Conceptual Modeling to Support Forward Modeling of NMR Data for Machine Learning*

D. Levi Craft[1 [0000−0003−3077−3402]] and Michael R. Gryk[1 [0000−0002−3483−8384]]

UConn Health, Farmington CT 06030, USA
darcraft@uchc.edu
gryk@uchc.edu

Abstract. For predictive and experimental methods alike, discovering the structure and biological mechanisms of proteins is vital to our fundamental understanding of life. Driven by the vast number of solved protein structures through x-ray crystallography and Nuclear Magnetic Resonance, as well as advances in machine learning and neural networking that enable us to predict a protein's structure based solely on its amino acid sequence, this project lays the groundwork for predicting what would be the observed experimental NMR data of a protein based on its structure. Outlined is our ongoing conceptual model, implemented as relational databases, used for our workflow-based approach to solving the Forward Modeling problem of NMR. This approach will support ongoing machine learning approaches in predicting protein-ligand binding mechanisms and other kinetic studies.

Keywords: Conceptual Modeling · Machine Learning · Big Data · Forward Modeling.

1 Introduction

1.1 The Protein Folding Problem

Proteins are fundamental to life and their unique structure provides an insight into the mechanisms through which they perform their vital functions. Determining a protein's structure solely from its sequence of fundamental building blocks, amino acids, has been a grand challenge in biology for nearly 50 years [1]. The biennial Critical Assessment of protein Structure Prediction (**CASP**) experiments was established as a way to enlighten the biological community of advancing methods for protein structure prediction. Moreover, CASP cultivated the assimilation of these approaches through scholarly communications to further propagate new methodologies.

One such approach, AlphaFold, has been recognized by the organizers of CASP to be an answer to this grand challenge, namely the 'protein folding

* Supported by a UCONN Convergence Awards for Research in Interdisciplinary Centers (CARIC) grant, PI: Dr. Jeffrey Hoch

problem'. In the most recent CASP assessment, AlphaFold predicted the structure of proteins based solely on its sequence of amino acids with a resolution of approximately 1.6 1 Å [2]; i.e., the width of an atom. AlphaFold has expanded upon the existing 17% of the total number of human protein sequences experimentally-determined to predict the structure of 95% [2]. These human proteins are included in the initial release of 350,000 predicted structures, spanning 20 genomes, determined using their neural network model [2].

1.2 Experimental Validation

Nuclear magnetic resonance (**NMR**) is a non-invasive and non-destructive technique that reports on a nucleus of interest's local magnetic environment, allowing it to be used to investigate the structure, dynamics, and kinetics of a wide range of biological systems [3]. Unique to NMR is its ability to detect molecular motion in proteins and other polymers over a wide range of time scales. However, NMR is an inherently complex technique and most scientific domains, including the field of biological NMR (bioNMR), are heavily reliant on computational software and methodologies developed by experts of the field.

With the introduction of a repository full of theoretically predicted structures determined by AlphaFold, NMR has a unique opportunity to advance its role in biomedical science. This coincides with the Big Data to Knowledge NIH (**BD2K**) initiative [4]. NMR can not only be used as a method of validating AlphaFold predictions, but also the combination of repositories for AlphaFold and spectra of assigned proteins in the bioMagResBank (**BMRB**) [5] can be used to predict the NMR spectra of a protein given its structure. This ability to predict NMR spectra and having predicted protein structures would advance the ability of NMR to investigate protein binding dynamics involved in ligand binding assays and other studies that report on the mechanisms of biological structures.

1.3 Forward Modeling

Until recently, the work of a scientist is to solve inverse problems, in which experiments are conducted, observations are made, and models are inferred from the underlying basis of observations. The reverse pathway, as demonstrated by AlphaFold, involves using theory to predict observations that would be observed if the models are correct, is called forward modeling. In addition to validating models, forward modeling has applications to machine learning (**ML**) when there is insufficient empirical data to train a ML system. This paper outlines progress being made to develop a conceptual model of NMR data and computational workflow for forward modeling of experimental NMR data from compositional, structural, and dynamical models of biological molecules.

