Inferring mechanisms of toxicity from differential genomics and semantic knowledge representations

by

I. J. Tripodi

B.S., University of North Texas, 2004

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy with a certificate in Interdisciplinary Quantitative Biology

Department of Computer Science

2020

This thesis entitled:
Inferring mechanisms of toxicity from differential genomics and semantic knowledge representations
written by I. J. Tripodi
has been approved for the Department of Computer Science

Dr. Robin D. Dowell

Dr. Lawrence E. Hunter

Dr. Ryan Layer

Dr. Orit Peleg

Dr. Daniel Larremore

Date ____________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Tripodi, I. J. (Ph.D., Computer Science, IQ Biology)

Inferring mechanisms of toxicity from differential genomics and semantic knowledge representations

Thesis directed by Dr. Robin D. Dowell and Dr. Lawrence E. Hunter

This thesis explores a combination of genomics analysis and semantic knowledge representation useful for computational toxicology. I first discuss a novel approach to infer differences in transcription factor (TF) activity between biological conditions, utilizing a variety of omics assays. This method for genome-wide exploratory analysis in exposure studies is valid for a simple and inexpensive protocol (ATAC-seq). It can also be used to study transcriptional perturbations by toxicants that may target protein receptors, at very early time points and with a fine time resolution, via more sophisticated protocols (GRO-seq, PRO-seq). Detecting changes between biological conditions is only half the problem. By integrating public databases, I provide ways to relate the highlighted TFs in different dimensions, thus expediting downstream analysis. Further exploiting the power of ATAC-seq, I show that there are inherent signatures in the peaks from ATAC-seq signal that, combined with the underlying sequence, can be used to predict the presence of nascent transcription or histone modifications at that genomic coordinate.

In addition to genomics analysis, I explore applications of semantic knowledge representation. I demonstrate how a consistent integration of data from public databases and open biomedical ontologies can be used to infer novel drug-drug interactions, chemical-protein relations, or enrichment of mechanisms of toxicity. A significant portion of computational toxicology work focuses on the prediction of *outcomes*, rather than the generation of mechanistic *explanations* for said outcomes (thus sometimes being perceived as a “black box”). I show it’s possible to produce putative explanations for our predictions of cellular toxicity modes of action from experimental data. The mechanism enrichment strategy accounts for the sequential order in which measured biological events happen. Here, the measured phenomena are changes in gene expression, however this mechanistic inference framework can be adapted to other types of mechanisms beyond toxicology.

Dedication

To all who strive for pushing science forward, always challenging current dogma. To my family, and especially to the memory of Kojak, one of the best friends one could ask for, which is the kind that irreversibly changes the course of your life.

Acknowledgements

I would like to thank both of my graduate advisors, Robin Dowell and Larry Hunter, from whom I've learnt far more than many technical (and incredibly useful) concepts: I've learnt how to think like a scientist from day one. I am also deeply grateful to the Interdisciplinary Quantitative Biology program, for the invaluable experience it provided me, allowing me to become "bilingual" in molecular biology and computer science. I would like to thank Adam Norris and Sujeet Bhat for briefly tempting me to seriously consider the academic world as an option, through exceptional teaching. I would also like to thank Manuel Lladser, for teaching me how to *really* read a scientific paper, and introducing me to the PhD program that would become the most important step I've taken in my professional career. I want to thank Andrea Stith, for believing in me and giving me the opportunity to show what I was capable of, and extend this acknowledgement to Kristin Powell and Amber McDonnell for always listening and helping navigate the administrative labyrinth. I'm grateful to Hristo Aladjov, for the opportunity to get involved in computational toxicology before even applying to graduate school. I'm also thankful for my lab mates for their support, especially Tiffany Callahan and Margaret Gruca for being outstanding collaborators, and Jacob Stanley for the encouragement and great conversation while decompressing at the rock climbing wall. I'd like to thank my friends outside of the science world, for the cheering and patiently listening to my gripes about publishing manuscripts. Finally, and most importantly, I would like to thank my family for their support and patience throughout this process, especially my wife Maru through the stress and permanently busy weekends, and my daughter Sofía who was born during the course of my doctoral studies, and turned me into a time management ninja and an incredibly proud dad.

Contents

Chapter

Appendix

Tables

Table

Figures

Figure

PH.D. THESIS

Ignacio J. Tripodi

Department of Computer Science

Interdisciplinary Quantitative Biology

University of Colorado, Boulder

May 2020

Chapter 1

Introduction

The disciplines of genomics and transcription network analysis, on one hand, and semantic knowledge representation, on the other hand, seem to be worlds apart. The former focuses on the analysis of the intricate regulatory networks in each cell of a living organism. The latter, on the other hand, deals with proper ontological representation of entities and the relationships between them. A goal of my thesis was to bring them together with a meaningful application to computational toxicology.

Mechanistic inference has been an area of artificial intelligence that was barely been explored, and a good target for combining these disciplines. Much of the research in computational biology is centered around determining outcomes or states from experimental data[1], such as whether a chemical compound is toxic to a particular type of body tissue. The act of inferring mechanistic behavior goes beyond that outcome, and attempts to explain *why* such an outcome occurs, and which are the series of steps that took place at varying levels of abstraction. What follows is an introduction to differential analysis (with a focus on studying differences in transcription factor activity), current approaches to integrate biomedical ontologies in a meaningful way, and the state of mechanistic inference as a nascent field.

1.1 Background: transcription, and transcription factors

The central dogma of molecular biology states that particular DNA regions, known as genes, are transcribed by an RNA polymerase enzyme into single-stranded RNA. There are three known types: RNA polymerase I, II, and II, though the majority of transcription is attributed to RNA polymerase II. Polymerases are large enzymes formed by many components, and RNA polymerase II in particular has an initiation form (when physically attaching to DNA), an elongation form (during the actual transcription phase), and a separate termination form (after synthesizing the end of the RNA transcript and detaching from DNA). The resulting RNA is initially rather unstable but undergoes various maturation modifications, such as polyadenlyation and splicing, to become a message RNA (mRNA). The mature message is a "template" for another sophisticated molecular machine known as a ribosome, to be translated into a protein. DNA is tightly wound around other proteins and RNA, forming a molecular complex referred to as chromatin, which adjusts dynamically to expose certain regions to DNA-binding proteins and the transcription machinery. Some of the critical components of chromatin structure include histones, protein complexes around which DNA strands are tightly bound like a spool, forming a nucleosome. Histone methylation, a process that involves methyl groups being transferred to some of the histone proteins, can alter transcription either way by helping uncoil nucleosomes or making regions of the genome even less accessible to transcription.

Transcription factors (TFs) are key pieces of the cellular machinery. These proteins control many cellular functions, from gene transcription regulation to differentiation targets. The ability to detect which TFs significantly changed their activity between two biological conditions genome-wide, has been an ongoing challenge for over two decades. Moreover, the methods used to generate hypotheses about the underlying biological mechanisms, based on the results from differential assays, have been evolving in parallel to increasing degrees of sophistication. We discuss in this review how both sides of the analysis have improved over time, the various approaches taken to detecting changes in TF activity, and those taken to generate hypotheses based on the resulting list of differentially-active TFs or genes.

A TF can be defined as a protein that binds to DNA, and alters transcription. By "binding" we imply a biochemical interaction that results in the TF becoming temporarily attached to one or both DNA strands via hydrogen bonds or Van Der Waals interactions (Fig. 1.1). This happens thanks to high affinity between one of the TF's domains (conserved portions of the TF protein sequence) and the bound DNA region's motif (DNA sequence pattern). Binding can occur anywhere in the genome, including promoters (a region just upstream of a gene's beginning coding sequence) or enhancer sites (distal regulatory regions that also control levels of gene expression). TFs recognize sequence motifs characteristic of these sites, with varying degrees of specificity. Each of these motifs is represented by a $4 \times n$ position weight matrix (PWM) that denotes the experimentally-determined probability of each nucleotide in an n-long sequence that the TF typically binds to. Bound TFs can then either recruit RNA polymerase or prevent its binding to DNA, thereby altering transcription. Surprisingly, evidence indicates that not all bound TFs alter polymerase activity nearby. Consequently, the binding property of a TF is distinct from its activity as a transcriptional regulator [2]. Knowing how TFs change in activity provides a fundamental piece of the genetic transcription puzzle: the most immediate response to a perturbation.

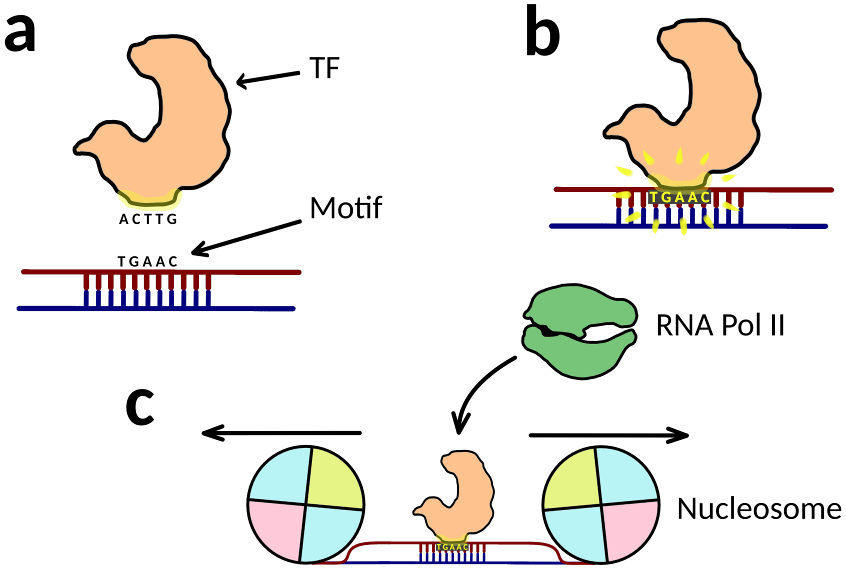


Figure 1.1: **TF binding to regulatory regions.** TFs recognize sequence motifs in DNA **(a)**, to which they can physically bind **(b)** and cause a displacement of nucleosomes, making the chromatin structure accessible and/or recruit RNA polymerase **(c)**.

Two important methods to understand the changes in regulatory activity of cells in different biological conditions are the analysis of differential gene expression, and differential TF activity. We present first an overview of differential expression, and proceed further in depth into differential TF analysis due to is complexity, principally from the lack of direct and relatively easy measurements of functional TF activity.

1.2 Differential gene expression

When a mRNA is present within a cell, it is said to be “expressed”. Expression profiles dictate most aspects of cellular behavior, such as cellular response to a stimulus or differentiation from a stem cell into specific cell types that eventually constitute the different tissues of an organism. The levels of expression vary continuously over time, and can present a very large dynamic range. These levels are crucial to quantitatively compare cellular behaviors, and the analysis cannot be restricted to those RNA species displaying multiple orders of magnitude more than others, since very low levels of expression of certain proteins acting in concert can also be biologically significant.

The most common method to understand the changes in cellular behavior after a biological perturbation, such as exposure to a small molecule, environmental changes, or comparisons between the same cells at different stages, has historically been a differential assay. This technique performs a high-throughput assay at two or more conditions, and examines the differences in results between both experiments. An example of this approach would be to quantify translated RNA first in unperturbed cells cultured in vitro, then again on the same cells 30 minutes after exposure to a drug, and contrast the differences in abundance levels of detected mRNA at target genes or genome-wide. This technique is of course not limited to gene expression, and can also be applied to other high-throughput assays. Another example of a differential assay could be to examine chromatin accessibility at these same two conditions, and compare the co-localization of accessible chromatin peaks to regions of interest, genome-wide. Understanding the changes of expression or chromatin accessibility could result in a way to understand responses to cellular perturbations.

Many techniques allow to interrogate the expression levels of the tens of thousands of distinct RNA species in a cell, each with a certain degree of noise and bias. Two of these tools commonly used to examine gene expression (i.e. the collection of mRNAs present in a cell) are hybridization-based microarrays[3] and RNA-seq[4]. Microarrays were the earliest developed method, which utilized a collection of probes to which the captured single-stranded RNA would anneal. A limitation of microarrays is its dependence on pre-defined probes, since one can only detect what is specifically being looked for. Hybridization techniques also suffer from a lower dynamic range than the sampling-based short read sequencing approaches. In various fields of research, microarrays have been largely replaced by RNA-seq, a method to capture available mRNA allowing to map where these reads came from, genome-wide. Despite their technical differences, both assays essentially measure the quantity of available translated RNA in cells. These assays are also usually referred to a "transcriptomics", a misnomer since they are steady-state measurements of mature RNA rather than point-in-time transcription readouts.

One of the most common methods to perform differential gene expression is to compare the expression levels of mRNA across two conditions, and perform a statistical analysis to determine the significance of this fold-change. It's worth pointing out that fold change alone is not a reliable method to determine the most significant changes, and should only be used when no experimental replicates are available. Limma[5] is currently one of the most popular tools to compare the output from two sets of gene expression replicates. When used in combination with robust multichip analysis (RMA), which assumes each readout is a combination of the true signal and Gaussian noise, one can detect more accurately which are the genes that present the most significant changes in expression.

1.3 Differential TF activity

Transcription itself is a heavily regulated process where TFs play a crucial role in altering RNA polymerase activity. An organism requires a sizable set of proteins in widely varying abundance levels, across hundreds of cell types and at many different stages of development, as well as in response to external or internal stimuli. Cellular functions are controlled by a feedback network involving over a thousand TFs, which influence the transcription of themselves, other TFs, and other types of proteins that participate in the regulatory process. The effect of a TF is further controlled by factors within the vicinity of that TF, including the motif of the TF's binding site on a DNA strand, its copy number, the chromatin accessibility of that site, and histone methylation.

The mechanism by which each TF alters transcription is still not entirely understood, and is believed to be influenced at least partially by DNA accessibility. It is currently unclear whether the activity of some TFs depends on an open chromatin state, or actively contribute to the current chromatin conformation. There is, however, a strong correlation between TF binding and chromatin accessibility, and the stronger the chromatin accessibility signal is, the stronger its association with more frequent TF occupancy on that region.

Besides binding to DNA with certain motif specificity, TFs recruit critical components of the transcription machinery, modify these components, or open the chromatin structure for exposure to enzymes responsible for transcription. The details of the act of binding itself are not entirely characterized, either: only a small fraction of sites matching a TF's sequence motif are actually occupied any given time, regardless of cell type[6]. Environmental conditions and cell type are large determinants of which of these sites are bound. From those that are bound, only a fraction alter transcription (according to reporter assays). To actually act as part of an enhancer complex and strongly contribute to changes in gene expression, a TF generally requires cooperative action by other nearby-bound TFs and cofactors (proteins that do not bind DNA directly but rather associate with TFs, and are crucial in forming the resulting regulatory complexes). Grossman et al showed how disrupting just one of the TF motifs associated with a regulatory complex had a strong effect on transcription[6]. Furthermore, a TF's binding activity and enhancer activity may be regulated independently.

1.4 Inference of TF activity by changes in binding

The desire to capture which TFs have altered function after a particular biological perturbation has been a subject of research for several years. Unfortunately, just capturing TF protein abundance in the cell nucleus is not sufficient to detect changes in activity[7]. An early assumption was that we could study changes in TF activity by comparing TF binding quantitatively between two conditions. The most common method to study the binding of a particular TF has been the use of chromatin immunoprecipitation, followed by sequencing (ChIP-seq). ChIP-seq utilizes a high-affinity antibody that binds to the TF in question, and after cells undergo lysis (breakdown of cellular membrane), the TFs are precipitated out of solution using these antibodies and separated from the rest of biological material. The fragments of DNA attached to these proteins are then amplified via polymerase chain reaction (PCR) and sequenced. We can then calculate a TF's "occupancy" based on the ChIP-seq coverage across all sites matching its motif, genome-wide. The ENCODE project[8, 9] performed hundreds of these experiments in vitro on a variety of cell types, and provides an open-access database of each of these samples referenced in the Sequence Read Archive (SRA). A change in the number of sites a TF is bound to, between two biological conditions, can be used to infer changes in that TF's activity. Steinhauser et al[10] provide a throughout review on differential ChIP-seq analysis tools.

There are several challenges regarding this approach. High-affinity antibodies are not necessarily easy (or possible) to obtain for every protein. We could also be precipitating a protein that is acting as a cofactor and not actually bound directly to DNA. This assay also implies we have a prior knowledge of the TF in question, and only allows for one TF to be inspected at a time, easily becoming too time-consuming and cost-prohibitive to run for most known TFs on each biological condition. Thus, any technique that uses a single assay per condition to capture all changes in TF activity genome-wide, may be preferable in many scenarios. A recent review[11] listed a total of 1,639 known human TFs, 1,107 of which have been verified by at least one form of experimental evidence.

Finally, the fact that a TF is indeed bound does not directly indicate it's effecting a change. Most TFs have a binding domain and a trans-activation domain, which makes binding just one of the TF's functions. Therefore, binding is not informative by itself as to whether the TF's trans-activation domain is truly functional. The definition of what "functional binding" implies is still under intense debate. For the purposes of this review, we will consider functional binding of a TF when it binds to DNA at regions matching a certain motif for which it has affinity, and either 1.

recruits RNA polymerase, or 2. modifies certain components of the transcription machinery, or 3. opens the chromatin structure to make that region accessible by polymerase or other TFs, as it is speculated for "pioneer"[12] TFs.

1.5 Inference of TF activity by changes in expression

In order to study changes in TF behavior genome-wide, attention shifted to using expression changes as a proxy for differential TF activity. Segal et al[13] presented one of the earliest methods of differential TF activity analysis, utilizing microarray technology. Given that TFs alter gene expression, this was an attempt to infer changes in TF activity from changes in expression profiles. This assumed that the regulators themselves were transcriptionally regulated, and these changes in expression must be detectable (a fundamental condition for various other subsequent methods). This pioneering technique used a combination of yeast TF motif data and ontology annotations to create "modules" of genes. The concept of modules refers to groups of genes that are related (regulated in a coordinated manner) to a common biological function, like "galactose metabolization". The possible transcriptional behaviors for each module are modeled as regression trees, and the most appropriate tree for each module is selected via the expectation-maximization algorithm. Unfortunately, in this implementation the modules were mutually exclusive (each gene could only belong to one module), which was a limitation of this approach as there exist relations between these functional units in all organisms.

