UNIVERSITY OF
CAMBRIDGE
Department of Computer
Science and Technology

Residue identity prediction from an amino-acid residue's atomic-environment with equivariant graph machine learning

Antonia-Irina Boca

King's College

May 2023

Submitted in partial fulfillment of the requirements for the
Computer Science Tripos, Part III

Total page count: ??

Main chapters (excluding front-matter, references and appendix): 10 pages (pp 5–14)

Main chapters word count: 467

Methodology used to generate that word count:

[For example:

```
$ make wordcount
gs -q -dSAFER -sDEVICE=txtwrite -o - \
   -dFirstPage=6 -dLastPage=11 report-submission.pdf | \
egrep '[A-Za-z]{3}' | wc -w
467
```

]

Declaration

I, Antonia-Irina Boca of King's College, being a candidate for the Computer Science Tripos, Part III, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

Signed: Antonia-Irina Boca

Date: April 18, 2023

Abstract

Write a summary of the whole thing. Make sure it fits on one page.

Acknowledgements

This project would not have been possible without the wonderful support of ... [optional]

Contents

Chapter 1

Introduction

This is the introduction where you should introduce your work. In general the thing to aim for here is to describe a little bit of the context for your work – why did you do it (motivation), what was the hoped-for outcome (aims) – as well as trying to give a brief overview of what you actually did.

It's often useful to bring forward some "highlights" into this chapter (e.g. some particularly compelling results, or a particularly interesting finding).

It's also traditional to give an outline of the rest of the document, although without care this can appear formulaic and tedious. Your call.

Chapter 2

Background

A more extensive coverage of what's required to understand your work. In general you should assume the reader has a good undergraduate degree in computer science, but is not necessarily an expert in the particular area you've been working on. Hence this chapter may need to summarize some "text book" material.

This is not something you'd normally require in an academic paper, and it may not be appropriate for your particular circumstances. Indeed, in some cases it's possible to cover all of the "background" material either in the introduction or at appropriate places in the rest of the dissertation.

1. Biological background
   (a) residue identity prediction
   (b) atomic environments
   (c) SMILES representation?
2. Machine learning background
   (a) neural networks
   (b) graph neural networks
   (c) equivariant graph neural networks
   (d) representing residues as graphs

2.1 Biological background

2.1.1 Proteins

Amino acids are the building blocks of proteins and play a fundamental role in various biological processes. These small organic molecules are essential for life, and their unique

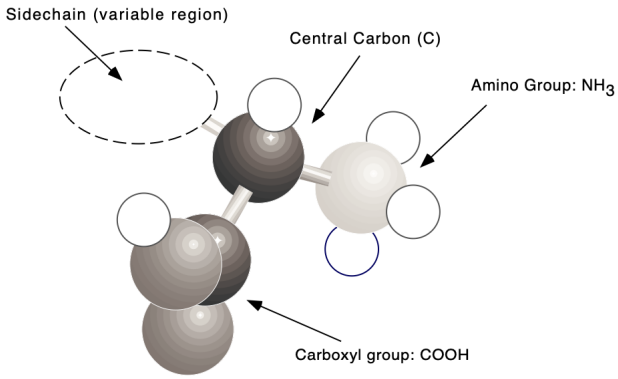


Figure 2.1: The basic chemical structure of an amino acid. Carbon atoms are black, Oxygen is dark grey, Nitrogen light grey, and hydrogen white. Image taken from [1].