2 Background

2.1 Protein Structure

A fundamental understanding of a protein's structure is necessary for the subsequent discussion. Proteins are composed of a sequence of amino acids. There

are 20 common amino acids, recognized by three letter and single letter abbreviations, each of which share a basic structure; depicted in **Fig. 1**. The functional groups common to amino acids are amino (-NH_2) and carboxyl (-COOH) functional groups, as well as a side chain or R group. In **Fig. 1**, Lysine's R-group is a methyl (-CH_3) group. An amino acid's R group is its identifying factor. Notice in **Fig. 1**, there is a common naming convention for the atoms of an amino acid.

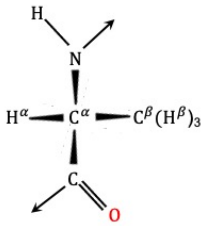

Fig. 1. Structure of Lysine, one of the 20 common amino acids.

A sequence of amino acids are bound together in sequential order through amide bonds between the carboxyl and amino groups, represented as arrows in **Fig. 1**; referred to as a proteins primary structure. This primary structure defines a protein's secondary structure through the local environments, hydrophobic vs hydrophilic as well as steric restrictions, defined by neighboring R groups. Common secondary structures are α helices and β sheets. The combination of secondary structures define the three dimensional conformation of a protein's tertiary structure.

2.2 NMR Workflow

BioNMR is inherently complex. A spectroscopist is required to solve the four inter-dependent problems in **Fig. 2**. The first of which is the ability to prepare a concentrated enough sample that produces a strong signal to assign each of the nuclei to a peak, as well as balance the resolution of the sample to resolve overlapping peaks. Once their primary NMR data is collected, the subsequent spectral reconstruction and analyses are performed iteratively for each of the experiments, in order to obtain biophysical characterizations of the system in hand. This process, reconstructing and analyzing spectra iteratively, can be thought of as the spectroscopist's workflow.

Depicted in **Fig. 2** is the classical approach of spectroscopists, solving the inverse problem of acquiring primary NMR data and through analysis of its spectral reconstruction, derive structure and molecular dynamics of the biological system. Inversely, this project's workflow will start with the structure of a protein and predict the spectra that would be observed. The input of this workflow will be depositions from the Protein Data Bank (**PDB**[1]) [6], representing

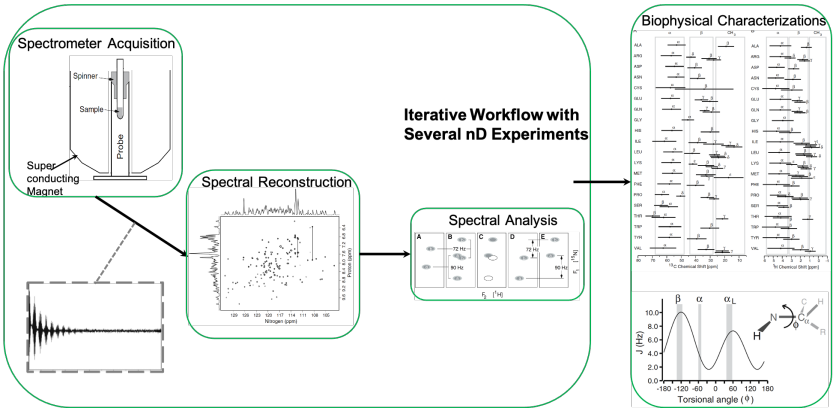

Fig. 2. Workflow taken by spectroscopists performing bioNMR, composed of its four unique challenges; images taken from [3].

Biophysical Characterizations in **Fig. 2**, which the three dimensional (3D) coordinates of a protein's structure solved either by NMR or X-ray crystallography. In the near-term, this project's output will be a predicted spectral reconstruction that will be compared to derived data deposited into the BioMagResBank (**BMRB**[2]).

