Graph-Structured Random Hermitian Matrices to Model Electron Dynamics During Protein Folding

Siddharth Rath

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Devin Mackenzie, Chair

Douglas Fowler

Arthur Barnard

Juan Carlos Idrobo

Program Authorized to Offer Degree:
Materials Science & Engineering

University of Washington

Abstract

Graph-Structured Random Hermitian Matrices to Model Electron Dynamics During Protein Folding

Siddharth Rath

Chair of the Supervisory Committee:
Devin Mackenzie
Department of Materials Science & Engineering

Random Matrix Theory is a well-known method in physics and mathematics to model quantum chaos by studying the statistics of energy states of both fermionic and nucleonic systems on the ergodic timescale. Additionally, recent advancements established Wigner's Surmise, Gaussian Orthogonal Ensemble symmetry and Quantum Unique Ergodicity in random band-matrices that model electrons interacting with one another along 1 and 2-dimensions; and associated the width of the band around the leading diagonal with the non-locality of the eigenvectors, thereby inferring the conductivity of the system under study. Quantum chaos has so far been modeled by random matrix theory in up to 2 dimensions, and exact or numerical solutions to dimensions ≥ 3 do not exist.

However, such methods have not yet been applied to the study of inter-electronic behavior during protein folding in 3 dimensions, where electronic interactions are critically important, causing a marked change in the conductivity of the protein as it folds. Protein folding is primarily entropy driven and involves weak hydrophilic-hydrophobic and Van der Waals interaction between residues and the backbone to create predominantly hydrogen bonds and other secondary bonds involving empty or half-empty π-orbitals. Additionally, the water molecules around the proteins are

known to adopt structures that are more regular than bulk water, to the extent that most of the water in living cells is structured. However, the resulting loss of configurational entropy when proteins and water molecules around them adopt regular stable structures, is chalked up to vague entropic compensation terms without clarification on the types of entropy that could compensate for the apparent reduction in entropy. As a result, our understanding of how protein sequences adopt unique structures and how such structures directly map to the protein's function is lacking.

While a mostly classical description of protein folding is inadequate to address the underlying physical mechanisms of specificity during molecular recognition and information flow during signal transduction cascades, a quantum mechanical description is currently computationally prohibitive. There is also the issue of decoherence at the temperatures at which biomolecules operate, due to thermal noise. Several problems also persist when it comes to predicting how a multitude of biomolecules specifically interact to execute signaling cascades leading to cellular functions. To determine underlying physics, Physics Inspired Neural Networks, PINNs, have recently taken center-stage. However, the fact that intrinsically disordered regions (IDRs) of proteins are involved in specific binding events, makes a simple sequence-structure-function predictive PINN untenable. The dynamic structures of IDRs necessitate a variational approach to protein structure prediction. Here we use random graph-structured matrices to circumvent the $<$ 3-dimension limit to model the electronic interaction network in a protein while it folds in an aqueous medium.

Assuming a) n points in 3-dimensional finite cubic Euclidean space of volume l^3 with coordinates $(x, y, z)\ \forall\ x, y, z \sim U(\{-(l/2), ..., (l/2)\})$ *i.i.d.*, where $U(\{-(l/2), ..., (l/2)\})$ is the uniform distribution over the set of real numbers between $[-l/2,\ l/2]$; b) Threshold value $\varepsilon \in \mathcal{R}^+$ which is the radius of the sphere in the finite cubic Euclidean space, within which all points interact; Leading to c) An adjacency matrix

$A_{n\times n}(\varepsilon) : A_{ij}(\varepsilon) = 0 \ \forall \ D_{ij} > \varepsilon$ and $A_{ij}(\varepsilon) = 1 \ \forall \ D_{ij} < \varepsilon$ where the pairwise Euclidean distance between the n points is given by the distance matrix $D_{n\times n} : D_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$; and d) A random real valued square Hermitian matrix $H : H^T = H$, we show that the continuous differential entropy, denoted as $CDE(A(\varepsilon))$, of the joint probability distribution of the nonzero elements of the leading diagonal and the lower triangle (or upper triangle) of the matrix $H \circ A(\varepsilon)$ (which directly represents the distribution of single point correlations of the eigenvalues of the matrix $H \circ A(\varepsilon)$) increases as the number of nonzero elements in the lower (or upper) triangle increases as a result of increasing ε. We also show that if we sample many random Hermitian matrices $H^1, H^2, ..., H^N, N \in Z^+, N > 40$, and denote the ordered set of eigenvalues of each matrix $H^k \circ A(\varepsilon), k = 1, 2, ...N$ as $\lambda^k_{H\circ A(\varepsilon)}$, then the standard deviation $\sigma_{sp}(\varepsilon)$ of the distribution of spacing between any two consecutive eigenvalues from all the ordered eigenvalue sets $\lambda^k_{H\circ A(\varepsilon)}$ first decreases then increases for increasing ε values. Moreover, the ratio between the $||L^\infty||$ and the $||L^2||$ norm of the eigenvectors $V^k_{\lambda_{H\circ A(\varepsilon)}}$ corresponding to the eigenvalues $\lambda \in \lambda^k_{H\circ A(\varepsilon)}$ approaches a lower bound of $1/\sqrt{n}$ as ε increases, indicating that the eigenvectors become more delocalized on the ergodic scale.

Our results indicate that the $CDE(A(\varepsilon))$ and $\sigma_{sp}(\varepsilon)$ as defined above can be used as metrics to optimize ε which is directly related to the number of nonzero elements in the lower (or upper) triangle of the adjacency matrix A. Our results also show that a) Reduction in configurational and conformational entropy during protein folding is associated with an increase in information entropy obtained from electron energy level statistics, and b) Reduction in uncertainty in electron energy levels is associated with delocalization length of electrons along secondary bonds. We also apply our numerical simulations to the case of protein folding where we use an all atom molecular dynamics simulation of a standard polyalanine peptide folding in water

to show that the numerical results hold for the proteins as well, paving the way to use random graph structured matrices as suitable representations of biomolecules to study their structure function relationships in greater depth. Our results posit that the nature of information transfer and flow during signaling cascades could be the certainty in electron energies of participating molecules. We directly infer the nature of information flow during signal transduction cascades and postulate about the role of quantum phenomena and decoherence in biomolecular function. Such notions add fuel to the idea that quantum mechanical phenomena are intricately and non trivially involved in protein structure-function relationships, and bolster the conjecture that water might be a necessary condition for life on earth.

The results show a mathematically accurate and physically relevant, first-principles based representation of proteins and other biomolecules that would be beneficial in machine learning approaches to predict secondary structures and in the nascent field of protein design for medical and technological applications. Therefore, we propose a future physics inspired variational autoencoder-generative adversarial network deep-learning model based on our results that could be applied as a more interpretable biomolecular structure and function predictor.

ACKNOWLEDGMENTS

I express sincere appreciation to the University of Washington (UW), and the Materials Science and Engineering (MSE) Department where I have had the opportunity to pursue interdisciplinary research. Gratitude is due to Prof. Mehmet Sarikaya (MSE, ChemE, OHS, MolES at University of Washington) who inspired me at the beginning of my PhD research and introduced me to the exciting study of peptides and proteins from a physics and materials science point of view, as well as providing me with computing resources to carry out my interdisciplinary research.

I am incredibly grateful to all of my Undergraduate students, especially Mr. Christian Frech (Physics, Carnegie Mellon University & MSE, University of Texas, Austin) and Mr. Jonathan Francis-Landau (Mathematics, UW); Graduate student colleagues Mr. John Hamann (MSE, UW), Mr. Oliver Nakano-Baker (MSE, UW), Mr. Jacob Rodriguez (MSE, UW), Dr. Tyler Jorgenson (MolES, UW), Dr. Deniz Yuçesoy (MSE, UW) and Dr. Richard Lee (MSE, UW); and friends, too many to name, for their academic and emotional support throughout the program. I also express enormous gratitude to my fellow graduate students in my dissertation support group who helped me stay focused while completing my manuscript, as well as staff members in the MSE department, Nichole Fernkes, Karen Wetterhahn, Bichtien Thach, Dawn Counts, Donald Obcena, and Yen Cone.