A method related to Segal's module networks approach was the TELiS database[14]. Here they also used microarray data, plus a database of putative TF binding sites calculated a priori. This analysis was promoter-centric. TELiS attempted to identify the motifs at the promoter sites of genes that were significantly up/downregulated in a differential microarray experiment. Based on these promoter regions, they attempted assertions of which TFs may have been more (or less) active. They used a population-based z-test statistic (as opposed to the traditional t-test), which was a step in the right direction as it more accurately predicted differences in activity at TF binding sites. They however relied only on fold-change as the measure of change (i.e. this type of analysis was only of total abundance of RNA), when there are other important factors like the number of sites involved in this change.

The idea of representing the many changing factors in transcription as matrices in a linear combination inspired tools like ISMARA[15], a linear model that performed network inferencing from gene expression data, from microarrays or RNA-Seq (or even ChIP-Seq), to infer ab initio the activity of principal TFs or miRNAs at each sample. The tool also generated hypotheses about the potential regulatory roles of the highlighted TFs. It represented the measured transcription signal as a matrix of promoters by samples, which was the result of a linear combination of binding site matrices (promoters by motifs) and motif activity matrices (motifs by sample), plus a noise component that will vary by experiment. The "signal" depended on the experiment: number of reads around the transcription start site (TSS) for ChIP-Seq, number of reads for each transcript for RNA-Seq, or probe intensity for microarray. This model used singular value decomposition (SVD) to determine how the different motifs contributed to the observed transcription signal. Just as many other methods, it employed a curated list of regulatory motifs and genome-wide promoter annotations. ISMARA was focused on promoter regions and thus ignored distal enhancers, which was one of its limitations. Another limitation was the linear representation of the signal to TF activity relation. Some scenarios presented complications, as in when multiple TFs would participate in the same regulatory event. It also assumed that the TF activity of activator or repressor is mutually exclusive, when in reality it could be context-dependent for certain TFs.

Similarly to ISMARA, and based on the assumption that if TFs regulate genes, they should bind within a certain proximity of promoter regions of said genes, TIGERi[16] implemented a probabilistic model built from a linear model proposed by Sanguinetti et al[17]. Here a set of gene expression measurements can be explained as a linear combination of a binary matrix representing the TF to gene mappings, a weight matrix representing the TF-gene interaction "strength", and a vector representing the concentration of each TF in the cell. This approach combined microarray data with a set of TF-to-gene relations (based on the simplification that assumes regulatory motifs are upstream of promoter sites, up to a fixed, arbitrary kbp distance), to infer the TF concentration levels and their respective “weights” that indicate how much they affect the regulation of a particular gene. In other words, they essentially attempted to infer TF activity from gene expression data, and the presumed TFs that regulate each of these genes.

A drawback of the common proximity assumption for gene regulatory regions is that many genes can be regulated by TFs binding to distal enhancer sites, which would fall outside the fixed window utilized to infer TF-to-gene relations. For more information on other TF activity inference approaches from expression data, see Bussemaker et al[18].

The general approach described in this section is not limited to expression data from microarrays. Other studies utilized tools like RNA interference (RNAi) to block the activity of certain TFs and study the co-expression scores of pairs of TFs[19] from a large expression profiling dataset using a worm model (Caenorhabditis elegans). RNA interference works by introducing synthetic RNA in the cell that hybridizes to its target (binds to proteins with an exposed domain featuring a complementary sequence to our synthetic RNA strand), thus blocking the target protein’s function. The mutated C. elegans strains used in this study expressed green fluorescent protein (GFP) for each of the various TFs analyzed. By trying to silence one TF at a time using RNAi, and using a rank-sum test, they determined which TFs were more significantly co-expressed. This information of TFs presumably acting in concert was then used to generate hypotheses about the possible regulatory network at play for a given set of experiments.

1.6 Inference of TF activity by changes in nascent transcription

An important caveat to gene expression assays is that they provide a steady-state snapshot of the regulatory network. This is akin to measuring the level of coffee in a coffee pot in a break room at some point during the afternoon: it has been partially emptied and refilled several times on a given day before our measurement took place. The quantities of mRNA obtained are the result of fluctuations in transcription (coffee being served and refilled, in our previous analogy), the availability of certain types of biomolecules during translation, post-translational modifications, and degradation. Consequently, changes in expression do not necessarily reflect changes in transcription directly.

Nascent transcription assays, however, are considered genome-wide true assays on transcription activity, and a new generation of methods shifted to using these techniques. Nascent transcription assays consist, in general terms, in isolating active polymerase activity and capturing the RNAs that are currently being transcribed. After mapping these reads to a reference genome, strand-specific "signal" can be analyzed for transcriptional activity. A true "differential transcription" analysis would focus on the nascent transcripts that are being synthesized by RNA polymerase in each experimental condition. Of course no protocol is ever perfect, and one caveat about nascent transcription assays is that they are rather time-consuming and prone to noise (expected when handling short, nascent RNA transcripts). A novel attempt to predict nascent transcription I developed, from a much simpler protocol (ATAC-seq), is described in Chapter 2 and Appendix B.5.

After the discovery that TFs functionally bound to DNA recruit RNA polymerase and, without apparent strand specificity, polymerase produces short, unstable transcripts known as enhancer RNAs (eRNAs) which are markers of TF activity[20, 21], a new wave of tools was published that made use of this feature. This behavior of RNA polymerase results in the synthesis of short RNA transcripts on both directions from the TF motif site (usually referred to as "bidirectional transcription"). The utilization of nascent transcription assays like GRO-seq[22], PRO-seq[23] or GRO-cap[20] allows for detecting these offset bidirectional peaks around a midpoint, a mark for putative functional TF binding sites. Furthermore, this feature is also present at the promoter region of protein-coding genes, marking the "initiation" phase of gene transcription (Fig. 1.2).

Two main tools were developed to make use of nascent transcription data to detect bidirectional transcription activity, subsequently linked to different aspects of TF activity. One of them was dREG [24, 25], which focused on detecting regulatory elements like promoter or enhancer regions. This approach uses support vector regression (SVR) to analyze a large training dataset of 50bp-wide intervals across the entire genome, using nascent transcription datasets for the same cell line (K562) but generated by different labs and researchers, to make the classifier more generaliz-

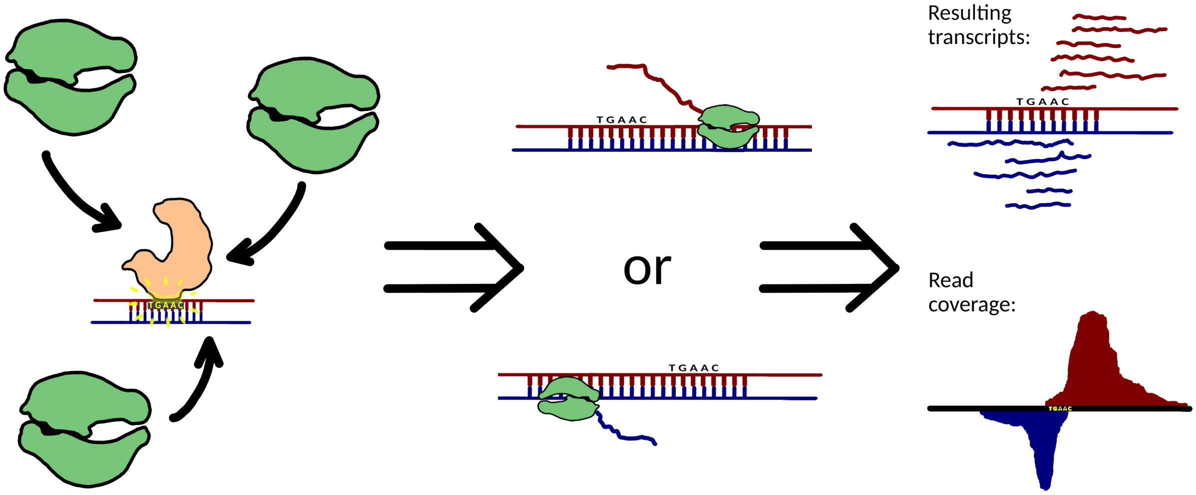


Figure 1.2: **Transcription initiation phase.** Upon binding, RNA polymerases are recruited to the motif site, which can bind on either strand and synthesize short, unstable RNA transcripts (known as “enhancer RNA” or eRNA). The reads captured by nascent transcription protocols are then mapped to the reference genome of the organism studied. The coverage in the vicinity of functional TF binding (a histogram of mapped reads at a single nucleotide resolution) displays a characteristic “bidirectional” peak shape.

able. Wang et al[25] focused the machine learning classification on negative examples (those that do not denote RNA polymerase activity) for the latest version of dREG, since they represent the majority of cases in real training data. Using a scoring scheme based on the number of reads in various bins around the putative peak's midpoint (dREG score), which follows a Laplace distribution, they use false discovery rates for their hypothesis testing framework to determine whether a peak is “positive” (i.e. classified as an enhancer site).

The other tool that made use of nascent transcription data was Tfit[26], which is a probabilistic mixture model of RNA polymerase II activity. This particular model uses expectation-maximization to estimate the parameters of a mixture model describing the different types of RNA polymerase II behavior (binding strand selection, transcriptional initiation, elongation, etc). The mixture it attempts to fit for bidirectional signal is of a Gaussian with an exponential function (an ”exponentially-modified Gaussian”), which mimics the shape of nascent transcription bidirectional signal on each strand, surrounding a TF motif site. For gene transcription, it incorporates a uniform function in the mixture to mimic the steady transcription that follows RNA polymerase

II loading at transcriptional start sites. This program can be combined with the Motif Displacement Score (MD-score) statistic[21], to use the regions of detected bidirectional transcription to predict changes in TF activity. The MD-score represents a measure of signal peak co-localization to recognized motif sites across the genome, for each distinct TF. The particular feature of offset bidirectional peaks around a center region can then exploited by this approach.

1.7 Beyond ChIP-seq and transcription: using chromatin accessibility and footprinting

For a TF to bind to DNA at a region matching its motif of affinity, we can assume that either the region must be accessible to the TF (i.e. not protected by nucleosomes or other proteins, as those from the nucleotide excision repair mechanisms[27]), or that the TF has the ability to open the chromatin structure to facilitate its own binding and other TFs' as well. This needs to happen because, simply put, the TFs and the proteins composing the chromatin structure can't occupy the same physical space. It also concerns steric hindrance at specific atoms forming the TF molecule (the congestion caused by the physical presence of ligands surrounding an atom, which blocks or at least slows down reactions at that atom). It is then a sensible choice to inquiry at regions of accessible chromatin for hints of where TFs may be bound, looking at their many target regulatory sites. The use of DNase I hypersensitive sites (DHSs) followed by sequencing (DNAse-seq[28]) was one of the first attempts to capture specific chromatin positioning (and indirectly, putative TF binding) genome-wide with a single assay, and marked a new way of approaching differential TF analysis. These DHSs are chromatin regions which are very sensitive to cleavage by the DNAse I enzyme. The latest generation of accessibility analysis is focusing on a different technique, the Assay for Transposase-Accessible Chromatin, followed by sequencing (ATAC-seq[29]). This experiment involves a mutated, hyperactive transposase to detect all open chromatin regions genome-wide. A clear advantage to using ATAC-seq is its simplicity, short duration and significantly smaller cell count requirements than nascent transcription assays.

DNAse-seq opened the door to leveraging the detected chromatin positioning for the identification of TF binding sites, from a single experimental high-throughput assay, before nascent transcription assays were developed. One of the earliest tools to make use of this feature was CENTIPEDE[30], which used a Bayesian mixture model to infer the likelihood of sites following a particular motif to be bound by a TF. This approach utilized regions overlapping ChIP-seq peaks for the TF in question as ground truth. Using the PWM match score for a particular motif as a prior, they estimated the probability of occupancy by a TF from the chromatin accessibility assay data (e.g. the number of reads around a candidate binding site). Therefore, for each putative motif site, CENTIPEDE calculated the probability of the observed measurements given the PWM-based prior. They noticed the distribution of DNAse-seq reads along each base position in the evaluation window was highly informative of TF binding. These characteristic shapes are known as DHS "footprints", referring to the signal flanking both sides of a putative TF binding site. This footprint also appeared to increase as ChIP-seq read depth increased.

A tool called Protein Interaction Quantitation[12] (PIQ), also provided the probability of TF occupancy of every putative TF motif genome-wide, based on DNAseI footprints. PIQ used PWMs for various motifs to scan for potential binding sites, then calculated a background model based on all DNAseI footprints, to "smooth out" the footprint signal. Using expectation propagation (a Bayesian machine learning technique to approximate complex probability distributions), they estimated regions of TF binding. An important aspect of this tool is that it demonstrated an in-depth analysis of "pioneer" vs. "settler" TFs, based on whether they were shown to open chromatin or required accessible, exposed DNA to bind in the first place. Part of the calculations thus also included a "pioneer index", as well as helper "chromatin opening index" and "social index" metrics. Contrary to previous approaches that were centered around promoters, PIQ excluded any regions adjacent to TSSs to avoid the bias typical of nucleosome-free regions around promoters. It's worth noting that those TFs that do not directly bind to DNA, but help via secondary binding (or those that provoke chromatin inaccessibility), will not be captured by this method.

An algorithm called Bivariate Genomic Footprinting (Bagfoot[31]) also relied on DNAse chromatin footprints at the motif site (indicative of direct binding) and at the regions flanking a motif (to account for interactions with other proteins that result in TF binding), which was helpful for footprint-lacking TFs. After bias correction, the genome-wide footprint depth (FPD) was calculated for each motif, by comparing the footprint "depression" regions to their flanking signals. Another statistic used was the flanking accessibility (FA), focused on the region 200bp on each side of the motif center. This bivariate data (FPD and FA) was displayed using a "bag plot", which could be thought of as a two-dimensional box-and-whiskers plot, on top of a scatter plot for all TFs. The ratio of footprint depth and flanking accessibility was the statistic used to predict changes in TF activity, comparing both biological conditions using a Chi-square distribution, and picking up outliers with a two-sample t-test. After studying the enzyme cut bias for DNAse-seq, they came to the conclusion that most of the motifs actually do not show a measurable footprint. This issue is combined with the fact that some TFs are bound to DNA very briefly (order of a few seconds), so they are extremely hard to detect via a DNAse footprints[32].

Bagfoot was written predominantly to be applicable to DNaseI but, since its publication, the majority of chromatin accessibility data has been focused on ATAC-seq. This made me wonder whether a difference in accessible, open chromatin (rather than footprinting) could be informative in changes of TF activity. Based on the observation that functional binding (detected by GRO-seq offset bidirectional signal, discussed above) overlapped peaks of open chromatin at close proximity, I leveraged a statistic developed for nascent transcription data[21] to capture differences in TF activity based on changes in chromatin accessibility. This motif displacement statistic provides in this case a measure of co-localization between ATAC-seq peaks and putative TF binding sites, genome-wide. To make this type of analysis widely available, I developed a publicly-accessible tool called the Differential ATAC-seq Toolkit (DAStk[33]), which is described in detail in Chapter 2 and the published manuscript in Appendix B.3. Whether all TF activity can be captured via open chromatin regions is still an open question.

1.8 Seeking concept enrichment from differentially active genes or TFs

All of these differential analysis tools provide a list of biological entities (genes, TFs or other proteins) that behave in a significantly different manner between two conditions. That list provides, however, only half of the information. The natural follow-up question to the subset of biological entities that have been up- or down-regulated is: What does this imply, in terms of cell functions? What higher-level biological process can explain this difference? In the particular case of nascent transcription, changes detected after a particular perturbation are so fast that they have to be primary (e.g. mechanistic) in nature. Now that we have a better picture of the dynamics of gene or TF activity and interactions, could we expand or validate the inferred mechanistic causes of these changes? As Callahan et al once cleverly stated, “There is an acute need to create tools for thought”[34]. The most popular way to generate hypotheses from the results of differential assays is to use domain-specific ontologies. There have been three major categories of ontology concept enrichment used to generate hypotheses from the experimental results: over-representation analysis, functional class scoring, and pathway-topology based enrichment. For a comprehensive review of these approaches, see Khatri et al[35].

A very common approach to date is still to seek enrichment of Gene Ontology[36, 37] (GO) terms, by detecting which ontology concepts are over-represented among the list of concepts denoted by significantly altered genes, based on a statistical test (generally Chi-squared, hypergeometric or binomial) at an arbitrary p-value cutoff. There are two concurrent versions of GO, full and filtered. The latter doesn’t contain any ”has part” or inter-ontology relations. Most post-data analysis enrichment tools use only the filtered version, so all these relations are generally missed. It’s also worth noting that the evidence code for the vast majority of relations is IEA[38] (computationally-inferred, and not curated by a human).

A general issue with all the approaches described here is that they require a list of differentially-expressed genes, not TFs. The public knowledge of the mapping from TF to genes is slowly improving, but we don’t yet have a comprehensive view of which of all TFs regulate which genes.

Some of the approaches to finding these mappings resort to looking for the first gene downstream of the putative TF binding site, up to an arbitrary distance cutoff (e.g. 10kbps), to guess the TF's regulation target. However, these regulatory regions could be far more distal from their gene targets. Moreover, most of these regulatory processes involve a combination of multiple TFs, which makes this a harder problem, as a single TF may participate in more than one complex, regulating more than one target gene. Other issues with mere over-representation include the assumption that these genes are independent (as well as the pathways they participate in), and that the "intensity" values for each gene (microarray probe intensity, high-throughput sequencing depth, MD-score) are often ignored.

The original independence assumption of ontology concepts ignored its hierarchical nature. Concepts closer to the root of an ontology tend to carry a lower information content, and there are multiple dependency pathways between some of the concepts. TreeHugger[39] took the hierarchical structure of ontologies in consideration, and assigned scores to concepts by weighting whether genes reference them directly or via subclassed children nodes. Using a t-test and Welch-Satterthwaite correction for unequal variances, they determined the most significantly enriched GO concepts. They still assumed that gene expression profiles were independent, which is likely not in concordance with the underlying biology.