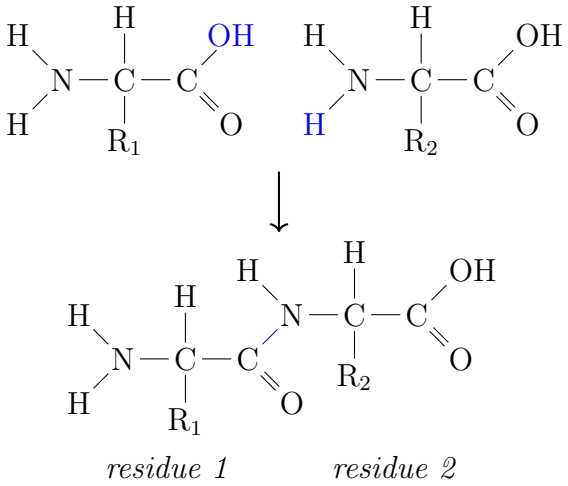


Figure 2.2: Two amino-acids are chained together through a peptide bond. The chemical reaction releases a water molecule (H_2O) in the process.

properties allow them to form complex and diverse structures within the body. All amino acids follow the same underlying pattern and consist of a central carbon atom (C), an *amino* group (NH_3), a *carboxyl* group (COOH), and a variable side-chain, as we can see in Figure 2.1. Chains of amino acids are formed through a chemical reaction that creates a *peptide bond*, as shown in Figure 2.2. The portions left of the original amino acids are called *residues*. Note that there are 20 naturally occuring amino acids.

Proteins are one of the most important macromolecules found in living organisms and they are involved in a vast array of biological processes. These large, complex molecules are composed of long chains of amino acids that are folded into shapes. Proteins play a variety of roles in the body, including catalysing chemical reactions, transporting molecules, providing structural support, and regulating gene expression. The diversity of protein structures and functions is vast, with some proteins consisting of just a few amino acids, while others contain thousands. The unique sequence of amino acids in each protein determines its three-dimensional structure and, more importantly, its *specific function*. Understanding the structure and function of proteins is essential for advancing our knowl-

edge of cellular processes and for developing treatments for a range of diseases caused by protein dysfunction.

2.1.2 Residue identity prediction

Residue identity prediction (RES) is a computational task in bioinformatics that involves predicting the amino acid residue at a particular position within a protein sequence. The accurate prediction of residue identity is an important problem in bioinformatics because it can provide insights into protein function, structure, and evolution. Knowing the identity of residues within a protein sequence can help to identify important functional sites and motifs, which can provide clues about the protein's function and interactions with other molecules.

Residue identity prediction is also important for *protein engineering*, as it can help researchers design proteins with specific functions. Machine learning approaches to RES have already been used to engineer plastic decomposing enzymes for higher thermal stability [2].

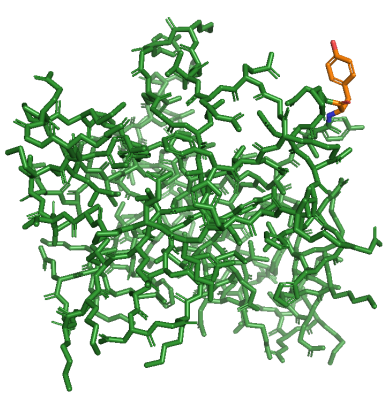
The ATOM3D dataset. One important benchmark dataset for RES is ATOM3D [3]. The ATOM3D collection is a compilation of datasets that contain the 3D structure of biomolecules, including nucleic acids, small molecules, and proteins. These datasets have been tailored to function as a benchmark for machine learning techniques that train on the 3D molecular structure of molecules in order to solve tasks such as molecular function prediction, ligand binding affinity, or protein-protein interactions.

The PDB format. ATOM3D datasets have a standardised format, in which each sample is a `.pdb` file. The `.pdb` format is a format provided by the **Protein Data Bank** (PDB), a database of 3D structural data of various biological molecules. Figure 2.3 shows an example of a molecule from the PDB.