I am particularly indebted to Dr. Hanson Fong (MSE, UW) whose guidance, financial help, and mentorship are perhaps the most tangible resources I have had access to. In addition, I express sincere and everlasting gratitude to Prof. Christine Luscombe (MSE, MolES, UW), Prof. Hadi Zareie(MSE, UW) Prof. Eric Shea-Brown

(AMATH, CompNeuro, UW), Prof. Thomas Trogdon (AMATH, UW, and the Graduate School Representative on my thesis defense committee), Prof. Kevin Jamieson (CSE, Allen School, UW), Prof. Jim Pfaendtner (ChemE, eScience, MolES, Chem, UW), Prof. David Beck(ChemE, eScience, UW), Prof. Guozhong Cao (MSE, UW), Prof. Aniruddh Vashisth (ME, UW), Prof. Lucien Brush (MSE, AMATH, UW), Prof. Lilo Pozzo (ChemE, MSE, UW), and my committee members Prof. Devin Mackenzie (MSE, ME, UW), Prof. Marina Meila (Statistics, UW), Prof. Douglas Fowler (GenomeSci, UW), Prof. Arthur Barnard (MSE, Physics, UW), Prof. Juan Carlos Idrobo (MSE, Physics, UW), and Prof. Armita Nourmohammad (Physics, UW) for their help, guidance, support, motivation, and commitment to my sustained academic and professional progress during the Ph.D. program.

DEDICATION

To my families, both related and chosen, and above all, to Nature Herself.

TABLE OF CONTENTS

PREFACE

The chapters in the document are preceded by the list of figures that lists the figure titles and their page numbers in the document's main text. The list of figures is followed by a glossary of terms that includes common definitions of many terms encountered throughout the document, and some definitions have been repeated inside the document's main text as well to contextualize the terms. The glossary is followed by Chapter 1, section 1.1 of which explains the current field in protein science, from the gaps in our fundamental understanding of how proteins work, to high throughout sequencing technologies, computational methods of structure prediction, data driven methods of structure prediction, and physics inspired methods of structure prediction that are ubiquitous in the literature. section 1.2 expounds upon the literature that led to the assumptions and hypotheses formed based on those assumptions, that then the rest of the document tries to test. Section 1.3 states the main problem statement that this body of work tries to address.

Chapter 2 details the current state of the art in data driven protein structure prediction using deep learning in section 2.1. Section 2.2 then introduces the VAE-GAN deep learning model, and how it was implemented in the current work by the author, for further probing of difficulties and limitations of unconstrained graph based deep learning models in protein structure prediction, especially for intrinsically disordered regions. Section 2.3 then shows the results of the VAE-GAN model for intrinsically disordered protein structure prediction and outlines the necessity of random matrices in helping deal with some of the limitations by explaining the physics of protein folding better.

Chapter 3 provides some background on what are random matrices, before describing how random band matrices were recently proven to display universality, quantum unique ergodicity, and a transition from a global spectral distribution that resembles a Poisson point process at low band width and localized eigenvectors, to Gaussian Orthogonal Ensemble symmetry class of random matrices with semicircle distribution at high band width and non local eigenvectors. This led to the major hypothesis by the author that the information entropy and other statistical measures of spectral distributions might explain the protein folding process as well as molecular recognition and signal transduction.

Chapter 4 talks about the hypotheses and in section 4.1 then specifies how such a hypothesis led to the idea of implementing a random graph-structured matrix model of electron dynamics during protein folding, and both analytically and numerically probe the global eigenvalue statistics and eigenvectors of the random graph structured matrices. Sections 4.2 to 4.4 describe the trends in Eigenvalue and Eigenvector statistics of random graph-structured matrices from numerical simulations and when applied to a model system of peptide folding and water at ambient conditions, which is followed by the major conclusions and discussions, as well as avenues for future research in chapter 5. The document ends with the bibliography and an appendix with details of how to obtain the code from GitHub repository, followed by a short vita of the author.

LIST OF FIGURES

GLOSSARY

AB-INITIO: 'From the beginning.' Ab-initio methods model phenomena from first principles, or fundamental physical theories, which are usually quantum mechanical in nature.

ADJACENCY MATRIX: A square (same number of rows and columns), and usually a symmetric matrix representing a finite graph's connectivity. The row/column indices represent the nodes, the matrix elements represent the presence/absence of edges between the nodes, or some value describing the edges. In the most basic case, a simple graph's adjacency matrix is a (0,1) matrix with zeros on the diagonal.

ALPHA FOLD: A deep learning algorithm developed by DeepMind, a Google subsidiary, that predicts a protein's 3D structure as seen in the Protein Data Bank, from its amino acid sequence information.

ALPHA HELIX: The α helix is the most abundant motif in the secondary structure of proteins and polypeptides. It is a right-handed helix conformation in which every backbone N–H group hydrogen bonds to the backbone C=O group of the amino acid located four residues earlier along the protein sequence.

AMINO ACID: Amino acids are organic compounds that contain amino (–NH_2 or –$[NH_3]^{+1}$ when solvated at neutral pH) and carboxylic acid (–[O=C–OH] or –$[O{=}C\text{-}O]^{-1}$ when solvated at neutral pH) functional groups, along with a side

chain (R–group) specific to each amino acid at its C_α atom that connects the amino and carboxylic acid functional groups.

BETA SHEET: The β-sheet or a pleated beta-sheet is the second most common motif of a regular protein secondary structure in 3D. Beta sheets are beta-strands connected sideways by at least two or three backbone hydrogen bonds, forming a generally twisted, pleated sheet. There are also beta turns and anti-parallel beta-sheets under the beta-structure umbrella.

BIAS–VARIANCE TRADE-OFF: In statistics and machine learning, the bias-variance tradeoff is the conflict in attempting to simultaneously minimize the bias and variance errors. The bias error comes from erroneous assumptions about the model prior to training, such as the system's linearity, while the variance error stems from hypersensitivity to small fluctuations in the training data set that might not be present in the test or benchmark data sets.

BINARY CROSS ENTROPY LOSS: It is a type of loss that is minimized during statistical learning. The BCE-Loss creates a criterion that measures the Binary Cross Entropy between the target and the output.

BROWNIAN MOTION: A classical mechanical description of random motion of particles suspended in a fluid medium.

BULK EIGENVALUE STATISTICS: In spectral analysis, statistical distributions and metrics of the median eigenvalue of a matrix or eigenvalues around the median eigenvalue, is called bulk statistics.

CHARMM: Chemistry at Harvard Macromolecular Mechanics (CHARMM) is a ubiquitously used set of partial differential equations and dynamical parameters

of such equations developed as a force-field to be applied in classical molecular dynamics simulations.

CLASSICAL MECHANICS: It is a physical theory that adequately describes the motion of macroscopic objects in a continuous, deterministic, and reversible manner.

COHERENT STATES: In quantum physics, the specific quantum mechanical state of a quantum harmonic system that displays oscillatory dynamics closest to a quasi-classical or classical description of a harmonic oscillator such as a mass suspended on a spring. Eigenvectors of the lowering operator usually denote such states, and they form an overcomplete family, i.e., there exist redundant eigenvectors, the removal of which will not impact the completeness of the set of eigenvectors.

CONNECTIVITY MATRIX: An adjacency matrix with ones on the diagonal, i.e., includes self-connection. They are sometimes used interchangeably. We refer exclusively to a connectivity matrix in this work, but we use the term 'adjacency matrix.'

CONTINUOUS DIFFERENTIAL ENTROPY: It is an analogous form of Shannon's entropy devised by Claude Shannon to describe information entropy of continuous distributions. It is defined for a probability distribution p(x) on a random variable x as:

$$\mathbb{H}(x)_{cde} = -\int p(x) log\{p(x)\} dx \tag{1}$$

It is a limiting case of the actual continuous information entropy, called limiting

density of discrete points defined as:

$$\hat{\mathbb{H}}(x) = -\int p(x) log \left\{ \frac{p(x)}{m(x)} \right\} dx \tag{2}$$

Where if there are N discrete points $\{x_i\}$ in i-dimensions, then as $N \to \infty$, the density of the N points approaches $m(x)$, an invariant measure.