GSEA[40] was a step in the right direction, by considering modest changes of groups of genes that are functionally related. In order words, GSEA measured the proportion of differentially-expressed genes in pathways, and how much correlation there was between these genes. Instead of performing the typical hypergeometric test and looking for GO enrichment on the top N genes (or those beyond some arbitrary p-value threshold), they concentrated on groups of genes that varied together significantly. The main idea behind this was that a few concerted changes in activity, varying a small percentage, would likely be more significant than a single gene varying manyfold in activity. For each of the gene sets (metabolic pathways, transcriptional programs, stress responses, etc) defined in their database, some manually-curated and others computationally-inferred, they attempted to determine whether the members of that set are randomly distributed throughout a ranked list of genes from a differential assay. Tools like GSEA present their own challenges, e.g. the pathways are analyzed independently, when in reality many genes are involved in more than one pathway. The individual gene "intensities" are still ignored within the same pathway, when different fold-changes should probably be weighted differently.

The current generation of enrichment and hypothesis generation tools take the network topology of ontologies into account for their analysis. These are similar to the functional class scoring methods, however they use the pathway topology from sophisticated knowledge-bases to compute the statistics for each gene. For example, ScorePAGE[41] adjusts the weight for each pair of genes based on the number of biochemical reactions needed to connect them in a pathway. Another tool, NetGSA[42], takes each gene's baseline expression into account and considers the change in network structure between biological conditions of the differential assay used (as sometimes the pathways will differ due to certain perturbations). The Signaling Pathway Impact Analysis (SPIA[43]) method incorporated information about the topology of pathways, to evaluate the significance of a particular pathway's enrichment in differential expression assays. This means that besides considering different genes in the context of a pathway (represented as a tree), it also accounted for where in the pathway they appeared, with changes in leaf nodes being less "disruptive" than changes in a node higher up the hierarchy.

MasterPATH[44] is a recent pathway enrichment algorithm to generate hypotheses of what molecular pathways may be at play, given a list of differentially expressed genes. This particular approach uses a combination of public databases of interactions between proteins, DNA and other molecules. It collects the shortest paths connecting all the genes involved and performs a large number of random iterations of differentially enriched genes, to detect how likely it is to result on each given pathway. The current shortcomings of these most sophisticated topology-based approaches are related to tissue specificity: many of these biological pathways are cell-type dependent. These models still don't consider interactions between pathways.

An interesting alternative to the use of ontologies was ChEA[45]. This was an attempt to obtain enrichment of TF targets using the list of genes obtained from differential expression experiments, using a manually-curated database of ChIP-related assays (ChIP-seq, ChIP-chip, ChIP-PET and DamID, which they summarized as "ChIP-X") rather than ontology annotations. The over-representation of TF targets for the given gene symbols was determined by a Fisher exact test (Bonferroni-corrected). A problem with this approach was that proximity was assumed to be the criteria of which TF regulates which gene (which we now know is incomplete). This problem of TF assignment to genes by proximity of sequence motifs to their TSS is still a largely unsolved one. Many distal sites can also regulate gene expression thru TFs after chromatin loops that bring both the enhancer site and TSS in close physical proximity. This project later became Enrichr[46], another tool that combined the above with other data sources like TF PWMs, histone mark assays, the gene-set TF library from ENCODE[8], and the microRNA gene set library from TargetScan[47].

All ontology enrichment tools ultimately produce a list of (often very disconnected) concepts that are considered highly relevant to the experimental results. A follow-up question would be how could we evaluate the hypotheses generated from concept enrichment. HyBrow[48] was an early attempt to aid with hypothesis evaluation by creating a custom ontology for yeast biology, also incorporating expression data from microarray assays. Even though not explicitly using semantic web technologies, the processes were modeled as triples (i.e. "acting agent" via "relationship" relates to "target agent"). They used a context-free grammar to define restrictions and temporal conditions (e.g. "protein X can only be the actor of this relationship when located in the nucleus", or "protein X is phosphorylated by this biochemical reaction, so it will remain phosphorylated in every future interaction unless there is a dephosphorylation step").

HyQue[34] was the natural successor of HyBrow, this time using semantic web technologies and SPARQL query language to search a triple store, and still revolved around yeast galactose gene networks like its predecessor. It features data from more ontologies, and more granular scoring metrics There are still some shortcomings, as longer pathways with mediocre evidence might score better than a much shorter one with solid experimental evidence. These and other hypothesis evaluation tools appear to be very narrow in scope, highlighting the need for a more general knowledge-base of biology, or at least one within the human biology scope (and therefore significantly more complex than yeast-focused). Another aspect that none of these pathway enrichment strategies address is the sequential nature of experimental data (i.e. taking into account the order of gene expression changes when seeking enrichment of a pathway). Chapter 4 and Appendix B.6 describe a novel approach to incorporate the sequential order of experimental variables into pathway enrichment.

1.9 Current pitfalls and the state of mechanistic inference

There is currently a wealth of knowledge in seemingly disparate and disconnected sources about different areas of biology (genes, proteins, pathways, chemicals, and biochemical reactions, to name a few). Seeking ontological enrichment in an aggregation of this information could paint an increasingly complete picture of how the significant genes andor TFs detected experimentally are related, in comparison to just utilizing the GO, or the Kyoto Encyclopedia of Genes and Genomes[49] (KEGG). Many interesting questions can be asked about each subset of TFs from the experimental results. Do they belong to the same pathway? If so, which one is upstream of which? Do they interact with the same cofactors? Do some of them (directly or indirectly) participate in the same biochemical process inside the cell, or across cells via some type of signalling? Much of the latter also applies to differentially expressed genes. These are questions that are normally answered after an arduous literature review and database searches. I show a type of question that can be answered using this type of ontological integration in Chapter 3 and Appendix B.1: finding novel drug-drug interactions.

We are in need of a systematic inquiry into how the existing knowledge representation of molecular biology and the mechanistic evidence from high-throughput assays complement or contradict each other. Rather than performing the analysis of experimental results and resorting to a single one-size-fits-all ontology enrichment, it's time to develop novel pathway enrichment algorithms and to seek over-representation in densely interconnected sources of knowledge that span beyond genes and proteins (biochemical reactions, disease phenotypes, cell specificity, etc). Seeking enrichment via the TFs of interest directly, rather than through the genes they're expected to regulate, could result in concept enrichment that is of higher relevance to differential TF activity assays. Another unexplored application of knowledge-bases is that to inform experimental expectations. Given a planned differential assay for certain cell lines, perturbations, and time points, one could generate hypotheses of what are the expected genes to see up/down-regulated, which TFs are expected to be more or less active, or which pathways we expect to be triggered. A protocol for discovery of perturbation mechanisms in molecular biology could greatly benefit from a two-way communication between omics data science and symbolic knowledge representations, so that both types of analysis could validate each other in a principled manner.

A portion of the relevant information useful for mechanistic inference is still contained in the literature but not curated into regularly updated databases. For example, much of the database curation of chemical to gene/TF interactions is focused on therapeutics. A relation extraction study on open-access literature targeted to a controlled vocabulary of chemicals (e.g. metals, potentially toxic industrial chemicals), and TF names or synonyms, could enrich our symbolic representation of interactions and help confirm or dispute experimental results. Just predicting whether or not a particular chemical is known to be related to a TF, and whether this relation is either upregulation or downregulation would be immensely valuable. I present in Chapter 3 and Appendix B.2 two different strategies to predict five different types of chemical/protein relations (or no relation mentioned), using features derived from the available text itself, as well as from a knowledge graph.

Finally, and perhaps most importantly, the grand challenge of mechanistic inference in biology is still largely incomplete. This is in part due to our limited understanding of every process in molecular biology, but also due to our inability to represent our existing body of knowledge in a way that is easy to query computationally, and within a certain scope (i.e. tissue specificity, as not all TFs present the same basal level of activity on every cell type, nor do all molecular pathways behave the same across all cell types). Experimental tools like nascent transcription assays offer such a fine time resolution that their results are mechanistic in nature, and these assays performed as a time series could provide a window into a causal sequence of steps post-perturbation. Linear causal models applied to TF scores (determined from experimental assays like nascent transcription or chromatin accessibility) of the same cells at various time points could elucidate possible mechanistic hypotheses. A good example of this approach can be found in Bay et al[50], where they studied gene regulatory networks with photosyntesis regulation in cyanobacteria as the use case. We could derive Bayesian priors from a knowledge-base to aid the decision of which could be the chain of causal events with the highest likelihood.

The use of artificial intelligence (AI) to mechanistic inference dates back to Dendral[51] and Meta-Dendral. These tools for hypothesis generation generally had a narrow scope. In this case, Dendral and its companion software Meta-Dendral focused on mass spectrometry analysis. Using a combination of statistical analysis to pick the most relevant peaks from the fragmented mass spectra, and prior knowledge of chemistry to infer constraints (e.g. ”the unknown molecule is probably of class X but definitely not class Y”), it aided molecular structure identification of unknown molecules. It also attempted to create general rules by correlating different types of features resulting from the mass spectrometry output of each studied molecule.

An exquisite example of taking this approach full circle was the implementation of a robot scientist[52] that could generate hypotheses from expression data but also take action on the generated hypotheses (via abduction, a logical inference mechanism), and conduct further expression assays, with enough iterations to reach scientific discovery. Though very narrow on scope (it focused on certain pathways of baker’s yeast regulatory network), it successfully generated novel hypotheses that were later validated. This was not the first work on hypothesis generation from molecular biology assay results. Hypgene[53] was presented a decade earlier by Karp et al, which used their earlier prediction tool (Gensim) to iterate over the predicted and real outcomes from bacterial gene regulation experiments. Hypgene was successful at hypothesis formation, reproducing most of the discoveries of a 15-year research program.

A few recent developments in the representation of mechanistic knowledge could significantly aid the task of mechanistic inference. Darden et al recently presented a diagramamatic method[54] to represent biological mechanisms and MecCog[55], a formal framework to describe them. MecCog focuses on the connections of genetic variants to disease phenotypes, and provides a way to represent ignorance, ambiguity and uncertainty anywhere in the mechanistic pathway.

Symbolic artificial intelligence has a relevant application to mechanistic inference, as I demonstrated in Chapter 4, with the use of rule-based systems and a knowledge-base that integrates the many sources of information we have. There are currently enough datasets to develop an expert system for differential analysis with a narrow focus (e.g. identification of metal toxicology pathways), that can combine our existing knowledge of biological processes, the known properties of bio-molecules and drugs, and pathways involving them, to generate putative explanations for the experimental results. A possible approach consists of using differential analysis between transcription assays in a time series, to detect significant TF activity changes (or changes in gene expression) at each time point. Using those results and a rich knowledge-base, we can identify "themes" that would describe the possible mechanisms at play. Rather than a disconnected list of enriched ontology concepts, it would be more valuable to biologists to produce a series of mechanistic explanations for the resulting changes to a perturbation.

Chapter 2

Expanding the limits of information we can extract from ATAC-seq assays

ATAC-seq is an assay to detect open chromatin regions (OCRs) that excels at simplicity as well as time and cell count requirements, therefore it's worth pushing the limits of what can be inferred from OCRs detected using this protocol. This chapter briefly summarizes published works of ATAC-seq applications for differential analysis in section 2.1, and a technique that detects nascent transcription as well as histone modifications from OCRs denoted by ATAC-seq peaks in section 2.2. The full papers can be found in Appendix B. This chapter also includes unpublished work, first regarding clustering of ATAC-seq peak positioning patterns with respect to TF motif sites in section 2.3, and later an in-depth analysis of OCR sequence clustering in section 2.4.

2.1 Differential analysis from accessible chromatin regions

Much is still unknown regarding how specific TFs become functional in enhancing or blocking gene expression. We have reasons to believe some of them actively open chromatin to make a DNA region accessible to the transcription machinery, however this is likely not a requirement for most TFs (in fact, DNA binding sites matching certain motifs may not even be required to be nucleosome-free at all). Given that ATAC-seq excels at requirements of sample size, duration and simplicity, its application as a differential analysis tool is very attractive, and extends to the clinical realm. In the study included in Appendix B.3, we used the motif displacement score (MD-score) first presented in Azofeifa et al[21] as a metric to contrast TF activity between ATAC-seq assays. This result provided the ability to infer changes in TF activity from a simpler and more cost-effective protocol than any nascent transcription one. The difference in activity between two contrasting biological conditions was originally captured using a two-proportion z-test, and recently updated to a bootstrap approach for improved statistical significance. To conduct this analysis, I developed the Differential ATAC-seq toolkit (DAStk[56]), which is described in detail in Appendix B.3 with illustrative use cases from public datasets. While originally developed for ATAC-seq data, this tool has proven to be useful to capture differential TF activity in other types of assay pairs like nascent transcription (GRO-seq, PRO-seq), or ChIP-seq.

2.2 Detecting signatures in omics data

There is still much to be learned from ATAC-seq in terms of functional TF binding, as there could be features in the assay's resulting "signal" that could help us discriminate between accessibility peaks over a functionally-bound TF, and likely unoccupied (yet open) regions. We made available a first-of-its-kind attempt to apply signal processing principles[57] to ATAC-seq output using both wavelet analysis and recurrent neural networks (RNNs). The goal of this study was to predict which ATAC-seq peaks overlapped a region that was actively transcribed, and which ones overlapped a histone mark associated to transcription. Owing to my background in electronics (and a penchant for audio engineering and signal processing), I often wondered about ways to analyze the coverage from high-throughput sequencing as a digital signal you would obtain from any other instrument. This led me to envision processing the "signal" from these peaks (as well as the sequence at the same coordinates) to detect nascent transcription or histone modifications. I evaluated the performance of multiple machine learning classifiers (support vector machines, random forests, ADAboost, and RNNs) and feature sources (ATAC-seq signal within a fixed window at a single nucleotide resolution, the sequence in that same window using different nucleotide encodings, and a combination of both) to find the configuration that resulted in the optimal classifier performance. All details are specified in the preprint included in Appendix B.4.

This study required a major data processing effort using assays from multiple public sources. In order to process the vast amounts of data from different protocols and improve reproducibility, I designed and implemented pipelines using Nextflow[58] framework to process both ATAC-seq[59] and nascent transcription[60] data (such as GRO-seq, PRO-seq, etc). One of them[61] has become part of the official nf-core[62] bioinformatics pipelines, which feature unit testing and continuous integration, and are only published after a throughout code review.

A more in-depth analysis of one of the RNN-based models to predict nascent transcription from ATAC-seq data is currently under review and included in Appendix B.5. In this manuscript we focused on the machine learning classifier configuration that yielded the optimal performance, and explored the biological implications of our findings. Returning once more to audio signal processing analogy, and akin to interpreting the ATAC-seq peaks' "songs", we used the "music" (the coverage background as a signal) and "lyrics" (the underlying genomic sequence) to predict the presence of nascent transcription. This consisted in representing each ATAC-seq peak as a combination of nucleotide embeddings, concatenated to the signal level (the mean numbed of mapped reads) for each nucleotide. These findings can open the door to future exploratory analysis to detect signatures of other biological events in fixed-window regions of the genome.

2.3 Birds of a feather: How different TFs share common motif displacement profiles

A different subject I studied was the change in motif location patterns respective to ATAC-seq peaks. Using the FIMO[63] scanner and the consensus TF motifs from the HOCOMOCO[64] (version 11) database, I generated a list of all possible (yet high-confidence) motif binding sites for each TF. Then, for each TF I collected all ATAC-seq peaks whose midpoint was located within 500bp of each motif site. The histogram of all peak midpoints both upstream and downstream from a motif site, normalized between 0 and 1, is what we call a "barcode" (more visually intuitive if we plot the histogram as a heat map). After observing common patterns for different TFs, I decided to attempt clustering these barcodes that could reflect similar biological or molecular behaviors, due to co-localization pattern. Using the K-means clustering algorithm with Chebishev distance, I was able to distinguish 4 clusters of common ATAC-seq signatures. These very different histograms could reflect specific biological behaviors, which are yet to be determined.

When plotting the distribution of putative TF binding sites relative to every ATAC-seq peak as a one-dimensional heatmap (also known as ”barcode plot”, Fig. 2.1) we notice a few common patterns. When TFs are active, we see co-occurance of their sequence motifs with the midpoint of OCRs denoted by ATAC-seq peaks. On a histogram where a position of zero is at the middle (the OCR's center), and the number of nucleotides upstream/downstream indicates the evaluation window, this looks like a normal distribution. Some TFs, however, present a binding pattern that is much ”tighter” than the rest, which could be represented by a normal distribution with a much smaller standard deviation. A third observed ”offset” pattern consists of motif sites a certain number of nucleotides upstream and downstream of the OCR's center, but not directly on top. There is still no well-established biological explanation for this pattern. Finally, when a TF's motif displacement distribution looks like a noisy uniform distribution, that TF is generally considered to be inactive. I used ATAC-seq data generated in the Dowell lab from HCT116 cells treated with Nutlin-3a (a drug that targets MDM2 and, by disrupting the MDM2-TP53 interaction dramatically increases the activity of TP53) to calculate MD-scores and study the difference between barcode patterns.

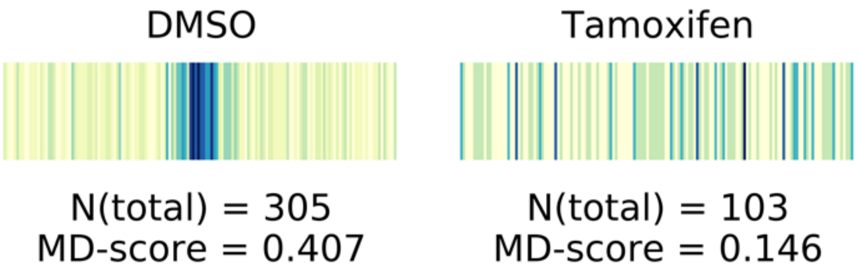


Figure 2.1: **Examples of barcode plots.** This alternative representation to histograms depict the density of regions of interest (e.g. ATAC-seq peak midpoints) relative to the motif location (at the plot center), representing a putative TF binding site. This particular example shows the distribution of binding locations for the PO5F1 motif for Mus musculus, in unperturbed conditions (left) and tamoxifen-treated cells (right).

