

# Integration of multi-omics data with artificial intelligence

Citation for published version (APA):

Ochoteco Asensio, J. (2022). *Integration of multi-omics data with artificial intelligence: studying the toxic effects on the post-transcriptional regulation*. [Doctoral Thesis, Maastricht University]. Maastricht University. <https://doi.org/10.26481/dis.20221109ja>

## Document status and date:

Published: 01/01/2022

## DOI:

[10.26481/dis.20221109ja](https://doi.org/10.26481/dis.20221109ja)

## Document Version:

Publisher's PDF, also known as Version of record

## Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

## General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.umlib.nl/taverne-license](http://www.umlib.nl/taverne-license)

## Take down policy

If you believe that this document breaches copyright please contact us at:

[repository@maastrichtuniversity.nl](mailto:repository@maastrichtuniversity.nl)

providing details and we will investigate your claim.

# Scientific and Social Impact

Currently, regulatory agencies rely on the use of animals for testing drugs for several reasons. First, the most accurate way of testing drugs would be testing them in humans, which is not ethically possible. Second, modeling everything that happens in our bodies in a Petri dish is nowadays unfeasible: the complexity of a human body requires 30 trillion cells to keep functioning. In addition, each of those cells also involves its level of complexity: even a simple yeast cell contains 42 million proteins<sup>1</sup>, which does not include other essential molecules like sugars and fatty acids. Even so, recent incentives (such as the regulation (EC) No 1223/2009 of the European parliament<sup>2</sup>) have pushed the scientific community to search for alternative testing methods without the use of animals. Specifically, the area of Toxicology is affected, as it studies the potentially toxic effects of drugs both before and after being released on the market. A popular method for studying the effects a compound can have in humans is by testing them *in vitro*, that is, by exposing human cells. Doing so helps to narrow the bridge between what is being tested (human cells outside the body) and the actual goal of the study (human cells inside the body) when compared to animal testing, where the model and the end goal belong to different species. Nowadays, the development of induced pluripotent stem cell (iPSC) technology allows reverting any human skin cell to another tissue type (such as cardiac cells that contain the same DNA as the donor of the skin cells), without the need for surgery or invasive biopsies. Although some drugs may kill cells by simply destroying the membrane that encapsulates the cell, most of them disturb the cell in more subtle ways. One of these ways is the deregulation of the number of proteins synthesized by a cell. Proteins are essential molecules, as they perform most of the cell functions. For this reason, disturbing or blocking their production can lead to the disruption or death of a cell and/or the ones that depend on it. The processes that lead to the making of proteins involve mainly RNAs, molecules that work as messengers from DNA to proteins. Therefore, in Toxicogenomics, studying how specific treatments can affect these molecules can help understand their mechanisms of toxicity.

In **Chapter 2**, we assessed how a recently discovered class of RNA, called circular RNAs (circRNAs), are disturbed in heart cells by known toxicants. These circRNAs have been hypothesized to regulate microRNAs (miRNAs) by letting the latter bind to the former. When messenger RNAs (mRNAs) are not bound to miRNAs, they can provide the instructions to produce proteins. When miRNAs are occupied binding circRNAs, they are not able to bind to mRNAs. For this reason, by assessing changes in the number of circRNAs, miRNAs, mRNAs, and proteins; we helped better understand how these compounds (that are still in use) are being toxic in the human heart without the use of animal testing. Going further into the thesis, we realized that knowing how many proteins are in a cell is crucial to

understanding how a drug (or disease, or any other perturbation) affects a cell. Unfortunately, the technology used for doing it, mass spectrometry, does not measure all proteins in a cell. Instead, researchers tend to use transcriptomics, which exhaustively measures the quantity of RNAs. Nevertheless, the number of RNAs and the number of proteins do not always perfectly correlate with each other.

In **Chapter 3**, we designed an equation to estimate how many RNAs are available to produce proteins. We did that by counting the total number of mRNAs and subtracting the ones that will be affected by miRNAs, but only those miRNAs that are not binding to other RNAs (like another mRNA or circRNA). Nonetheless, the formula, which is focused mainly on RNA molecules, demonstrated an added value for only a subset of proteins. As a result, in **Chapter 4**, we went a step further. We built a large dataset with RNAs and their corresponding proteins and trained a machine learning model to predict the latter. Machine learning algorithms “learn” how to predict values by looking at how other similar values behave. Using our data, our model predicted well the increases and decreases of proteins, which can help others to predict how many proteins there are in a sample of cells based on how many there are in similar ones.

As mentioned before, in the area of Toxicogenomics it is of great interest to study the changes happening in a cell. Transcriptomics is exceptionally good at counting how many RNAs there are, consequently using this technology helps us understand which molecules change in quantity due to a specific cause. The statistical tools used to detect changes, though, do not work without fault. For this reason, experts in this technology can manually detect these errors. On the flip side, this manual curation is pretty time-intensive when taking into account the number of genes to be evaluated, and requires specialist knowledge to do so. That is why, in **Chapter 5**, we again trained a machine learning model. In this case, we taught the model to recognize the profile of genes that are typically of interest to the researcher. We built several of them with different characteristics and selected the best one, which we named ‘AutoRel’. Even though AutoRel was not flawless, it showed improvements by removing genes that were not of interest.

In an era of increasing societal pressure against animal testing, added to the inherent shortcomings of animal assays, regulatory agencies need to reevaluate their historical procedure of risk assessment. With the rapid development of methodologies that allow the analysis of the complete set of biological entities in a cell exposed to any substance, regulators will need both a better understanding of all the complex interactions behind molecular biology and powerful data analysis tools to integrate them. The work of this thesis contributes to this necessary transition toward a next-generation risk assessment. This is achieved by the discovery of new changes that happen when a toxic compound affects a cell, predicting protein measures that are usually unknown with artificial intelligence, and filtering results in an automated way to

65 have a better understanding of the changes that occur in a cell. In aggregate, this contributes to assess  
66 more accurately how toxic a drug is, making the use of treatments more safe and reliable.

## 67 Bibliography

68

69 1 Ho, B., Baryshnikova, A. & Brown, G. W. Unification of Protein Abundance Datasets Yields a  
70 Quantitative *Saccharomyces cerevisiae* Proteome. *Cell Syst* **6**, 192-205 e193,  
71 doi:10.1016/j.cels.2017.12.004 (2018).

72 2 *Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November*  
73 *2009 on cosmetic products (Text with EEA relevance),*  
74 <http://data.europa.eu/eli/reg/2009/1223/oj> (2009).

75