COVALENT BOND: A type of bond between atoms that is crudely defined as sharing of electrons among them. Quantum mechanically, the probability of locating each shared electron from a specific energy level is equal around the nuclei of both atoms.

CROSS ENTROPY: In information theory, the cross-entropy between two probability distributions p and q over the same random variables, measures the expected number of bits needed to identify an event drawn from the random variable if a coding scheme used for the random variable is optimized for an estimated probability distribution q on a sample of such events, rather than the true distribution p on the population of the events. It is defined as:

$$\mathbb{H}(x)_{cross} = -\sum p(x) log_b \{q(x)\} \tag{3}$$

where the base b depends on the nature of the discrete random variable x.

DEEP LEARNING: It is a popular subclass of machine learning methods where artificial neural networks are ubiquitously used to automatically estimate the necessary representation of training data to predict the features or classify raw data with reasonable accuracy.

DELOCALIZATION: Eigenvectors are said to be fully delocalized if they do not possess significant modes anywhere in the array. A fully localized eigenvector

has a significantly large mode at a certain point. In quantum mechanics, the eigenvectors of a system of particles are the wavefunctions of particles associated with the eigenvalues (allowed energy levels). Delocalization pertains to the uncertainty in measuring the particle's location once its energy is fully known. The particle exists as a wave, smeared across the space it occupies, with the crests of the wave associated with a higher likelihood of locating the particle.

DNA: Deoxyribonucleic Acid or DNA is the sequence of nucleotide: Adenosine, Guanine, Cytosine, and Thymine, that exists as a double helix, wound into chromatin fibers and organized into chromosomes, that contains the genetic code for protein translation and gene expression within cells.

EDGE STATISTICS: In spectral analysis, the statistical distributions and metrics of the largest and smallest eigenvalues are called edge statistics. In the case of random matrices, the bulk eigenvalues follow gaussian ensembles and semicircle distributions, while the edge eigenvalues follow the Tracy-Widom distribution.

EIGENVALUE SPACING: The difference between consecutive eigenvalues arranged in ascending order.

EIGENVALUES: Eigenvalues are scalar factors that scale the Eigenvectors of a matrix. Physically speaking, upon a linear transformation of a set of basis vectors in Euclidean space, specific vectors within the space only get stretched or compressed and do not rotate. The amount of stretching or compression is the eigenvalue associated with the eigenvector. In quantum mechanics, the eigenvalues of the Hamiltonian (a matrix that quantifies the total energy of every particle) indicate the energies of each state in the superposition of individual wavefunctions.

EIGENVECTORS: Eigenvectors are defined as vectors in a vector space defined by basis vectors that only get scaled but not rotated upon a linear transformation of its basis vectors.

ERGODICITY: In mathematics, ergodicity is the idea that, given enough time, a point in a moving system (dynamical or stochastic process) will uniformly and randomly visit every possible location in the space that the system exists in. This work uses ergodicity because enough time has indeed passed from a moving electron's point of view.

FERMI-DIRAC DISTRIBUTION: It is a type of distribution obeyed by quantum mechanical particles that obey Pauli's exclusion principle that no two particles with half-integer spins can have a set of all identical quantum numbers. If two such particles exist in the same quantum state, they must have different spins, which measures their angular momentum. Such particles with half-integer spins are therefore termed fermions. Electrons are the most commonly encountered fermions. For a system of identical fermions in thermodynamic equilibrium, the average number of fermions in a single-particle state i with energy ϵ_i and chemical potential μ is given by:

$$\bar{n_i} = \frac{1}{e^{\epsilon_i - \mu/k_B T} + 1} \tag{4}$$

Where k_B is the Boltzmann's constant of value $k_B = 1.380649 \times 10^{-23} JK^{-1}$

FORCE FIELD: The force field refers to the functional form (partial differentials) and set of parameters utilized to compute the potential energy of a classical system of atoms or coarse-grained particles in molecular mechanics, molecular dynamics, or Monte Carlo simulations. The parameters are usually estimated from empirical evidence or computed from quantum mechanical principles.

GAUSSIAN ENSEMBLES: The most studied types of random matrices used to model energy Hamiltonians with and without time-reversal symmetry (Gaussian Orthogonal and Gaussian Unitary ensembles, GOE and GUE, respectively) or with time-reversal symmetry but without rotational symmetry (Gaussian Symplectic Ensemble, GSE). The distribution of the eigenvalues of the random matrices belonging to the different Gaussian ensembles is invariant under either Unitary, Orthogonal or Symplectic conjugation.

GENERATIVE ADVERSARIAL NETWORK: It is a sub-type of unsupervised deep learning models where two artificial neural nets compete in a zero-sum game. The *generator* network generates candidates that could belong to the original population of data samples, while the *discriminator* network evaluates whether the candidates are generated by a network or sampled from the true distribution of the population of data points.

GLOBAL SPECTRAL STATISTICS: Analysis of all eigenvalues of a matrix together at the same time. In random matrix theory, the semicircle distributions of Gaussian ensembles are a type of global spectral statistics (empirical spectral measure), that give the density of states.

GRAPH CONVOLUTION: A type of semi supervised variant of convolutional neural network class of deep-learning models on graph-structured data represented usually as weighted connectivity matrices.

GROMACS: It is a free, open-source molecular simulation software developed at the University of Groningen, Netherlands between 1991-2000, that can run on both CPUs and GPUs.

HAMILTONIAN: In quantum mechanics, the Hamiltonian of a system is a matrix-valued operator that describes the total energy of the system, both kinetic and potential., where the size of the matrix corresponds to the number of particles in the system. The eigenvalues of the Hamiltonian correspond to its spectrum or the set of the system's allowed total-energy states.

HERMITIAN MATRIX: Square symmetric matrices that are their own adjoint matrices. It is a self-adjoint operator on a finite dimensional complex vector space.

HYDROGEN BOND: Formed due to electrostatic attraction between a Hydrogen atom covalently bound to a more electronegative atom or radical or functional group and another electronegative atom bearing a lone pair (covalently unshared but still in the valence energy level) of electrons. From a quantum mechanical perspective, the electrons in the lone pair have a non-zero probability of being located on the hydrogen atom and its covalently bonded partner. Similarly, there is a non-zero probability of the hydrogen's electron being located on both electronegative partners, though there is a higher likelihood of locating the electron on the covalently bonded partners.

INFORMATION ENTROPY: the entropy of a random variable is the expected value of uncertainty inherent to the random variable's possible outcomes. If a discrete random variable X takes values in the set χ and is distributed according to $P : \chi \to [0, 1]$ the entropy is defined as:

$$\mathbb{H}_{discrete}(x) = -\sum_{i=1}^{n} P(x_i) log_b\{P(x_i)\} \tag{5}$$

where the base b depends on the type of the random variable.

INTERPRETABILITY: Statistical models such as those included in artificial intelligence and machine learning models are called interpretable when humans can reasonably understand the reasoning behind the decisions made by the models while relating input data to their learned representations and the predicted features.

INTRINSICALLY DISORDERED PROTEINS: A chain of amino acids that do not readily fold into any fixed or ordered three-dimensional secondary structure, usually in the absence of its interaction partners such as other ligands, receptors, and RNA. Intrinsically disordered regions in otherwise larger and structured proteins usually occur at or around their active sites. Such disordered regions fold quickly into well-organized conformations as soon as they interact with their specific partners and start the signal transduction cascade. The discovery of intrinsically disordered regions upended the static 'lock-and-key' molecular recognition model.

IONIC BOND: A type of bonding in chemistry that is purely electrostatic where electrically charged species attract each other with force described by Coulomb's law. However, from a quantum mechanical point of view, it results in the electron(s) of the 'positive' atom having a higher likelihood of being located on its counterpart 'negative' atom. The electrons are thus said to be 'transferred.' Such a system is more stable than individual atoms in terms of its ionization energy and electron affinity, both of which depend on the highest unoccupied and lowest occupied 'molecular' orbitals in the quantum sense.