Clustering histograms proved to be not a trivial task. There is existing research on this topic, generally linked to the area of image processing[65] (by clustering color channel histograms of images to identify similar ones). In this application, we are interested in clustering the histograms that represent each TF's binding proximity to a putative motif site. In order to find clusters of motif barcodes in a principled manner, I used K-means with four expected clusters, using the vector representing each motif's barcode as the features. Of all similarity metrics tested, Chebyshev distance resulted to be the most effective at discriminating among functional TF binding patterns. Four distinct patterns were identified by this algorithm (Fig. 2.2). Some ATAC-seq peak midpoints tend to appear at very close proximity to the TF motif sites (Fig. 2.2.a), and others appear within close proximity but at a higher variance of distance to the motif site (Fig. 2.2.b). Both of these cases correspond to TFs that are generally considered to be actively affecting transcription. This is also observed in nascent transcription data, where active TFs' motifs tend to have close proximity to the RNA polymerase II loading site (the midpoint of bidirectional transcription described earlier). It's worth noting how TP53 follows the first pattern mentioned, being specifically activated by Nutlin-3a, the drug used to treat these cells. The same is observed for CTCF, a known chromatin modifier, whose motif sites across the genome tend to co-localize at very close proximity to ATAC-seq peak midpoints. A third pattern is observed (Fig. 2.2.c), which corresponds to TF motifs that are not following any particular pattern, assumed to be the case for TFs that are not active in the current cell type and biological conditions. Finally, for a number of TF motifs a last pattern is clustered (Fig. 2.2.d), in which the location of open chromatin midpoints appears to be exclusively off-target, but at roughly the same distance in each direction. It is still unclear what is the biological explanation for this offset binding pattern.

2.4 Can OCRs be clustered by sequence bias?

A different unsupervised clustering task I researched was to attempt grouping nucleotide sequences corresponding to ATAC-seq peaks. The different types of numerical encoding I explored to cluster these sequences in a multidimensional space are listed in Table 2.1. For each of these encodings, I tested various dimensionality reduction techniques to visually detect any particular clusters among the genomic sequences. These algorithms were principal component analysis (PCA),

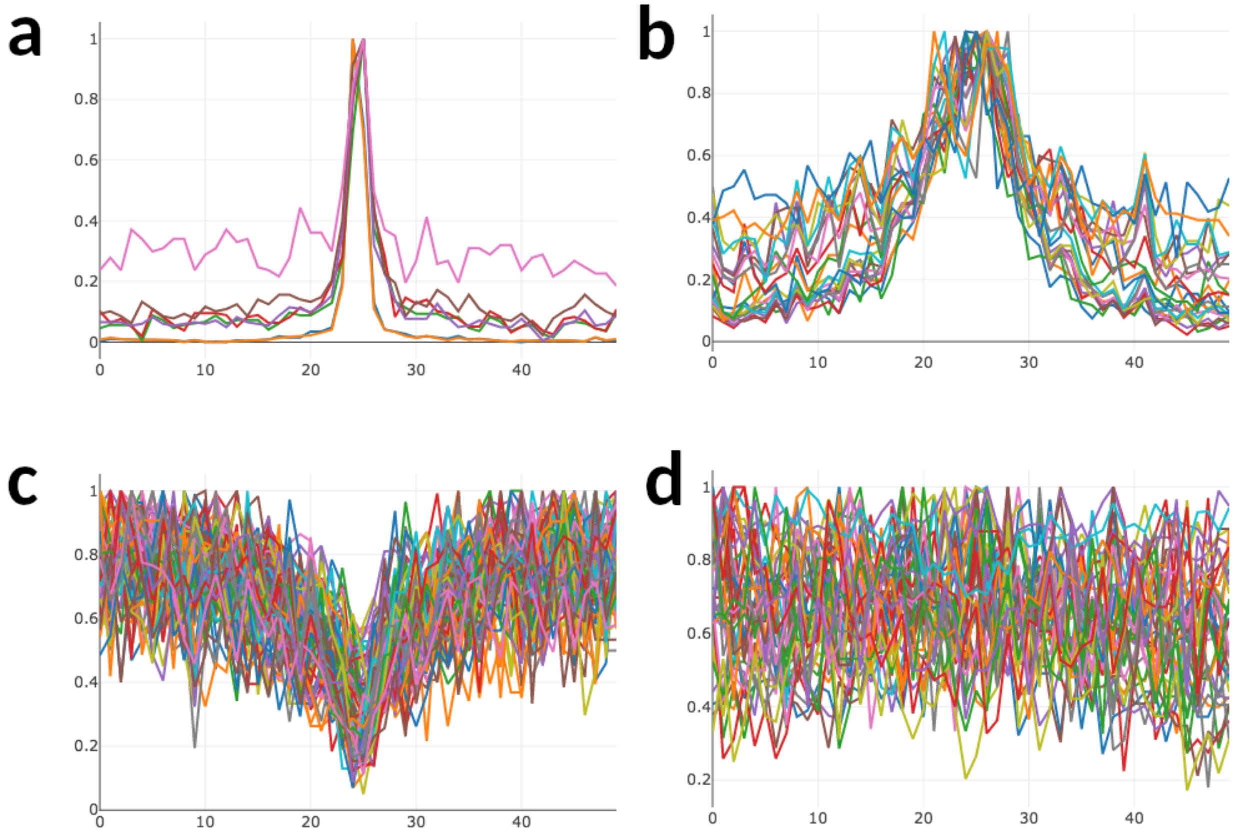


Figure 2.2: **Barcode clusters for different TFs.** Four clusters that illustrate the different TF motif co-localization patterns with respect to ATAC-seq peak midpoints.

T-distributed Stochastic Neighbor Embedding (t-SNE) with a perplexity hyperparameter of 10, and correspondence analysis (CA) where applicable.

The first step was to attempt clustering these sequences to assess whether we expected any clusters by chance alone, given the sequence bias observed at open chromatin regions (Fig. 2.3). To test this, I generated random sequences using a first-order Markov chain I trained on 400bp-long sequences centered at ATAC-seq peaks' midpoints, mapped to the GRCh38 reference human genome. This experiment yielded, unsurprisingly, no particular clusters in any of the tested configurations. With this result on the expectation from random sequences, I moved on to real sequences. I utilized 400bp-long sequences around the midpoint of ATAC-seq peaks in HCT116 cells in unperturbed conditions to attempt clustering real sequences. Unfortunately, no particular

Table 2.1: **Genomic sequence encoding types used.**

Description	**Values**
Basic (one number per nucleotide, no particular meaning)	A = 0 ; C = 1 ; T = 2 ; G = 3
Biochemical representation (relative weak or strong binding bias)	A = -1.5 ; C = 0.5 ; T = 1.5 ; G = -0.5
EIIP (distribution of free electrons' energies along the DNA sequence)	A = 0.1260 ; C = 0.1340 ; T = 0.1335 ; G = 0.0806
Atomic number	A = 70 ; C = 58 ; T = 66 ; G = 78
Random walk-like (moving to a G or C we add 1, and moving to an A or T we substract 1)	A = -1 to the last value ; C = +1 to the last value ; T = -1 to the last value ; G = +1 to the last value
GC bias	A = -1 ; C = 1 ; T = -1 ; G = 1
Di-nucleotide representation (each pair of bases is assigned a unique value)	AA: 0.0 ; AT: 0.06666667 ; AC: 0.13333333 ; [...] GT: 0.86666667 ; GC: 0.93333333 ; GG: 1.0
One-hot encoding (4-dimensional vector representation of each nucleotide)	A: $\begin{bmatrix}1\\0\\0\\0\end{bmatrix}$; C: $\begin{bmatrix}0\\1\\0\\0\end{bmatrix}$; T: $\begin{bmatrix}0\\0\\1\\0\end{bmatrix}$; G: $\begin{bmatrix}0\\0\\0\\1\end{bmatrix}$

clusters were detected for the real sequences, either. No obvious clusters were observed for any of the encoding strategies using PCA (Fig 2.4), t-SNE (Fig. 2.5) or CA (Fig. 2.6). This indicates that the sequence encoding and dimensionality reduction techniques employed were not sufficient to capture any particular trends among these sequences. In a previous study published as a preprint[57], we demonstrated that it was possible to discriminate among open chromatin regions (denoted by ATAC-seq peaks) that overlapped transcription from those that did not, by using only the underlying sequence. However, this approach employed an embedding of each nucleotide and a recurrent neural network.

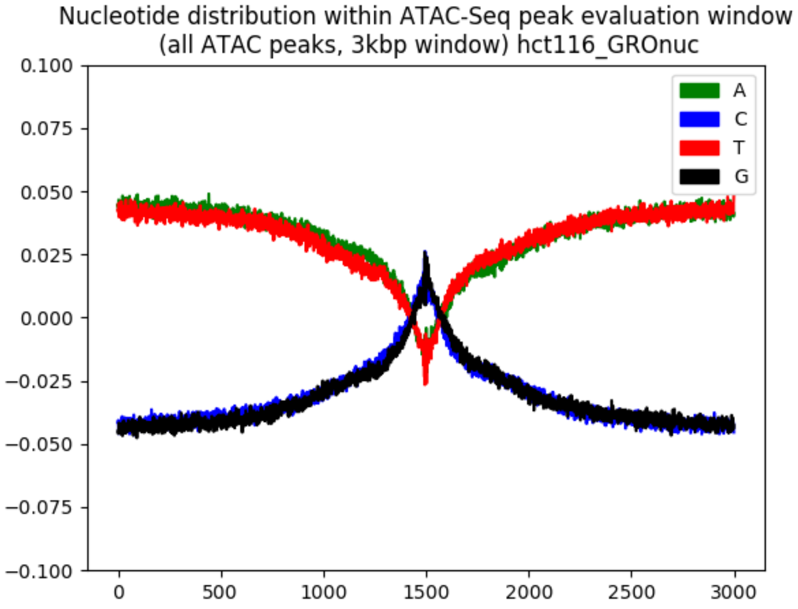


Figure 2.3: **Nucleotide bias at open chromatin regions.** We observe in this 400bp evaluation window the distribution of each nucleotide for every ATAC-seq peak in the HCT116 cells dataset. There is a clear difference in G/C concentration around the midpoint of all open chromatin regions, in contrast with A/T concentration.

2.5 In Summary

I demonstrated how we can detect differences in TF activity between conditions using only ATAC-seq data, by designing and implementing DAStk, a tool that is actively maintained and currently used by other labs across the world. DAStk can also be used to compare regions of interest from other types of data, as in detected bidirectionals from nascent transcription protocols. It is likely that for certain perturbations we may need to contrast both chromatin accessibility and nascent transcription. An in-depth analysis of the combination of nascent transcription and chromatin accessibility studies is needed, using a combination of statistical analysis tools and machine learning techniques, as both types of experiments may be required to gain a complete picture of TF activity at a given time point.

I have also shown how we can infer certain types of biological activity from regions of accessible chromatin determined using ATAC-seq, like histone modifications or bidirectional transcription.

To process the large number of datasets required for the studies mentioned in this chapter, I have created bioinformatics pipelines which I made publicly available to the scientific community.

2.6 List of publications and bioinformatics tools relevant to this chapter

Appendix B contains the collection of my first-author original research articles, pre-prints and conference proceedings mentioned in this thesis, which have been either published or are currently in the process of peer review. The manuscripts relevant to this chapter are listed below, in chronological order:

- I. J. Tripodi, M. A. Allen, and R. D. Dowell, "Detecting Differential Transcription Factor Activity from ATAC-Seq Data" Molecules, vol. 23, no. 5, p. 1136, May 2018.
- I. J. Tripodi, M. Chowdhury, and R. D. Dowell, "ATAC-seq signal processing and recurrent neural networks can identify RNA polymerase activity" bioRxiv, p. 531517, Jan. 2019.
- A. Pouikli, S. Parekh, M. Maleszewska, M. Baghdadi, I. Tripodi, C. Nikopoulou, K. Folz-Donahue, Y. Hinze, A. Mesaros, P. Giavalisco, R. Dowell, L. Partridge, and P. Tessarz, "Citrate carrier links intermediate metabolism to chromatin architecture and regulates osteogenesis in mesenchymal stem cells upon ageing" (currently under peer review)
- I. J. Tripodi, M. Chowdhury, and R. D. Dowell, "Combining signal and sequence to detect RNA polymerase initiation in ATAC-seq data" (currently under peer review)

My open-source bioinformatics tools and additional co-authored publications relevant to this chapter are listed below in chronological order:

- M. A. Allen, D. Thompson, K. McChesney, N. Parsonnet, I. J. Tripodi, and M. Melnick, "MyFavoriteTF: A web interface to identify transcription factor activity across cell-types" (2017), Website: `http://tf.colorado.edu/mytf/`, GitHub Repository: `https://biof-git.colorado.edu/hackathon/myfavoritetf`

- I. J. Tripodi and M. Gruca, "Differential ATAC-seq toolkit (DAStk)" (2018), GitHub Repository, `https://github.com/Dowell-Lab/DAStk`

- I. J. Tripodi, B. Busby, S. Tsang, J. Zhao, E. Floden, and C. Zhang, "ATACFlow: An ATAC-seq pipeline wrapped in NextFlow that can be run by Jupyter (ATACFlow)" (2018), GitHub Repository: `https://github.com/NCBI-Hackathons/ATACFlow/`

- I. J. Tripodi, M. Gruca, Z. Maas, "Nascent-Flow: Nextflow Implementation of the Dowell Lab Nascent Pipeline" (2018), GitHub Repository: `https://github.com/Dowell-Lab/Nascent-Flow`

- I. J. Tripodi and M. Gruca, "nf-core/nascent: Nascent Transcription Processing Pipeline" (2019), GitHub Repository: `https://github.com/nf-core/nascent`

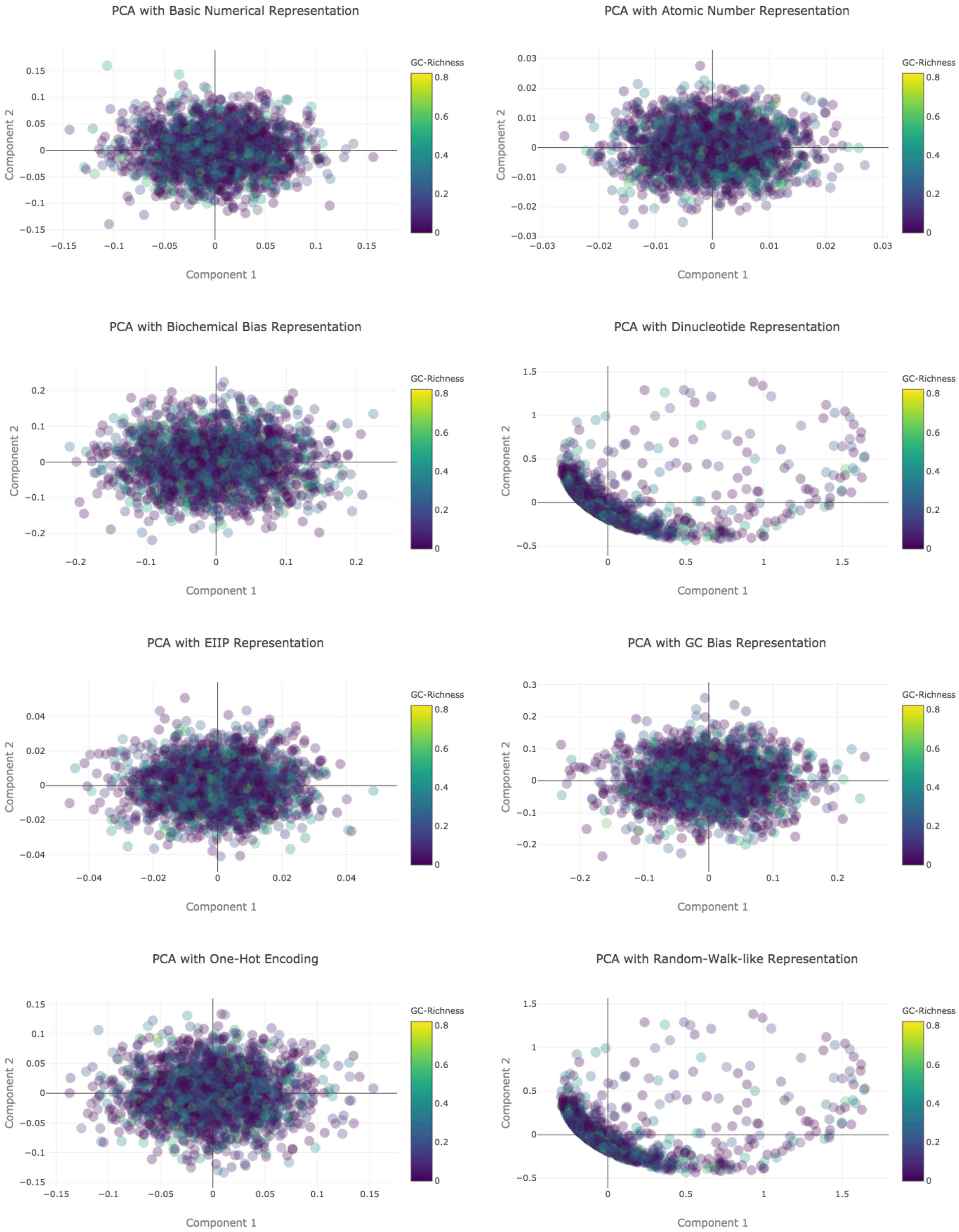
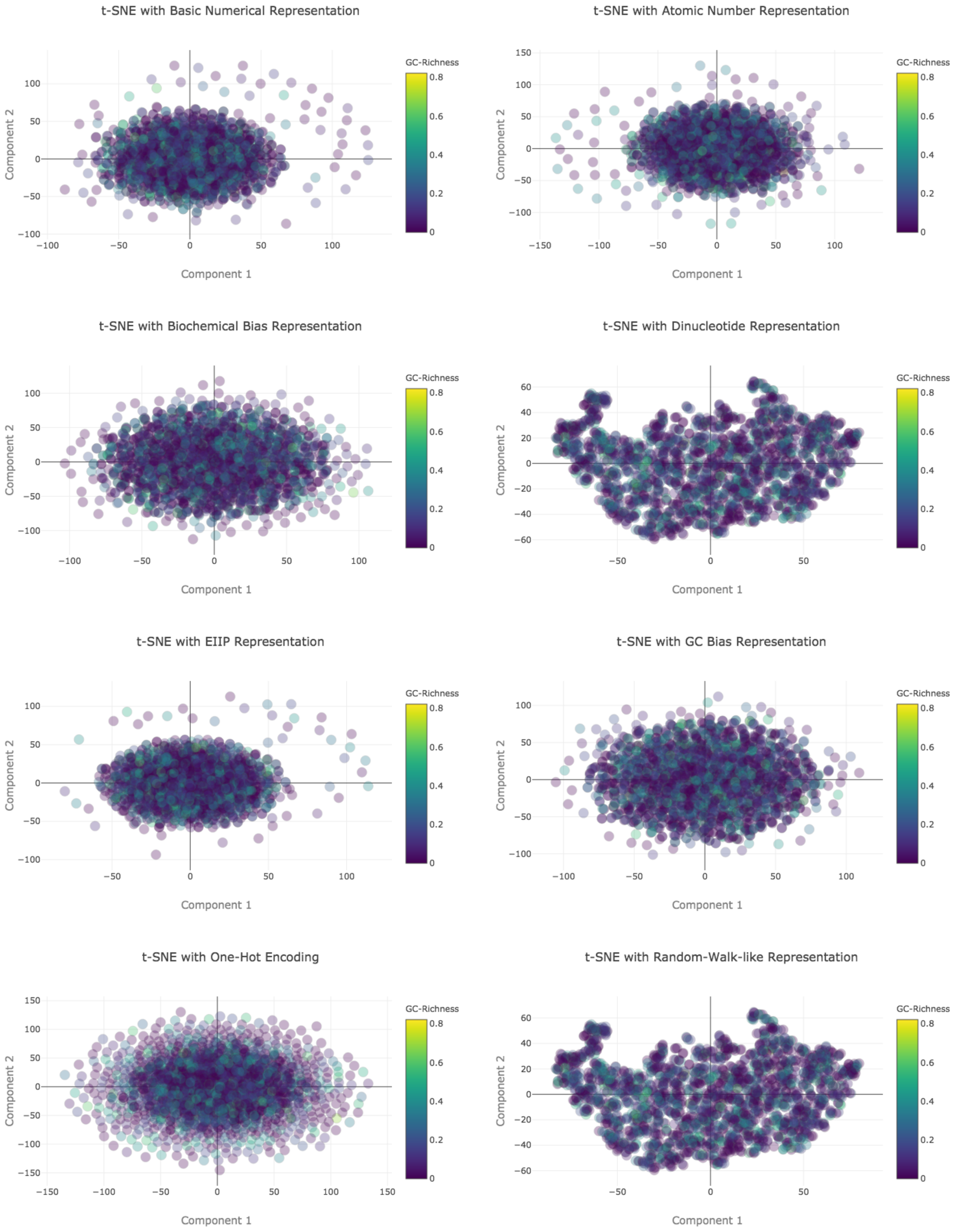