2.3 Biological Databases

Biological Databases containing experimental and dervied NMR data, as well as experimental and theoretical structural data will be used as input to this project's workflow as well as a means of evaluating its results. The following databases are instrumental to this projects success:

- PDB: An archive for 3D structural information of proteins, nucleic acids, and complex assemblies. Depositions contain ancillary information on the method of derivation, as well as the derived structural data. Each deposition has a unique identification code that appears in its downloaded filename as 'XXXX.pdb'. PDB files will be parsed for inclusion into this project's conceptual model. Note, the majority of PDB depositions are derived from X-ray crystallography, therefore light atoms, i.e., hydrogen atoms, are not present. X-ray crystallography's ability to determine the coordinate of an atom is dependent on its electron cloud. Therefore, hydrogens containing a single electron in their orbit, are undetectable. Instead, PDB depositions report the coordinates of heavy atoms, i.e., carbon, nitrogen, oxygen, common to an amino acid.
- BMRB: A repository of NMR data for biologically relevant systems. BMRB depositions follow their data dictionary, NMR-STAR [7]. Depositions containing ancillary information ranging from experimental conditions to thermodynamic data; depicted in **Fig. 3**. Each of the supergroups in **Fig. 3** area

collection of saveframes, which are themselves categorized by their collective categorical tags. At the smallest level, a piece of information such as the temperature at which an experimental set of NMR data is collected is saved as a tag-value pair.

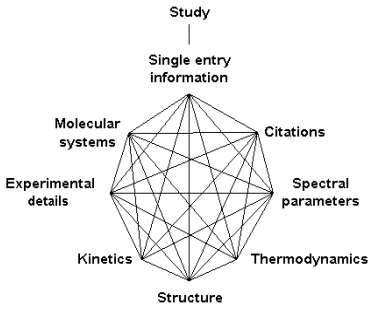

Fig. 3. BMRB's NMR-STAR Format: Depiction of a deposition's supergroups reproduced from the BMRB [5].

3 Conceptual Model of an NMR Experiment

The widely used nomenclature for representing atom specific data of polypeptide chains was established in the NMR community in 1998 [8]. This nomenclature for the twenty common amino acids follows the IUPAC recommendations, initially outlined in 1969 [9]. IUPAC's recommendations represent an implied data model. Efforts to make an explicit data model have been founded on IUPAC's implied model, thereby ensuring their implementation is scientifically sound [10].

Fox-Erlich *et. al* [10] developed an entity-relationship diagram to explicitly define IUPAC's implied model. A benefit of their model, paralleled in our own model shown in **Fig. 4**, was the designation of the chemical bond network. Though the bonding relationship of atoms was partially coded in the IUPAC naming convention, it was not explicitly defined. For example, referring back to **Fig. 1**, with a linear sequence of amino acids in consideration one can infer the organization of side chain atoms; i.e., C^{α} will always be bound to C^{β}. However, it can not be inferred from the atom nomenclature such as the $N - C^{\delta}$ bond in proline [10]. All bonding relationships in a protein were explicitly defined in Fox-Erlich's model [10].

Our preliminary model of an NMR experiment is built around BMRB's NMR-STAR format as well as an extension of Fox-Erlich's model [10]. Our conceptual model, **Fig. 4**, defines an element as the most basic component of a protein. From there, an amino acid is built of atoms ('AA_Atom' in **Fig. 4**) bound to one another. The bonds between elements of atom can be defined by their angles and torsion angles, 'AA_Bond_Angle' and 'AA_Bond_TorsionAngle', respectively. An individual amino acid is classified based on its chirality, 'AA_Chirality', which

informs on the clockwise order of an atoms around the α-carbon (see C_α in **Fig. 4**). With this basic construct of an amino acid, each of the 20 common amino