KL–DIVERGENCE: The Kullback-Liebler divergence, also known as relative entropy, denoted as $D_{KL}(P||Q)$ is a 'distance' measure between a probability distribution P and the reference distribution Q. For discrete P and Q on the

same probability space χ the KL-divergence is defined as:

$$D_{KL}(P||Q) = \sum_{x\varepsilon\chi} P(x) log\left(\frac{P(x)}{Q(x)}\right) \tag{6}$$

and if P and Q are continuous then:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} P(x) log\left(\frac{P(x)}{Q(x)}\right) dx \tag{7}$$

If the reference distribution Q is an invariant measure and consists of the density of infinitely many discrete points in space, then the KL-Divergence becomes the definition of continuous information entropy, as given by Equation 2.

LOCAL SPECTRAL ANALYSIS: The analysis of only a few eigenvalues and their spacings, instead of all the eigenvalues of a matrix together.

LOSS FUNCTION: The 'error' function chosen by the user to optimize the machine learning models. One minimizes the loss function, or cost function, which measures the 'distance' between actual and predicted labels in a machine learning scenario. Some examples include mean-squared loss and cross-entropy loss.

MOLECULAR DYNAMICS: A computer simulation tool used to simulate the time evolution of Newtonian motion of classically represented particles (atoms and molecules), where the force field calculates the total potential and kinetic energies of each particle at a given time and their gradients along the cartesian dimensions give the force necessary to calculate their positions in subsequent time frames. The model then minimizes overall energy to converge upon the equilibrium positions of the particles.

MOLECULAR ORBITALS: Similar to an atomic orbital, a molecular orbital describes the 'location' and momentum (energy) of the wave-like nature of the

electrons in a molecule, as opposed to an atom. Core molecular orbitals exist that are filled, along with partially filled molecular orbitals and empty or virtual molecular orbitals. A linear combination of atomic orbitals usually results in such molecular orbitals.

MOLECULAR RECOGNITION: It refers to the specific static or dynamic interaction between two or more molecules, usually one of them being a protein, through interactions that are neither covalent nor ionic. Such interactions involve Van der Waals interactions, hydrophobic/hydrophilic interactions or hydrogen bonds, and similarly weak interactions.

NUCLEAR MAGNETIC RESONANCE: A physical spectroscopy method where a weak oscillating magnetic field perturbs atomic nuclei suspended in a constant strong magnetic field, producing detectable electromagnetic waves with a characteristic frequency of the magnetic field at the nucleus. This method is helpful in the experimental observation of atomic resolution protein 3D structures by locating nuclei of the atoms in a folded protein by identifying the characteristic peaks in the electromagnetic spectrum.

PEPTIDE: A short amino acid chain that is up to 50 amino acids long when extended and is usually the result of proteolysis, i.e., it is a remnant from a digested protein. Some proteins are created small; in that case, they are called microproteins or micropeptides. In this work, we use the term peptide for any amino acid chain less than 50 amino acids in length, regardless of its origin.

PROTEIN: A long chain of amino acids, more than 50 in length, whose stable secondary and tertiary structures depend upon their specific sequence, and the structures affect specific functions in the body. All enzymes, antibodies,

ion transporting molecules, and several other functional biomolecules either are proteins or interact with proteins to render their functions.

PROTEIN DATA BANK: Usually referred to as the PDB, the database consists of thousands of protein sequences, their experimentally observed structures in 3D space, their associated biological functions, and other empirically measured properties. Primarily it includes X-ray crystallography and NMR structures.

QUANTUM CHAOS: A branch of physics that attempts to describe how chaotic classical systems evolve from fundamental phenomena that are necessarily quantum mechanical. Random Matrix Theory is a tool within the field of quantum chaos.

QUANTUM DECOHERENCE: The phenomena where the quantum nature of a system no longer exists upon any measurement, regardless of a conscious human observer. For example, the collapse of a superposed wavefunction upon collision with another particle. In such a situation, the quantum nature is shared with the other system, i.e., a type of quantum entanglement. Consequently, there is a net loss of the quantum nature of each system. Furthermore, as more systems get entangled, the quantum nature of their components collapses.

QUANTUM ENTANGLEMENT: A fundamental physical phenomenon occurs when a group of particles is generated, interacts, or shares spatial proximity in a finite space such that their individual quantum states are no longer fully describable independently of the other particles.

RANDOM BAND MATRIX: A random symmetric matrix where only a few rows and columns are filled out around the leading diagonal, and the rest of the matrix elements are zeroed out. The width of the band can be expressed as the number of filled rows or columns around the leading diagonal.

RANDOM COIL: A protein structure that cannot be described as alpha, beta, or any well-defined conformation. Intrinsically disordered regions of proteins usually occur as random coils in nature.

RANDOM GRAPH MATRIX: A random matrix multiplied elementwise with a connectivity matrix that represents random weights on a graph's edges.

RANDOM MATRIX: A matrix where each element is a random variable of the same type, described in the same support.

RESONANCE: In mathematics and physics, resonance is a positive interference between two waves that results in a much higher amplitude of the combined wave. In chemistry, a *resonance structure* of a molecule is defined as a superposition of several canonical structures because all structures of such a molecule exist equiprobably in nature. It is a direct result of the quantum nature and delocalization of electrons within a molecule, especially when empty molecular orbitals of similar energies are available.

RNA: Ribonucleic acid, which is a counterpart of the DNA. All genetic code in eukaryotes and the majority of prokaryotes exist as DNA, which is then transcribed into RNA, especially messenger-RNA, which then gets translated into proteins in the ribosome of the cells. Several viruses, however, lack DNA and can carry genetic information only as RNA.

ROSETTA: A molecular Monte-Carlo based simulation software initially developed by the Baker Lab at the University of Washington, Seattle. The force fields of the software are obtained from the protein structure data contained within the PDB.

SCHRÖDINGER EQUATION: A partial differential equation that describes the wavefunction of a quantum mechanical system, usually depicted in terms of a matrix-valued equation as:

$$\hat{\mathbf{H}}\psi(\mathbf{r},t) = \mathbb{E}\psi(\mathbf{r},t) \tag{8}$$

Where $\hat{\mathbf{H}}$ is the Hamiltonian of the system, and $\mathbb{E}$ are the eigenvalues of the Hamiltonian, and ψ are the eigenvectors, that change with location vector $\mathbf{r}$ and time t.

SECONDARY STRUCTURE: Physical 3-dimensional structures of peptides immediately formed upon synthesizing the sequence. Non-covalent bonds are predominantly the cause of secondary structure formation. Bulkier proteins usually form higher order structures in a hierarchical manner, known variously as tertiary and quaternary structures.

SEQUENCE: A biomolecule such as DNA, RNA, or protein's sequence is the sequence of monomers (nucleotides and amino acids) appearing from one end to another. A DNA sequence can be read forward or backward, with one side of a double helix being precisely complementary to the other side. RNA is usually single-stranded. Protein sequences are crucial in dictating what structures and functions the proteins will possess.

SIGNAL TRANSDUCTION: A cascade of chemical and physical signals mediated by molecular recognition within a cell along a complicated biomolecular pathway. Specific molecular recognition usually starts a signal transduction cascade across a specific pathway, resulting in a specified function. Correct secondary structures and positioning all molecules involved in such a cascade is critical. Multiple such cascades continuously occur in billions of cells in all organisms.

TIP3P: A three site water model that represents a water molecule as a usually rigid triangle. Each vertex has a point charge, and Lennard-jones parameters for Van-der-Waals interactions with other molecules within a cut-off distance. The angle subtended at the vertex corresponding to an Oxygen atom is about 104.5°. It has been shown that a TIP3P water model has highest efficiency with and incorporated into the CHARMM molecular mechanics software.

TRAJECTORY: In the sense of molecular dynamics simulations, a trajectory is a path in the energy landscape that is traced by the particles as they are brought to equilibrium. It is usually depicted as a plot with timestep as the x-axis and energy of the system or some other tractable metric (Dihedral angles usually) on the y-axis.

VARIATIONAL AUTOENCODER: It is a type of deep learning model, an artificial neural network, that attempts to estimate a parametric distribution that the input data comes from to sample a new data point that is convincingly from the actual distribution of training data. It is called an autoencoder because it can automatically encode massive amounts of detailed data in a few parameters.