Figure 2.4: **PCA plots for different sequence encoding strategies.** Each dot represents a 400bp-long genomic sequence at the center of an ATAC-seq peak. Dots were colored by their sequence's GC-richness.



t-SNE with Basic Numerical Representation

t-SNE with Atomic Number Representation

t-SNE with Biochemical Bias Representation

t-SNE with Dinucleotide Representation

t-SNE with EIIP Representation

t-SNE with GC Bias Representation

t-SNE with One-Hot Encoding

t-SNE with Random-Walk-like Representation

Figure 2.5: **t-SNE plots for different sequence encoding strategies.** Each dot represents a 400bp-long genomic sequence at the center of an ATAC-seq peak. Dots were colored by their sequence's GC-richness.

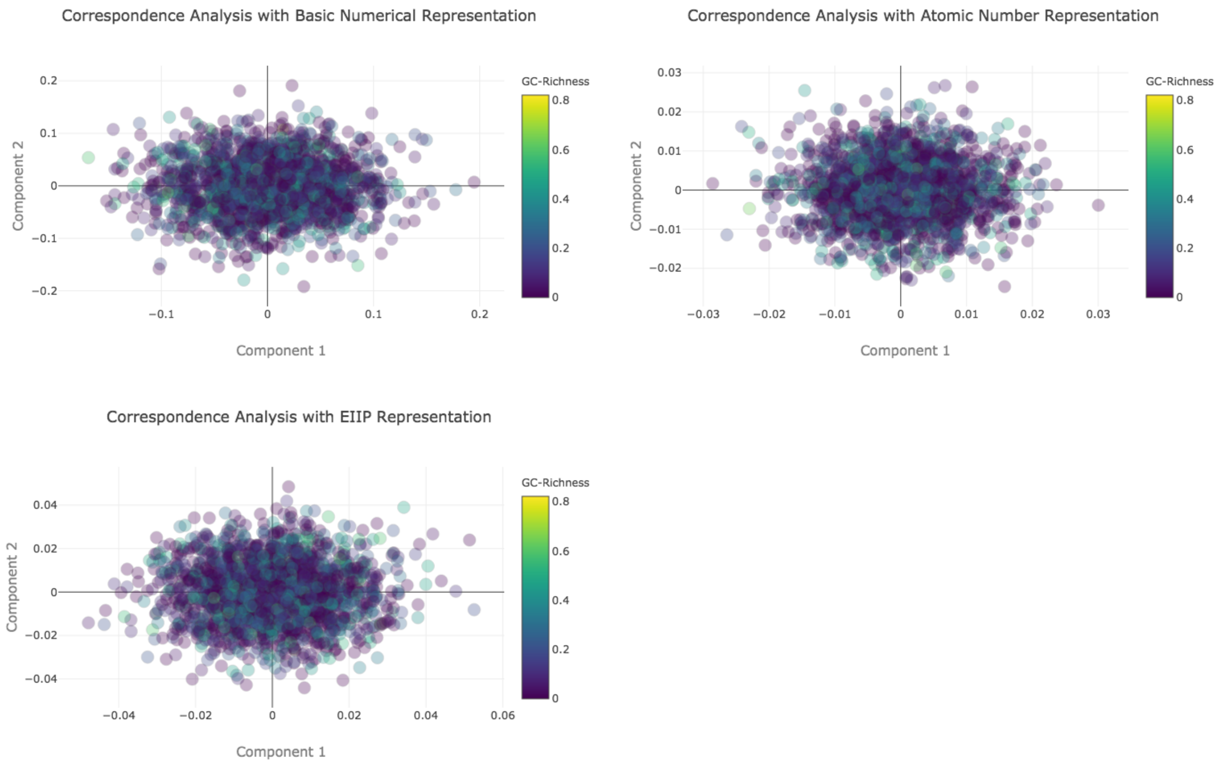


Figure 2.6: **CA plots for different sequence encoding strategies.** Each dot represents a 400bp-long genomic sequence at the center of an ATAC-seq peak. Dots were colored by their sequence's GC-richness.

Chapter 3

Exploring applications of semantic knowledge representations

Initially, this chapter briefly summarizes published manuscripts that describe applications of the Knowledge Base Of Biomedicine (KaBOB[66]), a semantically-consistent integration of biomedical ontologies and public databases. First, section 3.1 summarizes the use of integrated semantic knowledge to find novel drug-drug interactions. An application of KaBOB to the natural language processing task of relation extraction is later summarized in section 3.2. The publications can be found in Appendix B. I then describe unpublished work in section 3.3 where I revisited the relation extraction task, this time utilizing embeddings derived from either text, a new knowledge graph of biomedicine (PheKnowLator[67]), and a combination of both.

3.1 Hidden in plain sight: Inference of new biologically-relevant assertions in existing knowledge

While statistical machine learning applications in biomedical sciences keep improving their accuracy, it would be naive to ignore the wealth of existing human-curated knowledge already encoded in semantic data structures. The availability of open biomedical ontologies (OBOs[68]) and their many concepts denoting real-world entities and processes, as well as an abundance of human-curated and computationally-inferred relations between them, opens the door to a different type of analysis. One such analysis I conducted to search for novel drug-drug interactions[69] illustrates the richness of information that can be derived from interconnected sources of knowledge. The list of suggested drug-drug interactions and other details can be found in Appendix B.1.

Using KaBOB[66], a knowledge-base that integrates different public ontologies and databases, I looked for possible drug-drug interactions in its representation as a large directed, acyclic graph (Fig. 3.1). I gathered all pairs of drugs (each drug represented by a node in this graph) that intersected at the same Reactome[70] pathway step (generally, a biochemical reaction). Besides a list of drugs that merely activate/suppress the same targets (e.g. drugs that simply treat the same symptom or disease), I found pairs of therapeutics and other chemicals that could present potentially adverse interactions (e.g. participating in the same metabolic process involving a particular cytochrome P450 enzyme). This kind of result would be much harder to obtain (if not impossible) from statistical artificial intelligence (AI) alone. The drug-drug paths in the knowledge graph (KG) were also a use case in a different study[71] to generate abstractions of the OWL representation of semantically-connected knowledge.

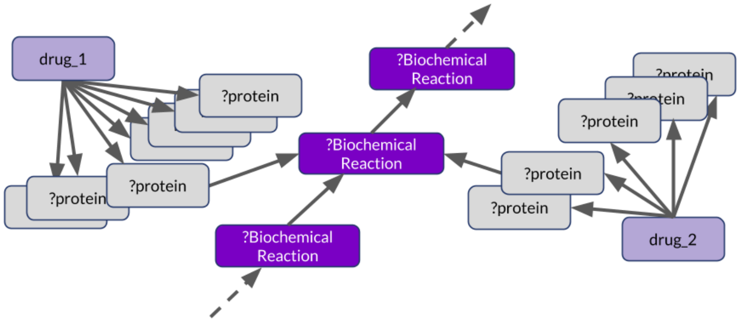


Figure 3.1: **Mining semantic knowledge for drug-drug interactions.** The approach explored in this publication looked for two chemicals (at least one of them a drug) that intersected at the exact same biochemical reaction.

3.2 Expanding our knowledge from the literature by relation extraction

Despite the vast numbers of relations between concepts encoded in these biomedical ontologies, new findings are published every day at a much faster rate than the ontology curators are able to keep up with. In order to capture these relations in an automated manner, I explored the natural language processing (NLP) problem of relation extraction[72] for BioCreative VI's shared task V, to identify how chemicals interact with proteins in an annotated corpus of scientific text. The goal was to predict the correct relation between these types of entities among six possible labels: up-regulation, down-regulation, antagonist, agonist, substrate, or simply no relation at all. The full description of the approach used can be found in the BioCreative VI conference proceedings included in Appendix B.2. The high-level architecture utilized is illustrated in Fig. 3.2.

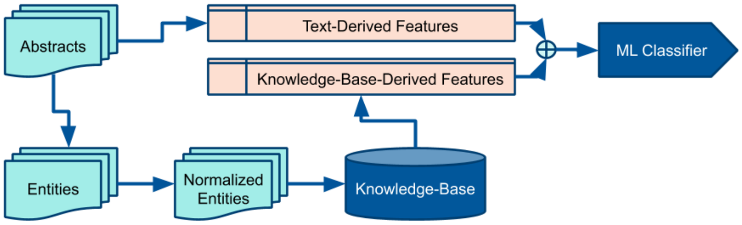


Figure 3.2: **First attempt at combining text- and knowledge-base-derived features.** The machine learning classifier used to predict the relation between the given entity pairs was given text-derived features as well as "bag of concepts" features as input.

I used KaBOB[66] for this task to create binary features in a "bag of concepts" approach, after mapping the given chemical to a ChEBI ontology[73] concept and the protein to a Protein Ontology[74] (PRO) concept. The list of parent concepts all through each ontology root node were set to 1, and the rest remained as zero. An inherent challenge of this approach was to accurately map a language token to a node in the KG that represents said concept. Finding ontological representations of both chemicals and proteins is challenging on one hand due to the variety of synonyms and naming conventions to describe the same molecule. We created a number of heuristics to find matches of the chemicals and proteins in the text to address this, as well as the use of synonym labels in ChEBI, PubChem[75], and PRO. On the other hand, the ChEBI ontology contains a limited number of substances and compounds, so sometimes there simply does not exist a node for a specific molecule. In this case, we imputed zeros for all features corresponding to the missing chemical andor protein. This one-hot concept encoding feature set was also combined with the language tokens along the dependency parse between the chemical and the protein. In this particular approach, the contribution of the knowledge-base-derived features was negligible in practice (likely due to their large number of dimensions), as described in the BioCreative VI workshop proceedings.

3.3 Relation extraction using word- and node-embeddings

I revisited this particular relation extraction task after working with multidimensional vector representations of ontology concepts (concept embeddings), generated from the PheKnowLator graph[67]. After the test labels from the BioCreative shared task were released, I had an opportunity to use this corpus for a different approach. As we demonstrated in a review[76] we recently published, natural language processing was a major focus area for knowledge-based biomedical data science. The questions I was interested in addressing were the same: First, can our predictions of chemical to protein relations from ontology concepts be as accurate (or nearly as accurate) than when using text-derived features? If we are able to classify relations significantly better than a baseline, it could indicate that there is intrinsic semantic knowledge about the chemical and protein in question that allows us to guess how they are likely to interact (e.g. chemicals of type X with certain characteristics tend to down-regulate proteins of type Y). Second, can our text-derived features be complemented with ontology-derived features to increase performance of the relation classifier? My alternative approach consisted in the use of embeddings, or dense vector representations of both word tokens and concepts. To this end, I used every chemical/protein word pair that was annotated in the BioCreative VI shared task V's corpus in the same sentence. If there was a relation annotation that linked these two words, I used it to label the relation. All other pairs were labeled as ”no relation” (that is, they just happen to be mentioned in the same sentence).

The two main components I used to encode the information about the entities we are trying to extract relations from, were a set of pre-trained word embeddings and a semantic knowledge graph from which I generated node embeddings. For word embeddings, I employed the BioWordVec[77] dataset which consists of 200-dimensional pre-trained embeddings from articles on PubMed (`https://www.ncbi.nlm.nih.gov/pubmed/`) and the clinical notes from MIMIC-III Clinical Database[78] (nearly a total of 5 billion tokens). This included not just biomedical terms but every other English word present in their corpus. For a chemical/protein pair in the corpus to be used in this new approach, it had to satisfy two conditions. First, both word tokens had to exist in the BioWordVec vocabulary so that there is a vector representation of each. Second, both tokens had to be mappable to an ontology concept. For chemicals, this required a mapping to the Chemicals of Biological Interest (ChEBI) ontology [73], and proteins were mapped to the Protein Ontology (PRO)[74] whenever possible. A great number of all possible chemical/protein pairs did not contain relation annotations, so in order to incorporate these in a balanced manner we randomly drew from this pool of "unrelated" pairs the same number of samples than our label with the largest support (1,847 total). This resulted in a total of 5,351 chemical/protein relations used to test performance, with a relation label distribution illustrated in Fig. 3.3.

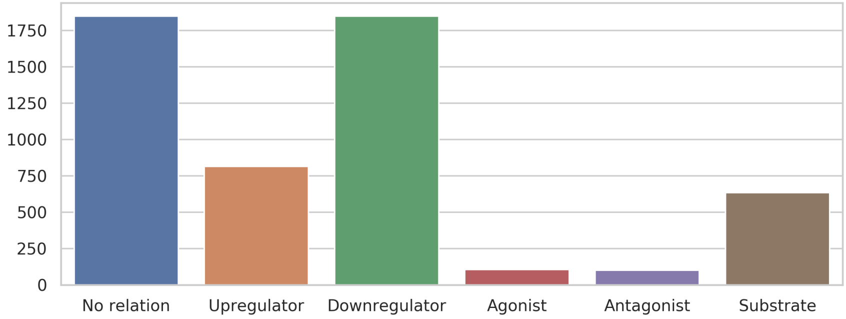


Figure 3.3: **Number of samples for each relation label.** This figure shows how the 5,351 sample labels are distributed, indicating there is a significant imbalance in the dataset with much fewer chemical/protein pairs with agonist/antagonist relations.

The mapping step required the use of aliases and multiple heuristics described in Tables 3.1 and 3.2, as well as the use of multiple sources to match an entity textual description to an ontology concept. For chemicals in particular, if none of the heuristics worked, we queried the PubChem[75] API service to obtain a SMILES string describing the substance. This consists of a string describing the atoms composing the chemical in question, and their structural organization. If a SMILES string was available for the chemical, we looked for the closest existing chemical in ChEBI with very high similarity (> 0.9) that we could use in place, assuming similar properties are due to a similarity in structure. In some cases, this method resulted in the exact same chemical to be matched (a similarity of 1.0), simply because the name was radically different than any of the known aliases in ChEBI.

Table 3.1: **Heuristics and data sources used for chemical concept mapping to ChEBI from the lowercase string annotated as a chemical.** Every possible combination of heuristic and source was used to attempt mapping a string to its corresponding ontology concept.

Heuristics	**Data sources**
• Verbatim	• ChEBI name
• Roman numerals in parentheses to number+ (e.g. "Fe(III)" to "Fe(3+)")	• ChEBI synonym
• Add a space between letters and a number (e.g. "TRP1" to "TRP 1")	• PubChem name to SMILES formula, formula to ChEBI ID
• Add a dash between letters and a number (e.g. "TRP1" to "TRP-1")	• PubChem synonim to ChEBI name
• Skip leading non-alphanumeric (e.g. "(-)-alprenolol" to "alprenolol"	• DrugBank name to ChEBI ID
	• ChEBML name to ChEBI ID • ChEMBL drug indication to ChEBI ID • PubChem name to SMILES formula, find closest compound with > 0.9 DICE similarity, search that SMILES formula to ChEBI ID

Using the dependency parser from the Stanford CoreNLP toolkit[79], I obtained the dependency tree for each sentence containing a usable chemical/protein pair. The tokens along the shortest path in this tree connecting the chemical and protein words has proven to be useful to predict the relation between them[72]. To use these as text-derived features, I took the words along this shortest path and, after discarding English stop-words, I averaged the word embeddings of the remaining ones to come up with 200 more numerical features. Averaging these dense vector representations of words is equivalent to calculating the hyper-dimensional centroid of these vectors. The features used as input to a machine learning classifier consisted of the concatenation of the word embeddings for the chemical, protein, and averaged shortest dependency path terms.

I tested a random forests classifier with 5-fold cross-validation, ensuring a balanced selection of labels on every occasion (via a stratified shuffle split). This way, while keeping a random selection for each training and test set I also ensured the distribution of label membership was kept very similar in each test. This strategy resulted in roughly the same number of test samples

Table 3.2: **Heuristics and data sources used for human protein concept mapping to PRO from the lowercase string annotated as a protein.** Every possible combination of heuristic and source was used to attempt mapping a string to its corresponding ontology concept. If any Greek character was used in the text annotated as a protein, it was replaced to the spelled out English character name before these heuristics were applied (e.g. “TNF-α” to “TNF-alpha”).

