paper into a PhD chapter: e.g. references to Supplementary tables/figures should be changed to the actual table number.*
Answer: Certain supplementary figures and files had not been included in the dissertation, but instead the reference contained a clickable URL which directs the reader to a downloadable file. Since this information is lost to readers of the printed dissertation, explicit captions have been added in a section for supplementary figures and tables, and contain a printable URL which direct the reader towards the file.

▷ *Some words are marked with an asterisk but the footnote? is missing.*
Answer: Section 3.4.2 uses an asterisk to denote special cases and does not refer to a missing footnote. This notation is introduced in the paragraph leading up to the usage thereof. An actual asterisk has been added to the introduction of the notation to improve readability of this section.

▷ *You mention consensus predictions, is this something you investigated more in depth? Would 'ensemble' predictions result in better performance? Is there a good way to aggregate the predictions?*

▷ *Monocle DDRTree seems to overestimate the topology based on Figure 3.5, is this somehow related to the way you have to force the algorithm into the common format?*

▷ *The dyngen described here seems to be less advanced than the one from Chapter 2?*

▷ *Small details / typo's: [...]*
Answer: These issues have been fixed, unless noted below.

▷ *Consistency: this chapter is US English*

1.6 Chapter 4

▷ *Figure 4.7: The figure is lacking what the colors represent.*

▷ *Figure 4.7: I see that even if you use the same dimensionality reduction; there is still a lot of variation between the methods on a very simple example. Could you clarify how often these methods disagree on trivial examples?*

▷ *Dyno offers a lot of guidance with respect to how to choose your methods, run them, etc. Did you consider support for features such as interpretation or automated error detection after prediction?*

▷ *This chapter does not really read as a scientific paper (yet). There are few comments I have since this is mainly an instruction manual + feature overview.* ⚠⧖

▷ *Small details / typo's: [...]*
Answer: These issues have been fixed, unless noted below.

▷ *Figure 4.7 is missing slingshot and paga label in part B*
Answer: Figure 4.7 does contain labels for Slingshot and PAGA. A small gap between part A and part B has been added, to improve interpretability.

1.7 Chapter 5

▷ *Comparing the preprint with the chapter in the thesis. Some minor changes seem to have been made to SCORPIUS: e.g. I believe in the pre-print classical Torgerson MDS is mentioned, while the more optimized Landmark MDS is used in the chapter and final version. There is also no mention of outlier removal. The main result Figure 5.4 is however identical. Did this change or other potential changes have no (measurable) effect on the initial analysis? I also noticed that the biological conclusions or hypothesis have been reduced in claim strength in the paper or have been removed. It would* ⚠

be interesting to know if this is due to new insights that have been found in the period 2016-> 2019 or mainly because the paper's target audience has shifted.
Answer: As mentioned, the preprint was originally published in 2016. In order to submit this work to a peer-reviewed journal in 2019, the manuscript and the software was rewritten to reflect developments within the field.

Regarding the methodology. Classical MDS has indeed been replaced with Landmark MDS, since this allows scaling beyond 10'000 cells. Principal curves has been modified to also allow scaling beyond 10'000 cells. These changes have little to no effect on any results. Outlier removal was removed because this step is now part of mainstream preprocessing pipelines such as Seurat. Initialisation of the curve (k-means clustering + shortest path) has been removed – while this part of the algorithm showed improved performance when applied to the datasets included in the bioRxiv benchmark, this was not the case when applied to a larger set of datasets.

Regarding reducing the biological conclusions. Section 5.2.2 regarding the functional modules found by SCORPIUS was reduced in character count, yet it makes the same claims. The main differences are that gene names and explanations of the biological functions are not discussed in detail, but are instead summarised in figure 5.4. The reasoning for these changes are indeed to adapt the text to the target audience. For example, detailed explanation of what actin polymerisation is, was removed; because immunologist readers already know this information and data scientist readers are presented with sufficient information when told that particular genes recovered by SCORPIUS are related to a biological process that is highly relevant to dendritic cell development, that is, actin polymerisation.

▷ *This chapter would offer information on how the developed algorithms are being used to discover new insights, so I'm unsure if it is beneficial if this chapter evolves more into a technical description of Scorpius as the former is missing from the thesis.*

▷ *Small details / typo's: [...]*
Answer: These issues have been fixed, unless noted below.

▷ *It would help to list the title of the manuscript on the pre-print server*
Answer: I fixed the DOI. This should be sufficient to find back the manuscript on the pre-print server.

1.8 Chapter 6

▷ *The abstract seems to be missing a sentence.*

▷ *A main challenge in network inference has always been choosing a suitable cut-off to determine if something is interacting or not. There are examples of regions in the network which are accurate but come at a much later stage in the ranking, at which time many spurious edges have been added to other parts of the network. I can only assume that this problem becomes even more challenging if one wishes to create case-wise GRN.*

▷ *The added value of the last paragraph in 6.3. seems small*

▷ *Why is Scorpius used to generate Figure 6.3? Is this not confusing, since I assume we are not working with single cell data?*

▷ *I'm afraid to ask, but I don't get the reference to the name bred.*

▷ *I don't understand how the approach in figure 6.5 is different from GENIE3 and how it results in a tensor instead of matrix/ranking of 'shuffle' variable importance scores. I believe I'm missing important information on how SCENIC works to be able to understand the modifications. I really struggle with this chapter: is the method executed for each profile or for all profiles together? Do you not end up with a single GRN for each of the profiles and as such why would bred be a true case-wise NI method and the Scenic approach would not? It would be nice to discuss this more during the defense.*

▷ *The readability of Figure 6.1 could be improved by using a black font*

1.9 Chapter 7

▷ *Small details / typo's: [...]*

Answer: These issues have been fixed.

1.10 Chapter 8

▷ *Does this contribution not better fit in an overall discussion of the thesis?*

▷ *Small details / typo's: [...]*
Answer: These issues have been fixed.

1.11 Chapter 9

▷ *As the candidate is not first author on this paper, I'm not sure if this paper should be included in the PhD book verbatim. Perhaps it should not be included or it could be merged with chapters 8-9-10 into a single discussion chapter. Maybe at the very least, the candidate should state a bit more explicitly why the content was included in the PhD book and what the personal contribution was.* ⚠

▷ *Author contributions is not correct (CS)*
Answer: ???

1.12 Chapter 10

▷ *The chapter contains a concise recap and critical view on the contributions in literature. It however does not discuss Chapter 9 (see comment above).*

▷ *The discussion could be improved by stating how the tools are allowing new biological insights. Only a small hint is provided in the last sentence of 10.4.*