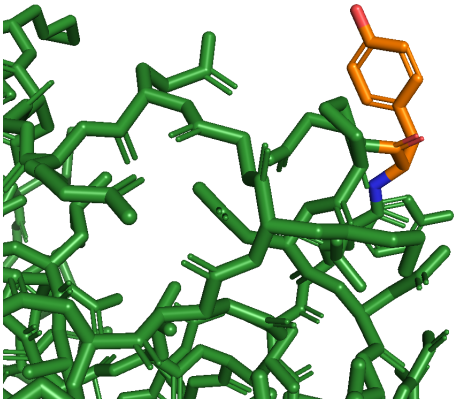
2.2 Machine learning background

First proposed by Rosenblatt [4], neural networks are a type of machine learning model that is inspired by the structure and function of the human brain. They are used for various tasks, such as image and speech recognition, natural language processing, and biomolecular prediction among others.

At a high level, neural networks consist of interconnected nodes or neurons organised into layers. Each neuron receives input data, applies a mathematical operation to it, and produces an output. These operations are typically weighted sums followed by activation functions that introduce non-linearity into the model. The outputs from one layer of



(a) The 1KDA molecule. The orange structure represents one amino acid residue.



(b) A close-up of the atomic environment around the orage amino-acid.

Figure 2.3: Example of a PDB molecule.

neurons serve as inputs to the next layer, forming a hierarchical structure. The resulting predictions made by the model are then scored using a *loss function*.

The training process in neural networks involves two steps. First, the derivative of the loss function with respect to each weight is computed, using the *backpropagation algorithm* [5]. This provides the gradient information needed for the next step. The second stage is the weight update, typically done using a gradient descent technique.

2.2.1 Graph Neural Networks

Maybe flesh this out later.

Graph neural networks (GNNs) are a type of neural network that operate on graphs and are able to capture structural information in the data by leveraging the existent relationships between entities. All GNNs use some form of *neural message passing*, where vector messages are exchanged between nodes and updated using a neural network [6].

Formally, for a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, at message passing iteration k, a hidden embedding $\mathbf{h}_u^{(k)} \in \mathbb{R}^{F_k}$ corresponds to each node $u \in \mathcal{V}$. The framework first creates a *message* $\mathbf{m}_{u,v}$ between a node u and its neighbour v by using a message function $\psi^{(k)} : \mathbb{R}^{F_k} \times \mathbb{R}^{F_k} \to \mathbb{R}^{F'_k}$:

$$\mathbf{m}_{u,v} = \psi^{(k)}\left(\mathbf{h}_u^{(k)}, \mathbf{h}_v^{(k)}\right) \tag{2.1}$$

These messages are then aggregated using operation $\oplus^{(k)}$; this operation is usually the sum or the average, but other versions exist as well. Note that in general function must be *permutation invariant*, since the neighbours of a node u do not have any intristic order. Finally, the aggregated message is combined with the node's own embedding $\mathbf{h}_u^{(k)}$ using function $\phi^{(k)} : \mathbb{R}^{F_k} \times \mathbb{R}^{F'_k} \to \mathbb{R}^{F_{k+1}}$, also called the *update* function, as seen in Equation

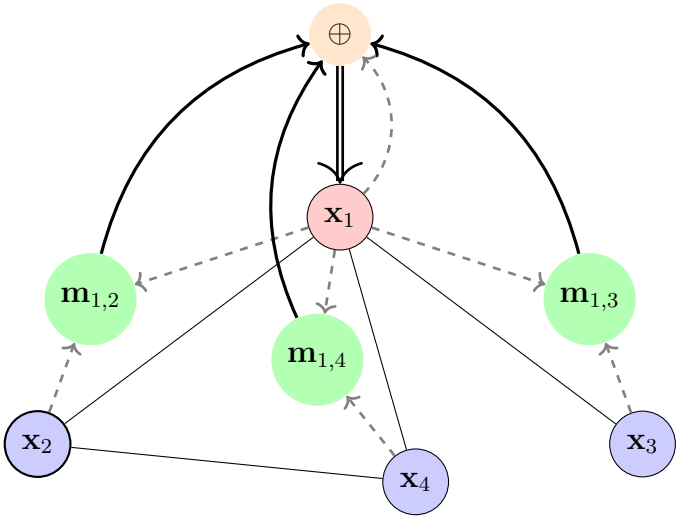


Figure 2.4: Diagram of the information flow of Equation 2.2. Node $\mathbf{x}_1$ has three neighbours coloured in blue. The messages between node $\mathbf{x}_1$ and each of its neighbours are aggregated using operator $\oplus$; the result is used to update the embeddings of $\mathbf{x}_1$. Note how in this diagram we are not interested in the edge between $\mathbf{x}_2$ and $\mathbf{x}_4$, since this is the update iteration for node 1.