Heuristics	**Data sources**
• Verbatim	• PRO name
• Greek letter name, to single letter (e.g. “TNFalpha” to “TNFa”)	• PRO synonym
• Add “-(human)” suffix (e.g. “TP53” to “TP53-(human)”)	• PRO name matches string before a dash (unless added by heuristic)
• Add “-protein” suffix (e.g. “TP53” to “TP53-protein”)	• PRO name matches string after a dash (unless added by heuristic)
• Add “-like-protein” suffix (e.g. “TP53” to “TP53-like-protein”)	• String matches gene name in Uniprot, link to PRO ID
• Add “-complex” suffix (e.g. “TP53” to “TP53-complex”)	
• Add “-related-protein” suffix (e.g. “TP53” to “TP53-related-protein”)	
• Remove dashes (e.g. “TNF-a” to “TNFa”)	
• Change “human” to “h” (e.g. “TP53-human” to “TP53-h”)	
• Add “h” prefix (e.g. “TP53” to “hTP53”)	
• Remove any non-alphanumeric character (e.g. “TNF-(a)” to “TNFa”)	
• Remove “human-” prefix (e.g. “human-TP53” to “TP53”)	
• Remove “human-” prefix and dashes (e.g. “human-TNF-a” to “TNFa”)	

on every validation fold, for the “Upregulator” (163 samples), “Downregulator” (370), “Agonist” (21), “Antagonist” (20), “Substrate” (127) and “No relation” (370) labels. The overall classifier performance is displayed in Fig. 3.4. Overall, incorporating concept features to the text-derived features did not increase performance in a significant way. A detailed illustration of precision and recall for each relation classified is displayed in Fig. 3.5.

Perhaps the difference in variance observed in Fig. 3.5 for different relation labels can be attributed to differences in how rich (or arbitrary) language is, to describe those relations between chemicals and proteins. The number of available samples to train some relation types certainly is a

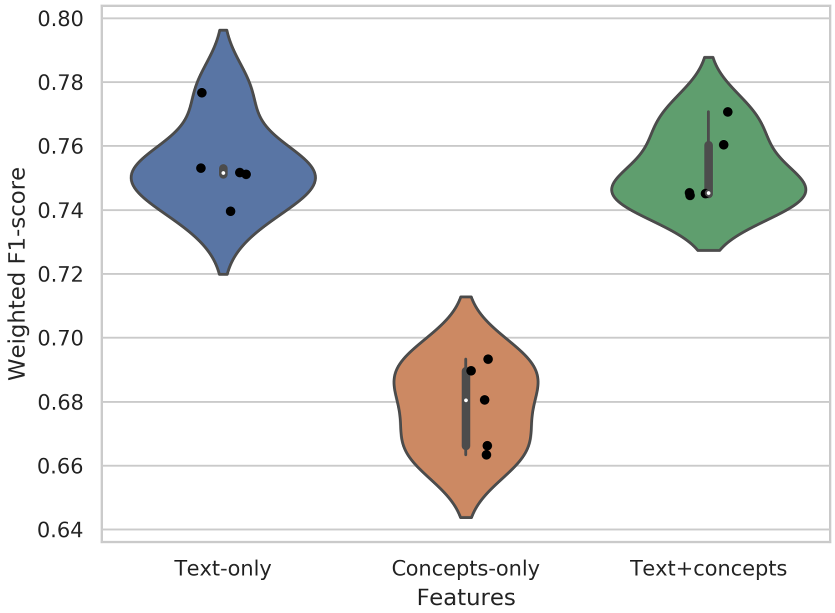


Figure 3.4: **Weighted F1-scores for each feature type.** Performance of the classifier using just word embeddings (left), just concept embeddings (center) and a combination of both (right).

relevant factor that affects classification performance on them. Another important factor affecting the classifier's ability to discriminate among these six possible relations is the presence of annotation errors (several of which were found in this corpus and fixed by hand for this experiment), either by incorrectly annotating terms as chemicals or proteins or by missing a correct annotation (and thus that chemical/protein pair being labeled as "not related").

A deep neural network configuration was also tested on this same dataset. This consisted of a dense layer with ReLU activation followed by a dropout layer, a smaller dense layer, another dropout layer, another yet smaller dense layer and a final softmax activation layer with 6 dimensions to determine the likelihood of each label. I explored a hyperparameter space of dropout rate [0.1, 0.2, 0.3], learning rate [0.01, 0.005, 0.001, 0.0005, 0.0001], and momentum [0.7, 0.8, 0.9]. However, the difference in performance using the optimal configuration was not statistically significant compared to that of the random forests classifier, which was preferred due to its speed and simplicity.

The addition of concept embeddings to the existing text embeddings did not appear to

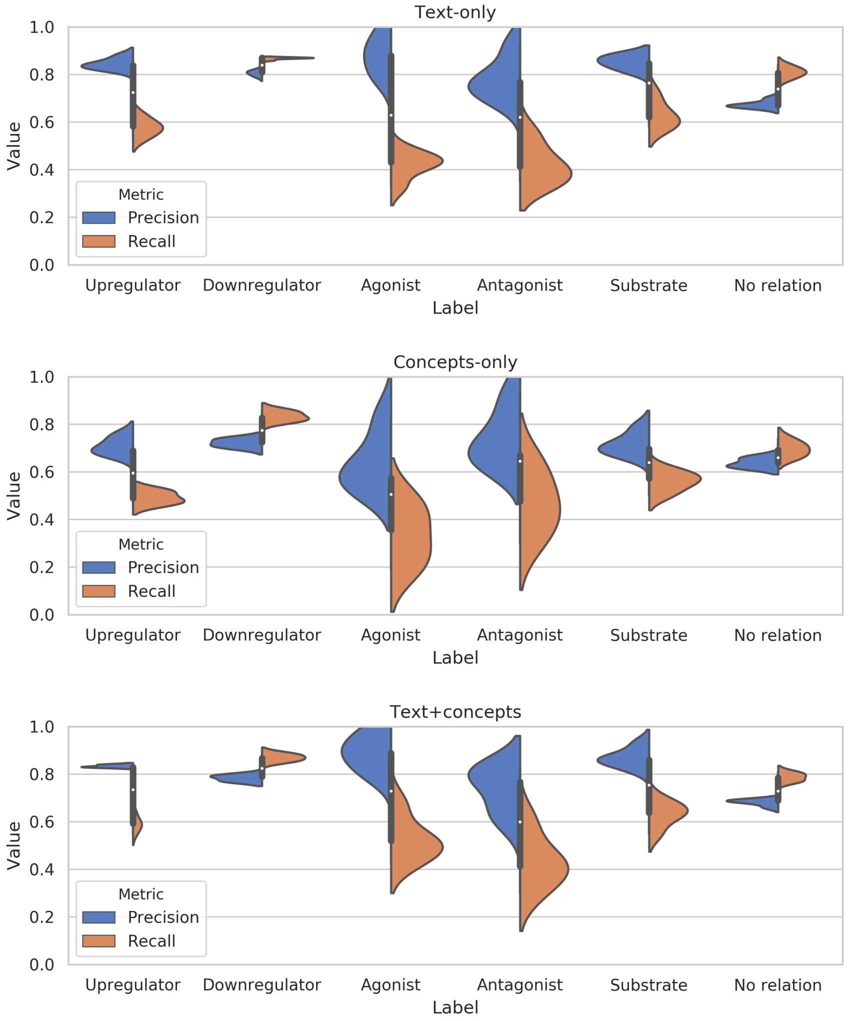


Figure 3.5: **Precision and recall for individual labels classified.** Performance of the classifier on each specific label, using just word embeddings (top), just concept embeddings (center) and a combination of both (bottom).

improve the classification performance in any significant manner. It just resulted in a slightly lower variance. While the attempted relation classification using just the concept embeddings did not perform as well as the word embeddings alone, it's worth noting how the predictive performance is still far better than random assignment. This suggests there are qualitative characteristics of the chemicals and proteins that allow us to predict the expected relation between them. As our existing knowledge of the relation between chemicals (and molecular components) expands, and so does their ontological representation, this exercise is worth revisiting.

3.4 In Summary

I have shown how we can utilize semantic graph representations of biomedical and chemical knowledge with targeted queries to generate novel hypotheses about drug-drug interactions. I have also shown how a rich integration of ontologies can be used to predict certain relations between chemicals and proteins, with the possibility to extend this analysis to relations between other types of biochemical entities.

In the relation extraction area, I have pushed the performance of chemical/protein relation inference by using word embeddings and dependency parsing. The question of how to best integrate word embeddings with concept embeddings is worth exploring further, as this approach is still at a nascent stage.

3.5 List of publications and bioinformatics tools relevant to this chapter

Appendix B contains the collection of my first-author original research articles, pre-prints and conference proceedings mentioned in this thesis, which have been either published or are currently in the process of peer review. The manuscripts relevant to this chapter are listed below, in chronological order:

- I. Tripodi, K. B. Cohen, and L. E. Hunter, "A semantic knowledge-base approach to drug-drug interaction discovery" in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 11231126.

- I. J. Tripodi, M. Boguslav, N. Hailu, and L. E. Hunter, "Knowledge-base-enriched relation extraction" ResearchGate. [Online]. Available: `http://www.biocreative.org/media/store/files/2018/BC6_track5_6.pdf`.

My additional co-authored publications and tools relevant to this chapter are listed below in chronological order:

- T. J. Callahan, W. A. Baumgartner, M. Bada, A. L. Stefanski, I. J. Tripodi, E. K. White, and L. E. Hunter, “OWL-NETS: Transforming OWL Representations for Improved Network Inference”, in Biocomputing 2018, 0 vols., WORLD SCIENTIFIC, 2017, pp. 133144.

- T. J. Callahan, W. A. Baumgartner, I. J. Tripodi, A. L. Stefanski, J. Wyrwa, “PheKnowLator: Phenotype Knowledge Translator” (2019), GitHub Repository, `https://github.com/callahantiff/PheKnowLator/wiki`

- T. J. Callahan, I. J. Tripodi, H. Pielke-Lombardo, and L. E. Hunter, “Knowledge-based Biomedical Data Science 2019”, currently under review.

Chapter 4

Bringing differential omics and semantic knowledge representation together

This chapter summarizes the published manuscript from my final thesis work in section 4.2, where I combined differential expression analysis with a knowledge graph that semantically integrated public databases and biomedical ontologies (PheKnowLator[67]) to infer mechanisms of cellular toxicity. Section 4.3 then describes how I expanded a differential analysis tool I developed (DAStk[33, 56]) to seek relations between the transcription factors (TFs) which are significantly changing between two biological conditions, from curated knowledge sources.

4.1 Background

After my experience with differential analysis in genomics and semantic knowledge representations, I decided to explore a branch of artificial intelligence that is still at a nascent stage: mechanistic inference. This area is a perfect fit for my interest in computational toxicology and its role in replacing animal models, as there is a great value in understanding the mechanisms of toxicity of chemicals. One application for this mechanistic inference task would be during the development of novel chemicals (including drugs), or to study compounds already on the market that result in severe adverse reactions for a subset of the population. Moreover, some mechanisms of cellular toxicity are actually beneficial to design novel oncological chemotherapeutics.

Much of the computational work in toxicology has revolved around determining whether a chemical is toxic or not, by itself or in a mixture, towards a particular type of human tissue (referred to as the "toxicity endpoint" in the literature). No prior study has, however, attempted to systematically identify which are the potential mechanisms by which these compounds result in an adverse outcome, for widely different chemicals in a variety of tissues. My final thesis project was the most directly applicable to computational toxicology, and consisted of generating mechanistic hypotheses of cellular toxicity from experimental data.

4.2 Mechanistic inference: Combining differential omics analysis with knowledge representation

Answering the “why” question of cellular toxicity posed a very interesting research topic, which I addressed by creating MechSpy[80], a mechanistic inference framework that uses gene expression data from publicly available microarray time series. From a mechanistic toxicology textbook[81], I curated a representation of several mechanisms of cellular toxicity as an ordered sequence of GO concepts that best depicted the expected molecular events. I used an extended version of the knowledge graph (KG) from the PheKnowLator[67] project, which integrated the most common OBOs and other information sources like Reactome[70] and the Cellular Toxicogenomics Database[82] (CTD), to which I added relevant edges from toxicology-related sources like the Adverse Outcome Pathway Wiki (AOPwiki[83]). I then employed a technique which originated in NLP (embeddings, or dense vector representations of words), in this case applied to nodes in the KG. Utilizing public gene expression time series in a variety of cell types post-exposure to different kinds of toxic chemicals, my inference framework generated a hypothesis for each of the three most likely mechanisms of toxicity predicted. For most of the time series tested, and particularly for those using chemical exposures at a high dose, the known mechanisms of toxicity were reflected among our top-three predictions (approximately 85% of the time).

The mechanistic inference process followed by MechSpy is described next at a high level. In a deductively-closed version of the KG (using the Elk semantic reasoner[84]), in which we added new edges between ontology concepts for a given list of acceptable transitive relations, I ran the node2vec[85] algorithm to produce node embeddings. This algorithm performs a number of random walks starting at every node in the KG, and generates a dense vector representation of each node that captures the semantic meaning of the concept it denotes. This way, we expect nodes like "BRCA1" (from the GO) and "breast cancer" (from the human phenotype ontology, or HPO) to be close in semantic space. This technique is analogous to the generation of word embeddings in NLP, based on neighboring terms instead of neighboring nodes. The metric I used to determine similarity between node embeddings was cosine distance (how close two vectors align in hyper-dimensional orientation), which is commonly used in NLP to determine if two word embeddings are close in meaning.

At each experimental time point, I performed differential analysis to determine the (up to) 100 genes displaying the most significant change in expression. After obtaining the node embeddings corresponding to these top genes and averaging them, I used this hyper-dimensional centroid vector to represent the changes in expression as a whole in that time point. The similarity between the gene centroid for each time point and each step (GO concept) in a mechanism of toxicity was then calculated, applying a penalty if the mechanism steps were enriched out of the expected sequential order (details in Appendix B.6 methods). The highest enrichment score obtained for each mechanism step was averaged to calculate a final enrichment score for the mechanism as a whole. Using a bootstrap calculation from random gene draws, I also determined an empirical p-value for this final score. The top-three mechanisms displaying significant ($p \leq 0.05$) scores were then presented as the most likely hypotheses.

For each of these three mechanisms, I looked for known relations between the most significant genes at each time point and their associated mechanism steps, to produce a narrative for a putative mechanistic explanation. A graphical representation of this explanation was also produced as an alternative, showing the gene expression changes in order and how these relate to each mechanism step (as illustrated in Fig. 4.1). Finally, we were able to experimentally validate our mechanistic predictions of mitochondria-mediated toxicity for two chemicals (chlorpromazine and adapin), at the same dose and using the same cell type than the public time series I evaluated. This work is described in detail in the submitted manuscript provided in Appendix B.6, undergoing peer review at the time this document was written.

4.3 Providing background to differential transcription factor activiy

As mentioned in a previous chapter, I created a bioinformatics tool called DAStk to present a list of TFs displaying the most significant difference in activity between two conditions. A natural follow-up question after obtaining this information is what do we know to be in common between some of these TFs, and whether there are biological processes or biochemical pathways that they participate in, that could hint at the underlying cell behavior. The normal course of action is to perform a manual literature search, which can be time-consuming and prone to missing relevant connections between TFs, due to the sheer volume of relevant studies for each highlighted TF in DAStk's results. To this end, I constructed an undirected graph that consisted of the combination of Reactome[86] and all relevant human protein information from Uniprot[87]. This allowed me to find paths connecting two or more of the significant TFs listed in DAStk's output, making that information immediately available to researchers.

To construct the lookup graph, I first used every human protein entry from Uniprot, including the GO annotations for each protein (of biological processes, molecular functions, and cellular compartments), information about cofactors, domains, and protein-protein interactions. This allowed me to link every node in the graph denoting a human protein to a GO concept, another protein, a domain described in PFAM[88], or a chemical cofactor in ChEBI[73]. Finally, I obtained the mappings from Reactome of Uniprot entries to all pathways and to all reactions, which allowed me to link biochemical reactions and pathways to their participating human proteins. The process to build this graph was scripted to be updated anytime with more recent annotations and protein entries. I made the graph available as a NetworkX object, to easily import it from the new DAStk release. The Uniprot entries include TFs among all human proteins, so the new DAStk tool searches for all shortest paths between each pair of nodes representing TFs from our results, at most two hops away from each other (or more if intersecting at a Reactome pathway).

This information not only saves many man-hours in literature searches, but also highlights non-obvious characteristics shared by sets of TFs. For example, we can provide information about a subset of resulting TFs that participate in the same pathway or process, that share a common cofactor, that have a similar binding domain, that are known to interact with each other, that have the same molecular function, etc. I illustrate this below with an excerpt of the kind of information we can obtain from a real example, where differential TF activity between cells in two conditions resulted in twenty TFs being highlighted as significantly changing in activity:

```
Transcription factors displaying a significant difference in activity:
CEBPB, CEBPD, CEBPE, ELK1, HLF, NFIA, NFIB, NFIX, NFYB, NRF1, TP53, ZNF180, ZNF341,
ZNF396, ZNF432, ZNF441, ZNF519, ZNF529, ZNF540, ZNF93

Here’s what we know about these TFs presenting significant activity changes (p=1.00E-03):

Direct interactions between each of these TFs:
----------------------------------------------
NFYB interacts with TP53
TP53 interacts with CEBPB

Other ways these TFs are related:
---------------------------------
CEBPB, CEBPE, ELK1, HLF, NFIA, NFIB, NFIX, NFYB, NRF1, TP53, ZNF180, ZNF341, ZNF396,
ZNF441, ZNF519, ZNF529, ZNF540, and ZNF93: located in nucleus
ZNF180, ZNF341, ZNF396, ZNF432, ZNF441, ZNF519, ZNF529, ZNF540, and ZNF93: has
component Zinc finger, C2H2 type
ZNF180, ZNF341, ZNF396, ZNF432, ZNF441, ZNF519, ZNF529, and ZNF540: has function metal
ion binding
ZNF180, ZNF432, ZNF441, ZNF519, ZNF529, ZNF540, and ZNF93: has component KRAB box
CEBPB, CEBPD, CEBPE, NFYB, TP53, and ZNF396: has function protein heterodimerization
activity
ZNF180, ZNF432, ZNF441, ZNF519, ZNF529, and ZNF540: molecularly interacts with
KRAB-ZNF / KAP Complex [nucleoplasm]
```

```
CEBPB, CEBPD, CEBPE, and HLF: has component Basic region leucine zipper
[...]
CEBPB and CEBPD: interacts with ATF4
CEBPB and CEBPD: interacts with CEBPA
CEBPB and CEBPD: participates in positive regulation of osteoblast differentiation
CEBPB and CEBPE: participates in cellular response to lipopolysaccharide
CEBPB and CEBPE: participates in defense response to bacterium
```

Since any kind of interaction common to the TFs in question is valuable, an undirected graph was a reasonable choice to perform these path searches. This graph was incorporated as part of DAStk release 1.0.0, with a tool to read DAStk's own differential analysis output file and find all shortest paths between every pair of nodes denoting the significantly changing TFs. In order to use the suggested motif sites in the tool documentation, I produced mappings from HOCOMOCO v11 motifs to Uniprot IDs, as well as a generic and more extensive mapping from common TF symbols to Uniprot IDs.