WAVEFUNCTION: A wave function is a way to describe a particle's quantum state mathematically. It is a complex-valued probability wave, which upon multiplication with its complex conjugate, results in the actual spatial probability of the property being measured, usually the location of a particle.

WIGNER'S SURMISE: Eugene Wigner postulated that the probability that the eigenvalues of a Hermitian random matrix (mean field, i.e., no element is zero) will overlap goes to zero, i.e., no eigenvalues are redundant, or degenerate. That postulate was proven accurate and became known as Wigner's surmise.

WISHART DISTRIBUTION: It is a generalization of the Gamma-Distribution to multiple dimensions defined over symmetric, non-negative definite random matrices.

X-RAY CRYSTALLOGRAPHY: In protein science, quickly cooling the crystallized proteins and then bombarding the crystals with X-rays, results in diffraction patterns that are studied to estimate electron densities in the crystal, mean nuclear positions, and types of bonds present within the crystals. It is the primary experimental technique used to determine protein structures and comprises the majority of structural data in the PDB.

Chapter 1

INTRODUCTION

Multitudinous and varied interactions between proteins, Deoxyribo Nucleic Acid (DNA), Ribo Nucleic Acid, and their conjugates with lipids, carbohydrates and other small molecules, mediates all functions in biological systems, such as cellular metabolism,[1,2] cell-regulation,[3] intracellular transport,[4–7] DNA/RNA replication,[8–12] cell-fate decision,[13–15] immune response to antigens,[16–19] and stimuli sensing and response.[20–23] The specific structures in 3-dimensional space attained by the biomolecules influences and regulates molecular recognition and binding of two or molecules, especially in the case of proteins and their interactions.[18,24] The ensuing protein complex modifies the participating proteins' geometric structures and initiates specific signal transduction cascades within the cells.[25–27] In such a cell signaling pathway, a biochemical signal hops across the biomolecular networks, resulting in a specific operation or action carried out at the end of the pathway, usually by another set of proteins.[19,28–30] Researchers not only attempt to model such biological networks but also endeavor to quantify the information that is carried around in a cell-signaling pathway. Nevertheless, despite an enriched understanding of the underlying physical mechanisms of how protein structural modifications result in a cascade of biochemical changes within a cell, the physical nature of the information passed along in a signaling pathway is still relatively mysterious.[26,27,31–37]

Moreover, attempts at ab-initio modeling of how protein sequences give rise to specific hierarchical structures cannot fully describe the causal relationship between the structure and the next step in the signal transduction cascade.[38–42] While misfolding of

proteins is a fairly commonly observed phenomenon, it's nonetheless a fact that most proteins fold correctly, most of the time in most of the cells. But, under conditions of undue stress, they can unfold or misfold, contributing to diseases such as Alzheimer's. How, and indeed, why, do most protein sequences adopt specific structures in almost all the cells in which they exist, and how and why might that effect precise execution of the signal transduction cascade along the entire signaling pathway is still not entirely answerable in a classical Brownian motion model.[36–38,43].

Furthermore, attempts at using data-driven models to predict how specific sequences might fold in 3-dimensional space and how such a folded structure might impact the protein's intended function are still limited by hitherto unexplored *ab-initio* physical mechanisms.[38,39,44,45] Understanding such mechanisms and mathematically describing the protein structures by incorporating the physical mechanisms of folding in a data-driven scenario is essential to designing end-to-end machine learning models, from sequence to function, without using conventional molecular modeling techniques. Such models hold the promise to not only predict how might a given protein sequence fold but also might shed light on the mechanisms that mediate the causal relationship between the protein-structures and the resulting signal-transduction cascade, as well as the nature of the information carried across the signaling pathways.

The current body of work body of work validates a mathematical description of protein structures based on Random Matrix Theory and establishes several metrics to investigate the underlying ab-initio physical mechanisms of protein folding, dynamic molecular recognition, and subsequent signal transduction (see fig 1.1). The resultant mathematical representation and metrics can be used in various data-driven protein structure prediction models. They can also be used to understand the fundamental physical mechanisms underlying many protein functions in living systems. Other sections in the present body of work describe the background, assumptions, and motivations behind the work undertaken and attempt to phrase the problem statement that the current body of work endeavors to tackle

succinctly.

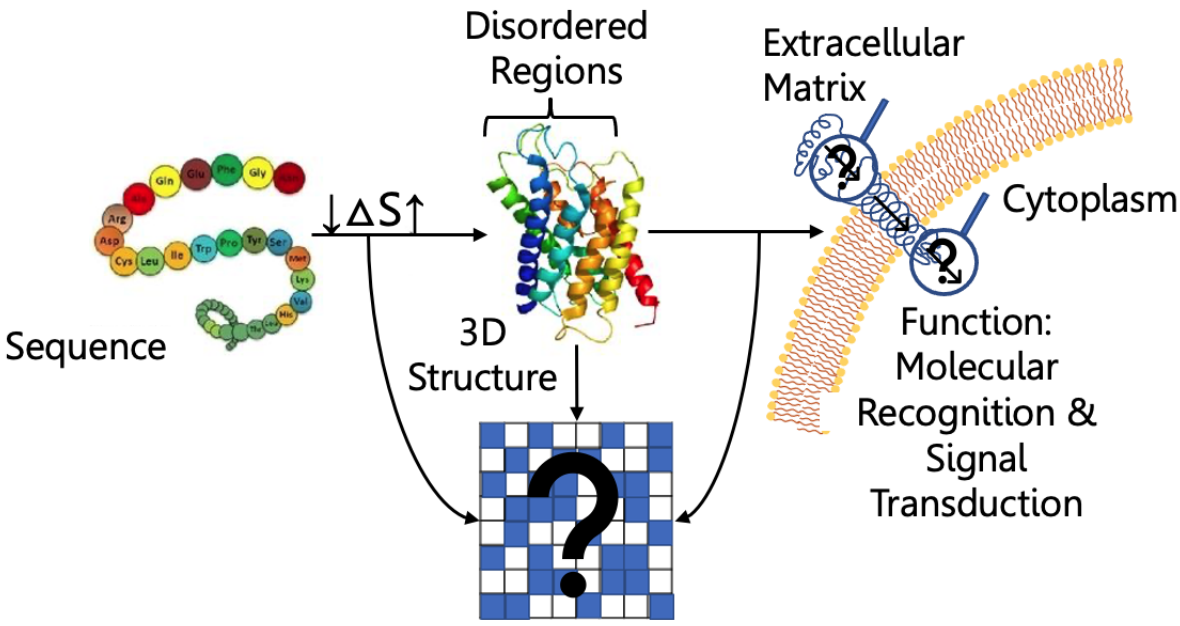


Figure 1.1: **Overall schematic of the project**: The figure shows the overarching summary of the current study. Random Graph Matrices (RGM) are used to infer physical explanations regarding the entropic compensation term during protein folding, specificity of binding during molecular recognition, and direction of information flow in an error-free execution of the signal transduction process. Establishing metrics, optimization of which in a data-driven model would dynamically evolve the connectivity of the molecular graphs towards physically real structures in 3D space.

1.1 Background

Proteins are large and flexible organic molecules comprised of several amino acids strung together in a particular sequence via a covalent 'peptide bond'. A peptide bond occurs when the –C(=O)–OH terminal of one amino acid and the NH_2–terminal of the next amino acid in the sequence react to create a –C(=O)-NH– bond while a water molecule is released as a byproduct. Such a reaction is catalyzed by RNA and

enzymes (which are themselves usually other proteins) within a cell's ribosome where previously transcribed messenger-RNA (they carry the RNA nucleotide counterparts of the DNA sequences that are exposed to transcriptase enzymes) are translated into proteins, as one of the last stages in gene expression, ubiquitously known as the central dogma of biology. As the chain of amino acids grows, it starts to fold and assume its secondary structure almost immediately depending on its sequence.[46–49] All naturally occurring proteins are composed of various sequences of amino acids, chosen from among the 20 natural amino acids. From a mathematical standpoint, therefore, there could exist 20^n unique protein sequences, where n is the length of a sequence in terms of the number of amino acids. Proteins perform almost all significant functions within and outside a cell, including mechanical support, transport, cellular metabolism, DNA/RNA replication, regulation of cellular processes and enzymatic catalysis, cellular reproduction, and signaling that effects in translating stimuli into cellular responses.[1–23]. proteins are also involved in error correction during DNA/RNA replication, thereby playing a role in mutations and subsequent evolution of species.