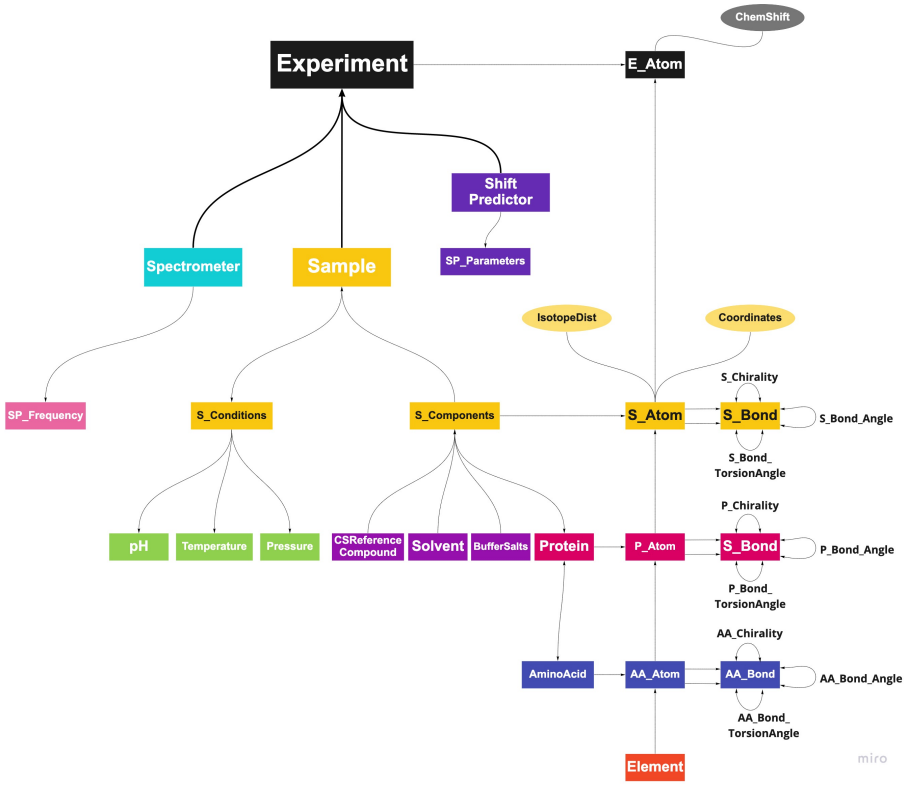

Fig. 4. Conceptual Model of an NMR Experiment based on Fox-Erlich's model of a protein [10]. Each NMR experiment modeled will is composed of a spectrometer, a sample, and a method of predicting the chemical shifts.

acids can be uniquely defined and referred to when defining the residue sequence of a protein. With this model of a protein in place, when a PDB deposition is parsed for the coordinates of atoms and are deposited in our own database, we augment its structural information by capturing their bond formation information. An example that exemplifies this significance, the pH of a sample defines which of the hydrogens in a given residue are present or absent; referred to as protonated or deprotonated. This pH dependency is residue specific as each residue has a unique pKA, defining the pH at which a residue will accept or donate a hydrogen. Recall, PDB coordinates are free of hydrogen atoms, whereas NMR data report frequency data for all atoms specified in the experimental configuration; i.e., a 3D HNCA experiment reports on H, C_α, and N. As a result, computational methods for protonation and energy minimization of the resulting protonated structures are used. However, these protonation methods often

do not accept pH as an input. This modeling of a residue's atom and their bonds will allow us to model NMR experiments at any specific pH by ensuring that only the anticipated hydrogens of a sample ('S_Atom') are present in our experiment ('E_atom'). **Fig. 4**'s schema will enable us to not only use BMRB depositions for evaluating our workflow's predicted spectra but also parse a protein's BMRB deposition to capture the experimental parameters, sample conditions and components, as well as spectrometer parameters to refine our predictions.

4 Workflow Approach

The development of our workflow will evaluate a given protein on an experiment-based approach. For a given protein, PDB ID, we have developed Python scripts to parse the BMRB and PDB depositions stored on the NMRbox [11]: a cloud-based computing platform for processing and analyzing NMR spectra. This addresses one of the existing shortcomings of BMRB depositions, corresponding PBD IDs are only captured if the author provides them. A query against a PDB ID in the BMRB thereby does not guarantee the capture of all corresponding PDB IDs. With this resolution in place, this workflow initializes on the structural data provided in a PDB deposition and will predict its spectra for a given NMR experiment. This prediction will utilize conditions corresponding to a specific NMR experiment in a BMRB and be evaluated against the reported spectra.