4.4 In Summary

I have demonstrated how we can integrate data from a time series of gene expression with a semantic knowledge graph, to generate mechanistic hypotheses of cellular toxicity. I have also designed a strategy for pathway enrichment that takes the sequential order of events into account. The code for MechSpy, the framework I created, is publicly available to the scientific community. This work should serve as a stepping stone for the curation of further mechanisms of toxicity, and ontology-based representations of adverse outcome pathways. It could also be applied to seeking enrichment of other types of mechanisms that can be described with linked ontology concepts, even beyond the area of biology.

4.5 List of publications and bioinformatics tools relevant to this chapter

Appendix B contains the collection of my first-author original research articles, pre-prints and conference proceedings mentioned in this thesis, which have been either published or are currently in the process of peer review. The manuscript relevant to this chapter is indicated below:

- I. J. Tripodi, T. J. Callahan, J. T. Westfall, N. S. Meitzer, R. D. Dowell, L. E. Hunter, "Applying knowledge-driven mechanistic inference to toxicogenomics" (currently under peer review)

My open-source bioinformatics tools and additional co-authored publications relevant to this chapter are listed below in chronological order:

- I. J. Tripodi and M. Gruca, "Differential ATAC-seq toolkit (DAStk)" (2018), GitHub Repository, `https://github.com/Dowell-Lab/DAStk`
- I. J. Tripodi, "MechSpy: Mechanistic inference of toxicity from gene expression time series and knowledge graphs" (2019), GitHub Repository, `https://github.com/ignaciot/MechSpy`
- T. J. Callahan, W. A. Baumgartner, I. J. Tripodi, "PheKnowLator: A repository for building biomedical knowledge graphs of human disease mechanisms." (2019), GitHub Repository: `https://github.com/callahantiff/PheKnowLator`

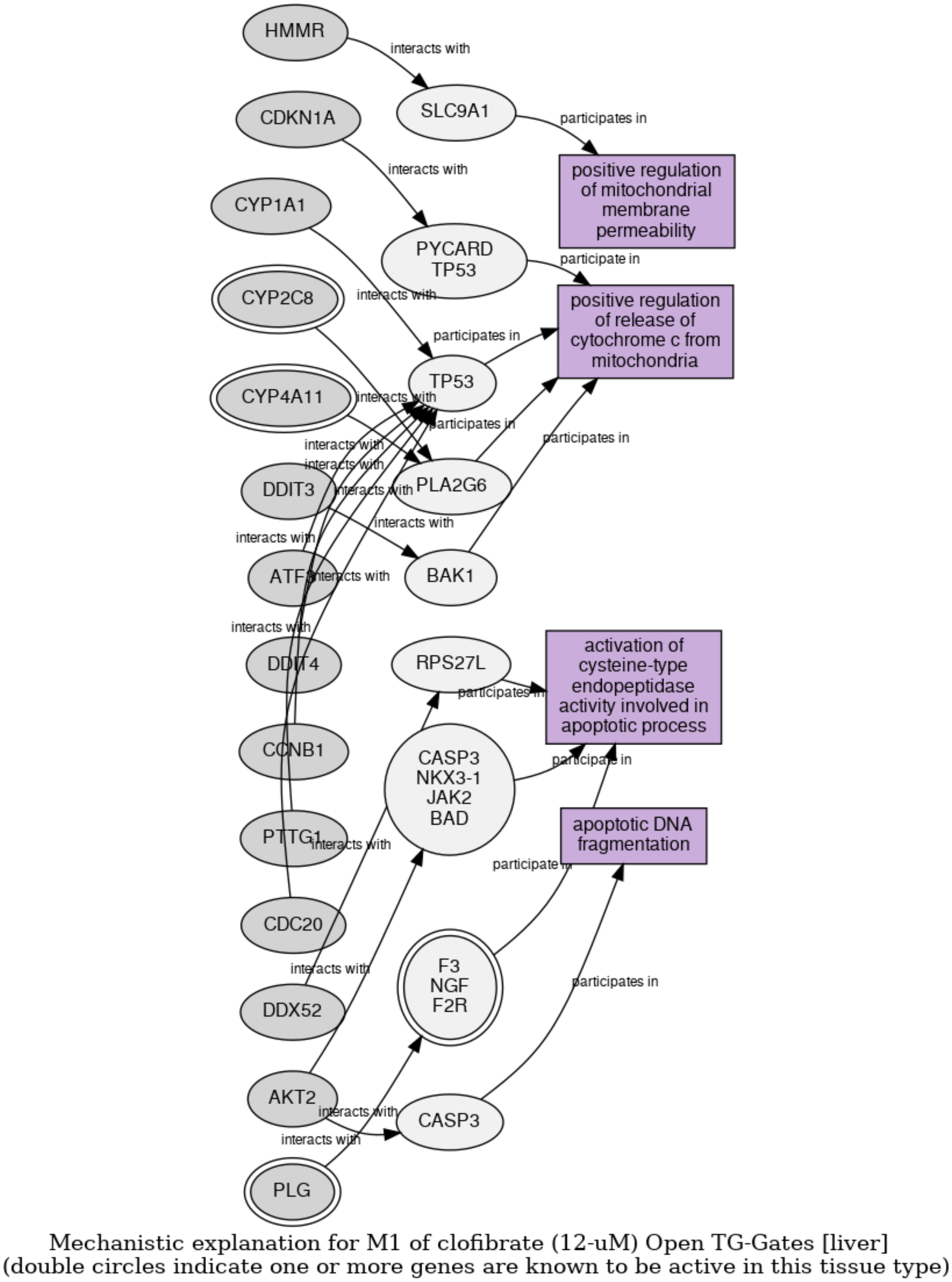


Figure 4.1: **Example of a toxicity mechanism inferred from an experimental time series.** This putative explanation for caspase-mediated apoptosis was generated by MechSpy for hepatocytes treated with a 12μM concentration of clofibrate, for which gene expression was assayed at three time points. The graph can be read from top to bottom, with nodes in dark gray representing genes with significant expression changes. Their known relations to each mechanism step (in purple) are shown.

Chapter 5

Future work

Combining data analysis from gene expression time series and a semantic representation of biomedical knowledge to infer biological mechanisms has been a significant step forward in a rather unexplored area. Many improvements of the current inference framework are possible, nonetheless, on the experimental, knowledge representation, and computational aspects. Section 5.1 describes possible enhancements in the knowledge graph utilized, and section 5.2 speculates about seeking enrichment of ontologically-sound descriptions of established adverse outcome pathways (AOPs), rather than high-level mechanisms. Then, section 5.3 proposes the use of different node embedding strategies, and section 5.4 discusses other sources of experimental time series that could be utilized for mechanistic inference of cellular toxicity.

5.1 Knowledge graph expansion to further sources

The knowledge graph (KG) of human biology[67] used for MechSpy would require an expansion to include information about TFs from the Uniprot database, to seek enrichment from TFs rather than (or in addition to) genes. It would also be highly informative to include which complexes they form (and which are the other protein or chemical components), as well as their respective interactions among them, and biochemical reactions and pathways in Reactome (the current entries in the KG correspond to proteins from protein-coding genes only). Independently of the experimental data source used, and even if we continue to use gene expression as the sole experimental input to MechSpy, the current KG used to derive node embeddings from is very genecentric. The expansion of the KG would still be beneficial to create a set of more informative gene node embeddings, if we continue relying on gene expression time series, because of the new relations between regulatory proteins. The addition of human protein nodes (concepts) could also include toxicity-specific sources related to proteins like the Toxin and Toxin-Target Database(T3DB[89]). Yet another benefit of incorporating all human proteins and their relations to the KG is the ability to find relations between TFs. This is a particularly attractive enhancement for DAStk, allowing to present ways in which the TFs significantly changing in activity relate to each other. In other words, it would allow the generation of mechanistic explanations for the differential analysis performed just like it's done for MechSpy. These TF/TF relations could highlight specific biochemical reactions and higher-level biological processes taking place, as well as explain the change of activity of certain TFs due to taking part in a larger molecular complex.

5.2 Alternative representation of mechanisms

The information in the KG derived from the AOPwiki could also be significantly improved. Currently, this consists in new edges between entities that are known to be causally upstream of others in a context of toxicology. A task that could significantly benefit the computational toxicology community beyond this thesis work, would be to properly curate every documented adverse outcome pathway (AOP) using proper ontology concepts and relations. This would require the use of Gene Ontology Causal Activity Models (GO-CAM[90]), as we would need concepts that don't currently exist in any of the ontologies. Besides the promise of more informative node embeddings from a KG that includes these new toxicology-specific relations, a different inference task could consist of seeking enrichment of these ontologically-sound AOPs, rather than the high-level mechanisms of toxicity I curated. In fact, the currently curated mechanisms of toxicity from Boelsterli's mechanistic toxicology textbook[81] would also benefit from using a GO-CAM representation rather than the current list of gene ontology concepts. Doing so would add specificity to these mechanisms, as we can be more precise about biological process inputs, outputs and cellular locations of each mechanism step.

5.3 Alternative methods to generate node embeddings and calculate mechanism enrichment

Alternative methods to generate the node embeddings could also be explored, as well. So far only DeepWalk and node2vec have been tested (the latter producing more informative vectors). We could evaluate the performance of HARP[91] or Walklets[92], using different configurations, to assess whether either is better at encoding semantic meaning of this KG. A grid search of various node embedding algorithms, embedding dimensions and specific hyperparameters could be used to determine the most useful semantic latent representation. It would not be surprising if the ideal configuration to generate node embeddings is dependent on the ontologies used to construct the KG. Therefore, a possible parameter in this grid search could be which ontology or database to exclude from the KG generation, within reasonable computational demands. This last experiment could, alternatively, be performed after the best embedding algorithm configuration is found: in this case, we could use it to determine which of all KG sources is the most impactful in our mechanistic inference process. I can verify this by recalculating the enrichment scores for all time series after excluding one knowledge source at a time from the KG, with the exception of the GO and whatever is used to represent the experimental changes (genes, TFs, etc).

Computational improvements are possible in the embedding and inference spaces. The node embeddings are currently being compared using the cosine distance between them. While this is a perfectly acceptable way to compare embeddings in high-dimensional space, it may be worth exploring a kernel method instead, for a more accurate comparison between these dense vectors. An alternative to the enrichment process can also be explored, by ignoring the temporal order of significant gene/TF changes and treating this as a topic modeling task. The "topics" to be modeled would be the mechanisms of toxicity, and I could use latent Dirichlet analysis (LDA) or an equivalent technique to rank the three most applicable topics to the pool of genes resulting from each time series.

5.4 Experimental alternatives

On the experimental side, a time series of a different type of assay is worth exploring. Better yet, a combination of assays at matching time points could be more informative than gene expression alone. Given that expression assays offer a steady state readout of available mRNA, I'm interested in exploring true measures of transcription via nascent transcription assays, mainly to track transcription factor (TF) activity over time. Not only the TF activity changes would be more mechanistically relevant, but also nascent transcription allows for a much finer time resolution, being able to detect changes in times as short as 10 minutes post-exposure. As explained in the background section, nascent transcription assays also suffer from a complexity that could be bypassed by using ATAC-seq, an open chromatin assay, as a proxy. Thus, I could alternatively employ a time series of ATAC-seq experiments post-exposure to a toxic chemical, and use DAStk to determine the most significant changes in TF activity at each time point, based on the motif displacement (MD-score) statistic. While the time resolution of ATAC-seq is still unknown, changes of the energy-dependent process of chromatin remodeling can be detected at least every 60 minutes.

Seeking enrichment of these mechanisms of toxicity using heterogeneous experimental sources, such as a combination of gene expression and chromatin accessibility, may require exploring different approaches to calculate enrichment scores. The natural first approach would consist of using all nodes from the various entities that presented a significant change at each time point (genes and TFs), and following the current enrichment score calculation strategy. Some alternative calculations to explore would include seeking separate enrichment scores from genes and TFs, and devising a weighting mechanism to combine both into a final score. The process of scoring a mechanism from multiple sources may use different weights depending on the current mechanism of toxicity being evaluated, as the contribution of one type of assay may be more important than in other mechanisms.

Chapter 6

Conclusion

I strongly believe the future of great science advancement is interdisciplinary, and I've had the fortune of being advised by two principal investigators in very different areas of bioinformatics. As an applied scientist, I have explored ways to solve current problems in both the genomics and semantic knowledge representation realms, finally bringing the expertise of both labs together in my final thesis project: mechanistic inference of cellular toxicity. The application I have always had in mind for every project I've worked on was its potential to be used in computational toxicology. There were many reasons that directed me towards pursuing doctoral studies in computational biology/bioinformatics, yet the strongest motivation came from making an advancement towards the "3Rs" paradigm[93] to **r**efine, **r**educe, and **r**eplace the use of animals as experimental models. Computational toxicology presented itself as field ripe with opportunity in this aspect, since mechanistic studies generally rely on animal models with varying degrees of efficacy[94, 95], often presenting translational issues to human health. By improving our existing in vitro and in silico models of mechanistic toxicity, we could eventually replace the existing requirements for in vivo models in lieu of methods more directly applicable to humans. Thanks to increasing advancements in computational and in vitro techniques, I believe it is our scientific and ethical duty to move towards a human-centric approach to toxicology.

I've demonstrated across the various studies during my dissertation the wealth of knowledge that can be inferred from differential analysis, when contrasting different types of omics assays performed in human tissue. I have also shown how novel artificial intelligence techniques employed with existing semantic knowledge of biology and biochemistry can help generate hypotheses of the biological mechanisms at play. While I believe this thesis work has made significant strides towards a formal mechanistic inference methodology, there is plenty of opportunity for improvement. My hope is that this study will serve as a stepping stone for future work on mechanistic hypothesis generation, not just for computational toxicology applications but other biomedical areas as well. As omics assays become increasingly accessible, and organoid models (which offer a much closer physiological response to in vivo models than 2D cell culture[96, 97]) become both easier and more inexpensive to produce, I envision this work to become the basis of computational mechanistic research.

Bibliography

[1] K. Taylor and L. Rego Alvarez. Regulatory drivers in the last 20years towards the use of in silico techniques as replacements to animal testing for cosmetic-related substances. Computational Toxicology, 13:100112, February 2020.

[2] Joseph F. Cardiello, Gilson J. Sanchez, Mary A. Allen, and Robin D. Dowell. Lessons from eRNAs: understanding transcriptional regulation through the lens of nascent RNAs. Transcription, 11(1):3–18, January 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21541264.2019.1704128.

[3] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science, 270(5235):467–470, October 1995. Publisher: American Association for the Advancement of Science Section: Report.

[4] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1):57–63, January 2009. Number: 1 Publisher: Nature Publishing Group.

[5] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7):e47–e47, April 2015.

[6] Sharon R. Grossman, Xiaolan Zhang, Li Wang, Jesse Engreitz, Alexandre Melnikov, Peter Rogov, Ryan Tewhey, Alina Isakova, Bart Deplancke, Bradley E. Bernstein, Tarjei S. Mikkelsen, and Eric S. Lander. Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proceedings of the National Academy of Sciences, page 201621150, January 2017.

[7] Bei Wei, Arttu Jolma, Biswajyoti Sahu, Lukas M. Orre, Fan Zhong, Fangjie Zhu, Teemu Kivioja, Inderpreet Sur, Janne Lehti, Minna Taipale, and Jussi Taipale. A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. Nature Biotechnology, May 2018.

[8] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447(7146):799–816, June 2007.

[9] An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature, 489(7414):57–74, September 2012.

[10] Sebastian Steinhauser, Nils Kurzawa, Roland Eils, and Carl Herrmann. A comprehensive comparison of tools for differential ChIP-seq analysis. Briefings in Bioinformatics, 17(6):953–966, November 2016.

[11] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. The Human Transcription Factors. Cell, 172(4):650–665, February 2018.

[12] Richard I. Sherwood, Tatsunori Hashimoto, Charles W. O'Donnell, Sophia Lewis, Amira A. Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K. Gifford. Discovery of directional and nondirectional pioneerf transcription factors by modeling DNase profile magnitude and shape. Nature Biotechnology, 32(2):171–178, February 2014.

[13] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics, 34(2):166–176, June 2003.

[14] Steve W. Cole, Weihong Yan, Zoran Galic, Jesusa Arevalo, and Jerome A. Zack. Expression-based monitoring of transcription factor activity: the TELiS database. Bioinformatics, 21(6):803–810, March 2005.

[15] Piotr J. Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan, and Erik van Nimwegen. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. Genome Research, 24(5):869–884, May 2014.

[16] Namshik Han, Harry A. Noyes, and Andy Brass. TIGERi: modeling and visualizing the responses to perturbation of a transcription factor network. BMC Bioinformatics, 18(7):260, May 2017.

[17] Guido Sanguinetti, Neil D. Lawrence, and Magnus Rattray. Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. Bioinformatics, 22(22):2775–2781, November 2006.

[18] Harmen J. Bussemaker, Helen C. Causton, Mina Fazlollahi, Eunjee Lee, and Ivor Muroff. Network-based approaches that exploit inferred transcription factor activity to analyze the impact of genetic variation on gene expression. Current Opinion in Systems Biology, 2:98–102, April 2017.

[19] LesleyT. MacNeil, Carles Pons, H. Efsun Arda, GabrielleE. Giese, ChadL. Myers, and AlberthaJ. M. Walhout. Transcription Factor Activity Mapping of a Tissue-Specific InVivo Gene Regulatory Network. Cell Systems, 1(2):152–162, August 2015.