Peptides are much shorter parts of the protein sequences, usually within $2 - 50$ amino acids long, but are more directly involved in specific functions carried out by the proteins and their conjugates such as peptidolipids, peptidoglycans, neurotransmitters, etc.[38,50,51] Peptides may be a part of a larger protein or may occur separately. Researchers mainly observe that short functional peptides occur around the active sites of the proteins, and they are typically intrinsically disordered (i.e., without a fixed, extremely likely, and well-defined secondary structure) due to their shorter lengths.[44,45,52,53] Literature labels amino acid chains of less than 15 amino acids as oligopeptides, while polypeptides contain up to 50 amino acids continuously strung in a particular sequence into an unbranched chain. In addition, some peptides, especially non-ribosomal ones, can be cyclic when the N- or C-termini join together

or to a residue or sidechain.[47,53,54]

Peptides play many biologically essential, indeed monumental, roles such as antimicrobials and antibiotics,[55,56] neuropeptides and neurotransmitters,[57–60] lipo-peptides and peptidolipids,[61] signaling peptides,[62] behavior and phenotype determining hormones,[63] compounds in venom,[64] immuno-peptides helping prevent diseases,[65] gastrointestinal peptides helping with digestion and nutrient absorption,[66,67] respiratory peptides helping with oxygen diffusion,[67] renal peptides helping filter blood,[68,69] cardiovascular peptides controlling optimum blood circulation,[70] and neurotrophic peptides,[71] vaccines, therapeutics, and antibiotic peptides to combat diseases.[71,72] Recently, a large class of small but functional proteins, termed miniproteins, microproteins, or micropeptides, have come to light.[43,73–77] Such terminology is necessary for literature to distinguish between traditional peptides–pared down remnants of longer proteins through proteolysis, and miniproteins born small in the Ribosome. However, for clarity and cohesion purposes, we use peptide as an umbrella term for any short chain of amino acids, regardless of its origins.

Levorotatory (left-handed) chirality is present in all naturally occurring amino acids except Glycine, which is not chiral as it does not have a side chain.[78–81] The sequence of amino acids (that comprise a protein or peptide) confers overall chirality on the bigger polypeptide chain itself. Proteins and peptides assume various secondary structures and longer proteins hierarchically fold into stable tertiary, quaternary, and even higher order structures in 3-Dimensional space, usually in aqueous media. Surrounding water molecules and hydrogen bonds within the protein or peptide, as well as with the solvent, significantly impact the secondary and hierarchical structures.[82–84]

1.1.1 Secondary Structures

While the primary structure of a protein is its network of covalent bonds along the backbone (that consists of the peptide bonds stringing the amino acids together) and within the sidechains, the secondary structure of the proteins denotes the 3-

dimensional conformation of segments of the protein which come about due to hydrogen bonds and other inter and intramolecular weak interactions among the side chains and the backbone, as well as with the surrounding water molecules. There are several well-defined types of secondary structures of proteins, α-helices and β-sheets being the most commonly described ones. Typically, secondary structure components spontaneously form as the protein folds into a hierarchical structure, immediately as the proteins get manufactured in the ribosome, leading eventually to a tertiary structure. Several proteins then fold together to form quaternary structures comprising a protein complex, with variously accessible active sites, modular sub-units with individual functions and trapped water molecules. Scientists empirically investigate the secondary and higher-order structures by spectroscopy methods such as Circular Dichroism,[85–87] X-Ray Crystallography,[88–91] cryogenic electron microscopy and Nuclear Magnetic Resonance (NMR),[92–95] the latter three methods forming the majority of data source in the Protein Data Bank (PDB).[96]

In X-Ray Crystallography,[97] lyophilizing the proteins (fast cooling leading to crystallization) and then bombarding the resulting protein crystals with high energy X-rays, results in diffraction patterns that are studied to estimate electron densities in the crystal, mean nuclear positions, and types of bonds present within the crystals. In NMR, a weak oscillating magnetic field randomly perturbs the atomic nuclei that are suspended in a constant strong magnetic field, producing detectable electromagnetic waves with a characteristic frequency (depending on the size of the atomic nuclei) of the magnetic field at the site of the atomic nuclei. This method is helpful in the experimental observation of atomic resolution protein 3D structures by locating nuclei of the atoms in a folded protein by identifying the characteristic peaks in the electromagnetic spectrum. Such experimental methods are slow and cumbersome but critical to the effort of studying protein secondary structures.

Formally defined as the pattern of hydrogen bonds, an algorithm called DSSP (Dictionary of Protein Secondary Structure)[48,49,98] that treats hydrogen bonds from a purely

classical electrostatic point of view usually assigns secondary structure labels to protein segments with prior empirical investigation. Accordingly, there are α-helices, 3_{10}-helices, π-helices, parallel or antiparallel β-sheets, hydrogen-bonded turns, β-bridges, and bends (see fig 1.2). In addition, protein segments or peptides with a higher degree of intrinsic disorder (more number of unstabke hydrigen bonds with water than the number of stable ones with itself) usually appear as random coils, an umbrella term for nonassignable structures. Fig 1.3 shows the differences between the major classes of protein secondary structures

α-helices are the most common type of secondary structure motifs found in proteins ubiquitously in all organisms. It is a right-handed conformation wherein every–N-H group on the backbone is in a hydrogen bond with the backbone –C=O group of the amino-acid four residues earlier in the sequence. Each amino acid corresponds to 100° in the helix, and there are 3.6 residues per turn, with a helical pitch of about 5.4Å.[48,49,104] Left-handed helices are usually not possible in any naturally occurring protein on earth except proteins where there are abundant achiral Glycine residues.[105] There are other helical structures, such as the 3_{10} and π-helices, which possess a similar pattern of hydrogen bonding, but the hydrogen bonds exist between –N–H and –C=O groups on the backbone, three (3_{10}) and five (π) amino acids over along the sequence respectively.[104]

The α-helix, the most common protein secondary structure, is also the most stable conformation under ambient biological conditions.[106,107] It is a very tightly packed structure with all the side-chains pointing outward, utterly accessible to the solvent, and roughly pointing towards the protein's N–terminus. Due to their primary covalent and secondary hydrogen bonding patterns, α-helices have a macro dipole moment from the N– to the C–terminus, resulting in relatively high conductivity.[69,108–114] Although lengths of α-helices can range from four to forty residues at a time, literature presumes that short peptides cannot undergo enough stabilizing phenomena to com-

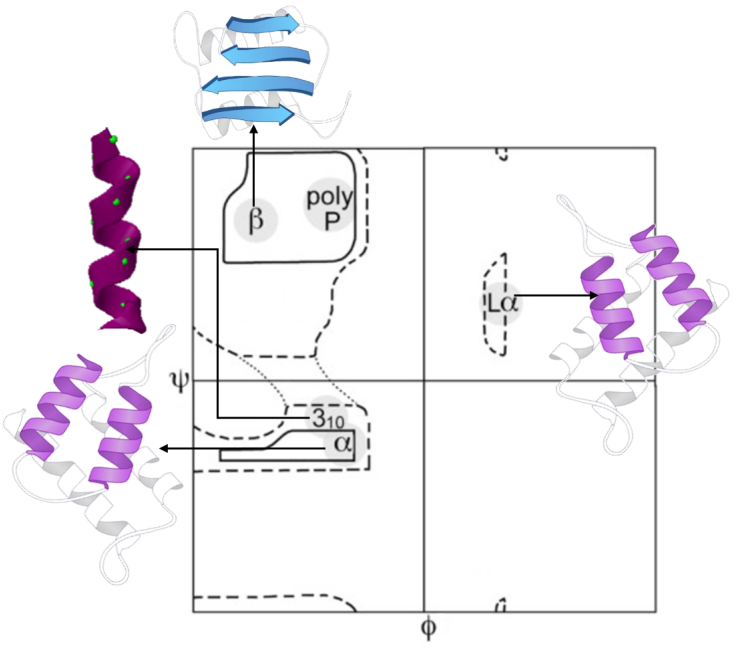


Figure 1.2: **Types of protein secondary structures**: Schematic showing the different types of secondary structure in proteins and their location in the coordinate space defined by the dihedral torsion angles subtended at the backbone (the Ramachandran plot[99–101]), ϕ and ψ, each of which takes values in the range $(-180°, +180°)$. Based on whether an amino acid in the structure determined through the X-Ray Crystallography, NMR or CryoEM[88] experiments possesses (ϕ, ψ)dihedral angle pairs that fall in the close vicinity of the ideal (ϕ, ψ) values of $\alpha, \beta, 3_{10}, \pi$ or other structures, are annotated as such, and visualized by several open source or commercial software such as PyMol[102] and Chimera.[103]