An emphasis of our workflow is our ability to report on the provenance of our predictions. For example, **Fig. 4**'s Shift Predictor category is composed of tens of competing software programs for predicting a protein's chemical shift and the specific input parameters used for each prediction. This in addition to the numerous protonation methods based on differing computational models, all of which we will evaluate using case studies, backed by NMRbox's virtual machines computational power. **Fig. 4** will be incorporated into a PostgreSQL [12] relational database. Our ability to categorize and report on the exact conditions of comparisons to be made will ensure their validity and transparency to the biological NMR community.

5 Conclusion and Future Work

We will develop our workflow based on the several hundred thousand predicted structures. These theoretical predictions represent the cleanest and most queryable structural data. Whereas, a PDB deposition can be comprised of the several hundred uncommon amino acids. These modified amino acids are often ignored in existing prediction algorithms, e.g., most of the common chemical shift predictors will not predict their chemical shifts. Additionally, seriality is not guaranteed for a PDB's amino acid sequence; residues can be missing from the protein's amino acid sequence as PDB depositions are derived data. Furthermore, PDB depositions have the possibility of containing error such as misprints of amino acid abbreviations, which prove difficult to detect and correct. Lastly, proteins are often multi-chained which are often incorrectly treated as single chains using prediction algorithms; incorrectly ignoring the effects of interactions inter-chains.

For these reasons, a subset of 2000 proteins predicted by AlphaFold, known to be single chains with corresponding BMRB depositions, will be used for the development of this projects workflow. With this completed workflow in place, edge cases will be accounted for and look-up tables will be created to handle the messy data that is common to experimentally derived PDB depositions.

The conceptual model developed for the atom and amino acid sequence, as well as the factors giving rise to an NMR experiment, will allow us to accurately model its expected NMR spectra. Our model is easily expandable, as is its associated PostgreSQL relational database, allowing us to incorporate additional factors and parameters discerned from literature review and metadata analysis of BMRB depositions. An expansion of our model to be made, will be to incorporate Gryk *et. al.*'s node representation of an amino acid in order to account for the 400 variants of amino acids [13]. Moreover, our model can without difficultly be extended to the prediction of time-domain, primary, NMR data.

References

1. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. Science **181**, 223–230 (1973).
2. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).
3. G. Rule and T. Hitchens, Fundamentals of Protein NMR Spectroscopy, 2005.
4. Bourne PE, Bonazzi V, Dunn M, et al. The NIH Big Data to Knowledge (BD2K) initiative. J Am Med Inform Assoc. 2015;**22**(6):1114.
5. "BioMagResBank", Eldon L. Ulrich et.al. Nucleic Acids Research **36**, D402-8 (2008)
6. wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D520–D528
7. Ulrich, E.L., Baskaran, K., Dashti, H. et al. NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. J Biomol NMR **73**, 5–9 (2019).
8. J. L. Markley, A. Bax, Y. Arata, C. W. Hilbers, R. Daptein, B. D. Sykes, P. E Wright, and K. Wüthrich. Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. Pure & Appl. Chem. **70**, 117-142 (1998).
9. T. Rules. IUPAC-IUB Commission on Biochemical Nomenclature (CBN) (1970) Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. J. Mol. Biol. **52**(1), 1–17 (1970).
10. Fox-Erlich S, Martyn TO, Ellis HJ, Gryk MR. Delineation and analysis of the conceptual data model implied by the "IUPAC Recommendations for Biochemical Nomenclature". Protein Science : a Publication of the Protein Society. 2004 Sep;**13**(9):2559-2563.
11. Maciejewski MW, Schuyler AD, Gryk MR, Moraru II, Romero PR, Ulrich EL, et al. NMRbox: A Resource for biomolecular NMR computation. Biophysical Journal. 2017;**112**(8):1529—34.
12. PostgreSQL http://postgresql.org
13. Heidi J. C. Ellis, Susan Fox-Erlich, Timothy O. Martyn, and Michael R. Gryk. 2006. Development of an Integrated Framework for Protein Structure Determinations: A Logical Data Model for NMR Data Analysis. In Proceedings of the IEEE Computer Society, USA, 613–618.