[20] Leighton J. Core, Andr L. Martins, Charles G. Danko, Colin T. Waters, Adam Siepel, and John T. Lis. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature Genetics, 46(12):1311–1320, December 2014.

[21] Joseph G. Azofeifa, Mary A. Allen, Josephina R. Hendrix, Timothy Read, Jonathan D. Rubin, and Robin D. Dowell. Enhancer RNA profiling predicts transcription factor activity. Genome Research, 28(3):334–344, March 2018.

[22] Leighton J. Core, Joshua J. Waterfall, and John T. Lis. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science, 322(5909):1845–1848, December 2008.

[23] Hojoong Kwak, Nicholas J. Fuda, Leighton J. Core, and John T. Lis. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. Science, 339(6122):950–953, February 2013.

[24] Charles G. Danko, Stephanie L. Hyland, Leighton J. Core, Andre L. Martins, Colin T. Waters, Hyung Won Lee, Vivian G. Cheung, W. Lee Kraus, John T. Lis, and Adam Siepel. Identification of active transcriptional regulatory elements from GRO-seq data. Nature Methods, 12(5):433–438, May 2015.

[25] Zhong Wang, Tinyi Chu, Lauren A. Choate, and Charles G. Danko. Identification of regulatory elements from nascent transcription using dREG. bioRxiv, page 321539, May 2018.

[26] Joseph G. Azofeifa and Robin D. Dowell. A generative model for the behavior of RNA polymerase. Bioinformatics, 33(2):227–234, January 2017.

[27] Radhakrishnan Sabarinathan, Loris Mularoni, Jordi Deu-Pons, Abel Gonzalez-Perez, and Nria Lpez-Bigas. Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature, 532(7598):264–267, April 2016.

[28] Shane Neph, Jeff Vierstra, Andrew B. Stergachis, Alex P. Reynolds, Eric Haugen, Benjamin Vernot, Robert E. Thurman, Sam John, Richard Sandstrom, Audra K. Johnson, Matthew T. Maurano, Richard Humbert, Eric Rynes, Hao Wang, Shinny Vong, Kristen Lee, Daniel Bates, Morgan Diegel, Vaughn Roach, Douglas Dunn, Jun Neri, Anthony Schafer, R. Scott Hansen, Tanya Kutyavin, Erika Giste, Molly Weaver, Theresa Canfield, Peter Sabo, Miaohua Zhang, Gayathri Balasundaram, Rachel Byron, Michael J. MacCoss, Joshua M. Akey, M. A. Bender, Mark Groudine, Rajinder Kaul, and John A. Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature, 489(7414):83–90, September 2012.

[29] Jason D. Buenrostro, Beijing Wu, Howard Y. Chang, and William J. Greenleaf. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current Protocols in Molecular Biology, 109(1):21.29.1–21.29.9, January 2015.

[30] Roger Pique-Regi, Jacob F. Degner, Athma A. Pai, Daniel J. Gaffney, Yoav Gilad, and Jonathan K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research, 21(3):447–455, March 2011.

[31] Songjoon Baek, Ido Goldstein, and Gordon L. Hager. Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. Cell Reports, 19(8):1710–1722, May 2017.

[32] Myong-Hee Sung, Michael J. Guertin, Songjoon Baek, and Gordon L. Hager. DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence. Molecular Cell, 56(2):275–285, October 2014.

[33] Ignacio J. Tripodi, Mary A. Allen, and Robin D. Dowell. Detecting Differential Transcription Factor Activity from ATAC-Seq Data. Molecules, 23(5):1136, May 2018.

[34] Alison Callahan, Michel Dumontier, and Nigam H. Shah. HyQue: evaluating hypotheses using Semantic Web technologies. Journal of Biomedical Semantics, 2(2):S3, May 2011.

[35] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLOS Computational Biology, 8(2):e1002375, February 2012.

[36] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology, May 2000.

[37] TheGeneOntologyConsortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research, 45(D1):D331–D338, 2017.

[38] Louis du Plessis, Nives kunca, and Christophe Dessimoz. The what, where, how and why of gene ontologya primer for bioinformaticians. Briefings in Bioinformatics, 12(6):723–735, November 2011.

[39] Daniel Jupiter, Jessica ahutolu, and Vincent VanBuren. TreeHugger: A New Test for Enrichment of Gene Ontology Terms. INFORMS Journal on Computing, 22(2):210–221, October 2009.

[40] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43):15545–15550, October 2005.

[41] Jrg Rahnenfhrer, Francisco S Domingues, Jochen Maydt, and Thomas Lengauer. Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data. Statistical Applications in Genetics and Molecular Biology, 3(1):1–29, 2004.

[42] Ali Shojaie and George Michailidis. Analysis of Gene Sets Based on the Underlying Regulatory Network. Journal of Computational Biology, 16(3):407–426, March 2009.

[43] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S. Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. Bioinformatics, 25(1):75–82, January 2009.

[44] Natalia Rubanova, Anna Polesskaya, Anna Campalans, Guillaume Pinna, Jeremie Kropp, Annick Harel-Bellan, and Nadya Morozova. MasterPATH: network analysis of functional genomics screening data. bioRxiv, page 264119, February 2018.

[45] Alexander Lachmann, Huilei Xu, Jayanth Krishnan, Seth I. Berger, Amin R. Mazloom, and Avi Ma'ayan. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics, 26(19):2438–2444, October 2010.

[46] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Maayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics, 14:128, April 2013.

[47] Benjamin P. Lewis, I-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of Mammalian MicroRNA Targets. Cell, 115(7):787–798, December 2003.

[48] S. A. Racunas, N. H. Shah, I. Albert, and N. V. Fedoroff. HyBrow: a prototype system for computer-aided hypothesis evaluation. Bioinformatics, 20(suppl_1):i257–i264, August 2004.

[49] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1):27–30, January 2000.

[50] S. D Bay, J Shrager, A Pohorille, and P Langley. Revising regulatory networks: from expression data to linear causal models. Journal of Biomedical Informatics, 35(5):289–297, October 2002.

[51] Bruce G. Buchanan and Edward A. Feigenbaum. Dendral and Meta-Dendral: Their Applications Dimension. In Bonnie Lynn Webber and Nils J. Nilsson, editors, Readings in Artificial Intelligence, pages 313–322. Morgan Kaufmann, January 1981.

[52] Ross D. King, Kenneth E. Whelan, Ffion M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature, 427(6971):247–252, January 2004.

[53] Peter D. Karp. Design methods for scientific hypothesis formation and their application to molecular biology. Machine Learning, 12(1):89–116, August 1993.

[54] Lindley Darden, Lipika R. Pal, Kunal Kundu, and John Moult. The Product Guides the Process: Discovering Disease Mechanisms, July 2017.

[55] Lindley Darden, Kunal Kundu, Lipika Ray Pal, and John Moult. Harnessing formal concepts of biological mechanism to analyze human disease. bioRxiv, page 350371, June 2018.

[56] Gruca M. Tripodi, I.J. Differential ATAC-seq toolkit. `https://github.com/Dowell-Lab/DAStk`, 2018.

[57] Ignacio J. Tripodi, Murad Chowdhury, and Robin D. Dowell. ATAC-seq signal processing and recurrent neural networks can identify RNA polymerase activity. bioRxiv, page 531517, January 2019.

[58] Paolo Di Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. Nature Biotechnology, 35:316–319, April 2017.

[59] Ignacio Javier Tripodi, Ben Busby, Steve Tsang, Jingjing Zhao, Evan Floden, and Chi Zhang. An ATAC-seq pipeline wrapped in NextFlow that can be run by Jupyter (ATACFlow). `https://osf.io/ucwrh/`, August 2018.

[60] Ignacio Javier Tripodi and Margaret Gruca. Nascent-Flow. `https://github.com/Dowell-Lab/Nascent-Flow`, December 2018.

[61] Ignacio J. Tripodi and Margaret Gruca. nf-core/nascent: nf-core/nascent version 1.0. `https://github.com/nf-core/nascent`, April 2019.

[62] Philip Ewels, Alexander Peltzer, Sven Fillinger, Johannes Alneberg, Harshil Patel, Andreas Wilm, Maxime Garcia, Paolo Di Tommaso, and Sven Nahnsen. nf-core: Community curated bioinformatics pipelines. bioRxiv, page 610741, April 2019.

[63] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. Bioinformatics, 27(7):1017–1018, April 2011.

[64] Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Ruslan N. Sharipov, Alla D. Fedorova, Eugene I. Rumynskiy, Yulia A. Medvedeva, Arturo Magana-Mora, Vladimir B. Bajic, Dmitry A. Papatsenko, Fedor A. Kolpakov, and Vsevolod J. Makeev. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Research, 46(D1):D252–D259, January 2018.

[65] Rosanna Verde and Antonio Irpino. Dynamic Clustering of Histogram Data: Using the Right Metric. In Selected Contributions in Data Analysis and Classification, Studies in Classification, Data Analysis, and Knowledge Organization, pages 123–134. Springer, Berlin, Heidelberg, 2007.

[66] Kevin M. Livingston, Michael Bada, William A. Baumgartner, and Lawrence E. Hunter. KaBOB: ontology-based semantic integration of biomedical databases. BMC Bioinformatics, 16:126, April 2015.

[67] T. J. Callahan. PheKnowLator. `https://github.com/callahantiff/PheKnowLator/wiki`, March 2019.

[68] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, The OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25(11):1251–1255, November 2007.

[69] I. Tripodi, K. B. Cohen, and L. E. Hunter. A semantic knowledge-base approach to drug-drug interaction discovery. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1123–1126, November 2017.

[70] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome Pathway Knowledgebase. Nucleic Acids Research, 46(D1):D649–D655, 2018.

[71] Tiffany J. Callahan, William A. Baumgartner, Michael Bada, Adrianne L. Stefanski, Ignacio Tripodi, Elizabeth K. White, and Lawrence E. Hunter. OWL-NETS: Transforming OWL Representations for Improved Network Inference. In Biocomputing 2018, pages 133–144. WORLD SCIENTIFIC, October 2017.

[72] Ignacio J. Tripodi, Mayla Boguslav, Negacy Hailu, and Lawrence E. Hunter. Knowledge-base-enriched relation extraction. In BioCreative VI Proceedings. BioCreative, 2017.

[73] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic acids research, 44(D1):D1214–9, January 2016.

[74] Darren A. Natale, Cecilia N. Arighi, Judith A. Blake, Jonathan Bona, Chuming Chen, Sheng-Chih Chen, Karen R. Christie, Julie Cowart, Peter D'Eustachio, Alexander D. Diehl, Harold J. Drabkin, William D. Duncan, Hongzhan Huang, Jia Ren, Karen Ross, Alan Ruttenberg, Veronica Shamovsky, Barry Smith, Qinghua Wang, Jian Zhang, Abdelrahman El-Sayed, and Cathy H. Wu. Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. Nucleic Acids Research, 45(D1):D339–D346, 2017.

[75] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. PubChem 2019 update: improved access to chemical data. Nucleic Acids Research, 47(D1):D1102–D1109, January 2019.

[76] Tiffany J. Callahan, Harrison Pielke-Lombardo, Ignacio J. Tripodi, and Lawrence E. Hunter. Knowledge-based Biomedical Data Science 2019. arXiv:1910.06710 [cs], October 2019. arXiv: 1910.06710.

[77] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific Data, 6(1):1–9, May 2019.

[78] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1):1–9, May 2016.

[79] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[80] Ignacio J. Tripodi, Tiffany J. Callahan, Jessica T. Westfall, Nayland S. Meitzer, Robin D. Dowell, and Lawrence E. Hunter. Applying knowledge-driven mechanistic inference to toxicogenomics. bioRxiv, page 782011, September 2019.

[81] Urs A. Boelsterli. Mechanistic Toxicology: The Molecular Basis of How Chemicals Disrupt Biological Targets, Second Edition. CRC Press, Boca Raton, FL, 2nd edition edition, June 2007.

[82] Allan Peter Davis, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly. The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Research, 47(D1):D948–D954, January 2019.

[83] AOP wiki. `https://aopwiki.org/`, 2019. Accessed: 2019-05-01.

[84] Yevgeny Kazakov, Markus Krtzsch, and Frantiek Simank. The Incredible ELK. Journal of Automated Reasoning, 53(1):1–61, June 2014.

[85] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. arXiv:1607.00653 [cs, stat], July 2016. arXiv: 1607.00653.

[86] Antonio Fabregat, Konstantinos Sidiropoulos, Guilherme Viteri, Oscar Forner, Pablo Marin-Garcia, Vicente Arnau, Peter DEustachio, Lincoln Stein, and Henning Hermjakob. Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinformatics, 18:142, March 2017.

[87] UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1):D506–D515, January 2019.

[88] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurlien Luciani, Simon C. Potter, Matloob Qureshi, Lorna J. Richardson, Gustavo A. Salazar, Alfredo Smart, Erik L. L. Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C. E. Tosatto, and Robert D. Finn. The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1):D427–D432, January 2019.

[89] David Wishart, David Arndt, Allison Pon, Tanvir Sajed, An Chi Guo, Yannick Djoumbou, Craig Knox, Michael Wilson, Yongjie Liang, Jason Grant, Yifeng Liu, Seyed Ali Goldansaz, and Stephen M. Rappaport. T3db: the toxic exposome database. Nucleic Acids Research, 43(Database issue):D928–934, January 2015.

[90] Paul D. Thomas, David P. Hill, Huaiyu Mi, David Osumi-Sutherland, Kimberly Van Auken, Seth Carbon, James P. Balhoff, Laurent-Philippe Albou, Benjamin Good, Pascale Gaudet, Suzanna E. Lewis, and Christopher J. Mungall. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nature Genetics, 51(10):1429–1433, October 2019.

[91] Harp: Hierarchical representation learning for networks. `https://github.com/GTmac/HARP`, 2018. Accessed: 2019-05-01.

[92] Don't walk skip! online learning of multi-scale network embeddings. `https://github.com/benedekrozemberczki/walklets`, 2017. Accessed: 2019-05-01.

[93] W. M. S. Russell and R. L. Burch. The principles of humane experimental technique. London: Methuen & Co. Ltd., 1959.

[94] Gail A. Van Norman. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is it Time to Rethink Our Current Approach? JACC: Basic to Translational Science, 4(7):845–854, November 2019.

[95] Thomas Hartung. Thoughts on limitations of animal models. Parkinsonism & Related Disorders, 14:S81–S83, July 2008.

[96] David Pamies and Thomas Hartung. 21st Century Cell Culture for 21st Century Toxicology. Chemical Research in Toxicology, 30(1):43–52, January 2017.

[97] Justyna Augustyniak, Alessia Bertero, Teresa Coccini, Diego Baderna, Leonora Buzanska, and Francesca Caloni. Organoids are promising tools for species-specific in vitro toxicological studies. Journal of Applied Toxicology, 39(12):1610–1622, 2019.

Appendix A

List of abbreviations

AI: Artificial intelligence

AOP: Adverse outcome pathway

GO: Gene ontology

HPO: Human phenotype ontology

KG: Knowledge graph

NLP: Natural language processing

OBO: Open biomedical ontology

PRO: Protein ontology

SMILES: Simplified molecular-input line-entry system

SNP: Single-nucleotide polymorphism

TF: Transcription factor

TSS: Transcription start site

Appendix B

First-author publications

The content of this appendix consists of the collection of my first-author journal research papers and conference proceedings mentioned above, in chronological order, which have been either published or are currently in the process of peer review.

B.1 A Semantic Knowledge-Base Approach to Drug-Drug Interaction Discovery

I conducted the knowledge retrieval, pathway analysis, and result evaluation. I analyzed the resulting drug pairs against public knowledge from RxNorm data with Larry Hunter. The Sparql query to find the paths between drugs intersecting at the same biochemical reaction was constructed with much help from William Baumgartner and Elizabeth White. All authors participated in the writing of the manuscript.

B.2 Knowledge-base-enriched relation extraction

I processed the provided corpus, generated the feature sets, and conducted most of the concept mapping to public biomedical ontologies, as well as most of the different machine learning classifier implementations. Mayla Boguslav conducted very helpful error analysis and incorporated new heuristics to increase out concept mapping performance. Negacy Hailu generated the dependency parses and implemented the neural network classifier. All authors participated in the writing of the manuscript.

B.3 Detecting Differential Transcription Factor Activity from ATAC-Seq Data

Mary Allen and Robin Dowell conceived and designed the experiments. I implemented DAStk and performed the experiments. I gathered additional public ATAC-seq datasets of the same cells in different biological conditions, conducted the literature search to validate new DAStk-derived results, and analyzed the results with Robin Dowell. All authors contributed to writing the paper.

B.4 ATAC-seq signal processing and recurrent neural networks can identify RNA polymerase activity

I envisioned the idea of approaching this open chromatin region classification as a signal processing task, conceived the experiments with Robin Dowell, implemented the signal processing code and most machine learning classifiers and data encoding schemes. I also processed all public datasets from the different high-throughput sequencing protocols. With Robin Dowell, I conducted the error analysis for both detection of transcription and histone marks, and designed the statistics displayed in the discussion. Murad Chowdhury implemented the recurrent neural network and nucleotide embedding. We designed all other machine learning scenarios with Murad Chowdhury. All authors participated in the writing of the manuscript.

B.5 Combining signal and sequence to detect RNA polymerase initiation in ATAC-seq data

In this in-depth reanalysis of the previous manuscript, focused solely on detecting nascent transcription, I designed the project with Robin Dowell. We designed the new training/validation/test dataset split and implemented the machine learning classifiers with Murad Chowdhury. I conducted the error analysis and cloud-based execution of all machine learning tests. All authors contributed to writing the paper.

B.6 Applying knowledge-driven mechanistic inference to toxicogenomics

I conceptualized and implemented the mechanistic inference framework, processed all public transcriptomics data, incorporated AOPwiki-derived edges to the knowledge graph, and curated all mechanisms of toxicity. Tiffany Callahan designed and implemented the knowledge graph, and conducted the deductive closure on it. Jessica Westfall performed the experimental validation of our predictions for some compounds without established mechanisms of toxicity. Nayland Meitzer helped with the literature review to label known mechanisms of toxicity for the chemicals we evaluated. Robin Dowell and Larry Hunter supervised the research. All authors but Nayland Meitzer contributed to the writing of this manuscript.