pensate for the entropy loss to sustain an α-helix.[115] α-helices predominantly occur in structural proteins such as keratins, muscle fibers, and membrane proteins inside the hydrophobic lipid bilayer surrounding the cells.[76,116,117] Their primary functions are helping enzymes bind to the DNA or RNA, helping membrane-proteins embed in the plasma membrane, and providing mechanical and structural support.[117] Since they ubiquitously exist in the integrin and actin protein complexes[118,119] and focal adhesion kinases,[120] they also participate in a big way in the cell's response to mechanical stimuli, which is a significant driver of cell migration along the tissues.[121,122] Despite

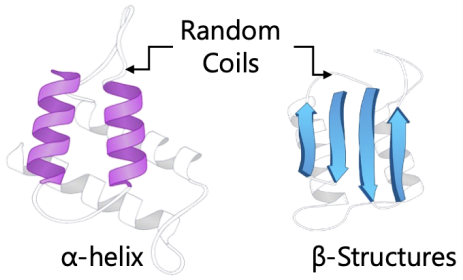
being the most stable and abundant secondary structure motif, α-helices only provide primarily mechanical and structural support to the proteins, a critical function that enables active sites to hide or become exposed to the solvent as necessary strategically.[123–127]

3_{10} and π-helices, on the other hand, are not as abundant; indeed, π-helices occur in less than 1% of all proteins,[124,126,128–132] while 96% of 3_{10}-helices have up to four residues in them.[129,130,133] Longer 3_{10}-helices are present only in specific transmembrane proteins of neurons that sense voltages and operate a voltage-gated potassium channel to control the firing of neurons.[134] π-helices also seem to transition easily into α-helices and vice versa by a single residue insertion or deletion in the sequence.[132] There is also evidence that π-helices, although rare, occur notably around active sites of specific proteins. The observation establishes their role as evolutionarily conserved motifs to effect phenotype changes due to point mutations in the genome.[135]

β-sheets are common secondary structure motifs, although not as common or stable as alpha helices in living systems.[106,107] It consists of beta strands connected horizontally by hydrogen bonds, resulting in a slightly twisted and pleated sheet. An individual beta-strand is about 3 to 10 amino acids long extended conformation. The sideways distance between the hydrogen-bonded partners in the beta-sheet is roughly 5Å.[48,49,104] β-sheets also have small dipole moments along the horizontal hydrogen bonds and are transversely conductive, although not as much as α-helices.[136] Sometimes found in supra-molecular assemblies of proteins leading to fibrillation and diseases such as Alzheimer's,[137–143] β-sheets also have a predominantly structural function.[46,144–148] Along with α-helices, they participate in a regular low-frequency oscillatory motion like an accordion,[149–151] whose purpose remains unknown.

Random coils, meanwhile, are not as ubiquitously found in proteins as α-helices and β-structures. They are usually in some conformation that is not stable and does not fall under the purview of any well-defined structure. The peptides and proteins that are intrinsically disordered[133,152,153] exist as random coils. They form up to 30% of

all protein structures and predominantly occur at active sites of the proteins.[44,45,52,53] They are primarily involved in molecular recognition and signal transduction, which proteins employ to render their functions, upending the notion that stable structures at active sites and their binding counterparts work as locks and keys.



Prevalence %	α > β > random coils
Stability	α > β > random coils
Conductivity	α > β > random coils
Location	α: Not prevalent at active sites β: Not prevalent at active sites random coils: Prevalent at active sites
Function	α: Mechanical support and tethering β: Structural support random coils: Molecular Recognition

Figure 1.3: **Comparison between α-helices, β-structures, and random coils**: The figure shows the major differences between the three main classes protein secondary structures in living organisms.

1.1.2 Molecular Recognition

Molecular recognition refers to the specific static or dynamic interaction between two or more molecules, usually one of them being a protein, through interactions that are neither covalent nor ionic.[154–156] Such interactions involve Van der Waals (VdW) interactions, hydrophobic/hydrophilic interactions or hydrogen bonds, and similarly weak interactions. Additionally, there is a growing repertoire of mounting incontrovertible evidence that water (the solvent) is, directly and indirectly, involved in molecular recognition.[157] It invariably mediates all biomolecules' biological and physicochemical functions, predominantly proteins.[158,159] Several biological molecular recognition events include ligand-receptor,[160] antigen-antibody,[161] protein-DNA,[162] RNA-Ribosome,[163,164] and sugar-lectin interactions.[165,166]

Molecular recognition has two types: (a) Static molecular recognition[156] involving

molecules with fixed geometries that fit together precisely, and (b) Dynamic molecular recognition[167,168] that involves shape-changing molecules that can filter out molecules that do not satisfy some condition for binding. Static molecular recognition is analogous to the fit between a key and a keyhole; it is a one-to-one complexation reaction between two molecules to form a supramolecular complex.[168] In dynamic molecular recognition, sometimes the binding of the first molecule to the first binding site of another molecule affects the interaction probability and strength of the third molecule with a second binding site of the second molecule, leading to cooperativity in binding between the molecules. In other types of dynamic molecular recognition in proteins, which usually occur at intrinsically disordered regions, the binding sites are not of fixed shape. They instead exist in multiple conformations that keep switching to filter out molecules that do not satisfy some specific condition for binding as a conformational proofreading mechanism.[169–171] The dynamic nature of molecular recognition is vitally significant since it provides a mechanism to regulate specific binding in biological systems. Such types of molecular recognition enhance the ability to distinguish between several competing target molecules via the conformational proofreading mechanism.[172] Dynamic molecular recognition is an active research area for application in highly functional chemical sensors and molecular devices.[133,173] However, Molecular recognition always triggers very specific signal transduction cascade within the cells.

1.1.3 Signal Transduction

Signal transduction pertains to transducing a signal from one type to another. In biology, it refers explicitly to the processes by which a stimulus changes into a physicochemical signal which moves within a cell as a series of biochemical cascades,[174,175] most commonly a reversible post-translational modification of the proteins (phosphorylation) catalyzed by other enzymes such as kinases.[176] Such a cascade ultimately results in a cellular response. The cellular response involves alteration of the expres-

sion of effector genes or activation/inhibition of targeted proteins.[177] At the molecular scale, cellular responses include changes in the transcription or translation of genes, post-translational modifications and conformational changes in proteins, and changes in their spatial location.[174,175,178] These molecular events control cell growth,[179] proliferation,[180] metabolism,[181–184] mechanotransduction,[185–190] and many other processes. In multicellular organisms, signal transduction pathways regulate cell communication in various ways.[190–193] Some common signal transduction pathways include the immunological complement system[194] and the insulin signaling pathway that controls blood sugar levels.[195–197]

With the advent of computational biology, analyzing signaling pathways and networks has become an essential tool for understanding cellular functions and disease, including signaling rewiring mechanisms underlying responses to acquired drug resistance.[198] However, so far, a complete understanding of the reasons behind the genesis of such precise execution of molecular changes at every step in every such biochemical signal transduction cascade, eludes scientists.[199–201] **Although successful in binding to counterparts, synthetically designed molecules that undergo molecular recognition events do not necessarily trigger any functionally important signal transduction cascades.**[202]**Furthermore, they are not necessarily specific to the binding site because there is a knowledge gap in the fundamental principles that lay beneath such superior specificity of molecular recognition and what is the nature of the stimulus that triggers precise signal transduction cascades.**[203] A statistical mechanical approach for molecular recognition that assumes conformational state changes of the random coil structures between the allowed states as the proof reading mechanism and the probability of each such state as the origins of an entropy driven recognition event therefore falls short when linking the recognition event with the signal transduction cascade. Simply getting a ligand to bind to the receptor of interest isn't enough to effect the proper signaling cascade in cells, and therefore there is a need to under-

stand the physical nature of the information that is passed along the cascade when the recognition event happens.