2.2:

$$\mathbf{h}_u^{(k+1)} = \phi^{(k)}\left(\mathbf{h}_u^{(k)}, \oplus_{v \in \mathcal{N}(u)}^{(k)} \mathbf{m}_{u,v}\right) \tag{2.2}$$

A visual representation of the information flow can be seen in Figure 2.4.

2.2.2 Equivariant Graph Neural Networks

2.2.3 Representing residues as graphs

Chapter 3

Related work

This chapter covers relevant (and typically, recent) research which you build upon (or improve upon). There are two complementary goals for this chapter:

1. to show that you know and understand the state of the art; and
2. to put your work in context

Ideally you can tackle both together by providing a critique of related work, and describing what is insufficient (and how you do better!)

The related work chapter should usually come either near the front or near the back of the dissertation. The advantage of the former is that you get to build the argument for why your work is important before presenting your solution(s) in later chapters; the advantage of the latter is that don't have to forward reference to your solution too much. The correct choice will depend on what you're writing up, and your own personal preference.

Chapter 4

Design and implementation

This chapter may be called something else... but in general the idea is that you have one (or a few) "meat" chapters which describe the work you did in technical detail.

Chapter 5

Evaluation

For any practical projects, you should almost certainly have some kind of evaluation, and it's often useful to separate this out into its own chapter.

Chapter 6

Summary and conclusions

As you might imagine: summarizes the dissertation, and draws any conclusions. Depending on the length of your work, and how well you write, you may not need a summary here.

You will generally want to draw some conclusions, and point to potential future work.

Bibliography

[1] Lawrence Hunter. Molecular biology for computer scientists. *Artificial intelligence and molecular biology*, 177:1–46, 1993.

[2] Hongyuan Lu, Daniel J. Diaz, Natalie J. Czarnecki, Congzhi Zhu, Wantae Kim, Raghav Shroff, Daniel J. Acosta, Bradley R. Alexander, Hannah O. Cole, Yan Zhang, Nathaniel A. Lynd, Andrew D. Ellington, and Hal S. Alper. Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature*, 604(7907): 662–667, April 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04599-z. URL `https://doi.org/10.1038/s41586-022-04599-z`.

[3] Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Brandon M. Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. ATOM3D: tasks on molecules in three dimensions. *CoRR*, abs/2012.04035, 2020. URL `https://arxiv.org/abs/2012.04035`.

[4] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 1958.

[5] Henry J. Kelley. Gradient theory of optimal flight paths. *ARS Journal. 30 (10): 947-954*, 1960.

[6] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Appendix A

Technical details, proofs, etc.

Appendices are for optional materials that is not essential to understanding the work, and that the examiners are not expected to read, but that will be of value to readers interested in additional, in-depth technical detail.

A.1 Lorem ipsum

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

A.2 Homo sapiens non urinat in ventum

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in

vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, At accusam aliquyam diam diam dolore dolores duo eirmod eos erat, et nonumy sed tempor et et invidunt justo labore Stet clita ea et gubergren, kasd magna no rebum. sanctus sea sed takimata ut vero voluptua. est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat.

Consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis

nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, At accusam aliquyam diam diam dolore dolores duo eirmod eos erat, et nonumy sed tempor et et invidunt justo labore Stet clita ea et gubergren, kasd magna no rebum. sanctus sea sed takimata ut vero voluptua. est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat.

Consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.