1.1.4 Computational Methods for Protein Structure Prediction & de-novo Protein Design

Understanding how protein structures come about is the first step to comprehending the principles governing specificity in molecular recognition and signal transduction precision. Various theories of protein folding exist, mainly from a classical mechanical and thermodynamical point of view, and several explicit computational models exist that fit the theories with varying degrees of agreement (fig 1.4). On the other hand, experimental methods of circular dichroism,[85–87] X-ray Crystallography,[88–91] and Nuclear Magnetic Resonance[92–95] can only tell us what the structures are but not how and why they formed nor how they guarantee the specificity and precision in the molecular recognition and signal transduction phenomena. Moreover, empirical methods of structure determination are cumbersome and painstakingly slow. They also cannot keep up with the pace at which modern sequencing technologies are discovering, identifying, and storing unmanageable amounts of the genome and proteome of multitudinous species.

Molecular Dynamics

One of the ubiquitous methods for secondary and tertiary protein structure prediction is Molecular Dynamics.[204,205] It is a simulation method for analyzing the motion of atoms and molecules by treating them as classical hard-spheres.[205] The algorithm numerically solves Newton's equations of motion for the system of interacting particles under investigation. It calculates the forces and potential energy gradients using inter-atomic and intermolecular potentials and force fields whose parameters come from either quantum mechanical calculations or empirical measurements.[206,207] MD is

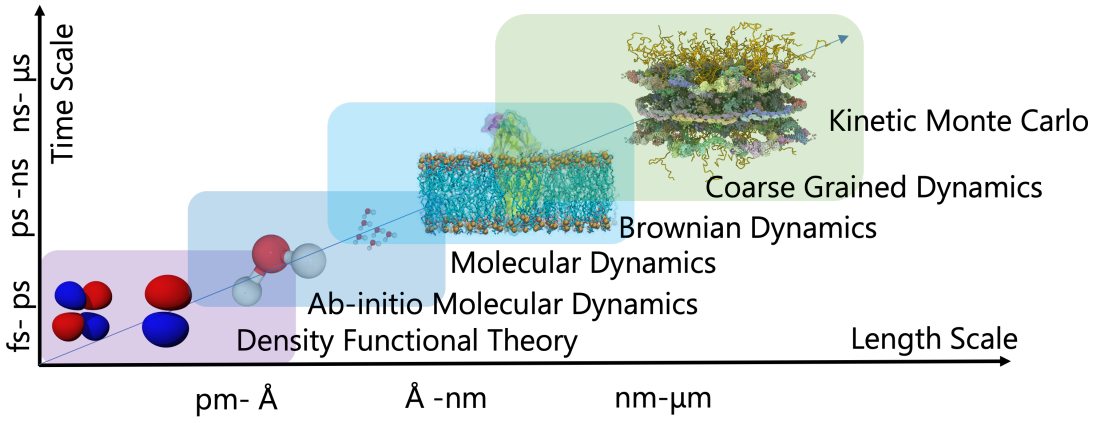


Figure 1.4: **Multiscale computational modeling of proteins**: The figure shows several computational models and the length and time scales they model at comparable computational costs.

often used to refine 3-dimensional structures of proteins and other biomolecules based on empirical conditions from X-ray crystallography[88–91] or NMR spectroscopy[92–95]. In addition, the method frequently analyzes the motions of macromolecules such as proteins and nucleic acids, which help interpret the results of biophysical experiments and model the interactions with other molecules, as in ligand docking.[205]

One chooses between explicit and implicit solvents for simulating molecules in a solvent.[208–210] The force field must calculate explicit solvent particles (such as the TIP3P,[208,211] SPC/E,[212,213] and SPC-f[214] water models) expensively, while implicit solvents[215–217] use a mean-field approach. Employing an explicit solvent is computationally expensive, requiring the inclusion of roughly ten times more particles in the simulation. However, the granularity and viscosity of explicit solvent are essential to reproduce many properties of the solute biomolecules.[208] In all molecular dynamics simulations, the simulation box size must be large enough to avoid boundary condition artifacts. One often treats boundary conditions by choosing fixed values at the edges (which may cause artifacts) or by employing periodic boundary conditions in which one side of the simulation loops back to the opposite side, mimicking a bulk

phase (which may also cause artifacts of a different type).[218]

The design of a molecular dynamics simulation considers the available computational power. One selects simulation size (n = number of particles), timestep, and total time duration so the calculation can finish within a reasonable period. However, the simulations must be long enough to be relevant to the time scales of the natural processes under study. The simulated time should match the natural process's kinetics to make statistically valid conclusions from the simulations. Most literature about the dynamics of proteins and DNA/RNA use data from simulations spanning nanoseconds ($10^{-9}s$) to microseconds ($10^{-6}s$). Several CPU days to CPU years are needed to obtain such simulation data. Parallel algorithms distribute the computational load among CPUs. With the advent of powerful parallel computing technologies like the Graphical Processing Units (GPUs), several organizations that maintain Molecular Dynamics codebases released several newer versions[219–222] that are GPU compatible. Although GPUs massively accelerate MD simulations, they still take several days to produce microseconds of data. Such time constraints created a need for enhanced sampling methods such as Replica-Exchange methods[223] and Metadynamics.[224,225]

Other limitations of MD are related to the parameters used in the underlying molecular-mechanics-based force fields.[206,207] For example, many MD simulation optimizes the protein's potential energy rather than the free energy, indicating that they may neglect all the entropic contributions to the thermodynamic stability of proteins' structures, including the conformational entropy of the polypeptide chains and hydrophobic effects.[226–229] Another critical factor is intramolecular hydrogen bonds, which are not explicit in modern force fields.[204,230–235] They exist as Coulomb interactions of atomic point charges, which is a crude approximation. Hydrogen bonds have a strong quantum mechanical nature.[236–238]

Further, MD algorithms calculate electrostatic interactions using the permittivity of free space,[239] although the surrounding aqueous solution has a much higher permittivity. Moreover, using the macroscopic permittivity (as empirically measured) at

short interatomic distances is dubious.[240,241] Finally, MD algorithms describe Van-der-Waals (VdW) interactions by Lennard-Jones potentials[242] based on a model only applicable in a vacuum. However, from a classical point of view, all types of VdW forces are of electrostatic origin and therefore depend on the dielectric properties of the solvent environment.[243] Standard MD simulations neglect the environment-dependence of VdW forces, leading to the development of several novel polarizable force fields, such as the interfacial force field.[244–249]

Computational Quantum Mechanical Models

Classical molecular dynamics usually only represents the ground state of the potential energy surface in the force field. First principles can provide electronic behavior for excited states and chemical reactions involving the making and breaking of bonds using a quantum mechanical method, such as density functional theory,[250–252] when a more accurate representation is needed. Density-functional theory (DFT)[251] is a quantum mechanical computational method that investigates the electronic or nucleonic structure (primarily the ground state) of many-body systems in specific systems. In DFT, functionals (functions of mathematical functions) of the spatially dependent electron density determine the properties of a many-electron system. DFT is the most popular method in condensed-matter physics, computational physics, and computational chemistry. Despite new improvements, there are still several difficulties[252,253] plaguing the use of DFT to describe adequately, (a) Intermolecular interactions, especially Van-der-Waals forces; (b) Excitations during charge transfer; (c) Strongly correlated systems, transition states, global potential energy surfaces, and point-defect interactions; (d) Calculations of the band gap; and (e) Calculation of ferromagnetism in semiconductors. The poor treatment of Van-der-Waals forces adversely impacts the precision of DFT in studying systems dominated by weak interactions or where weak interactions compete significantly with other effects, e.g., in biomolecules.[253] There-

fore, developing new DFT techniques designed to overcome the problem by alterations to the functional or by including additive terms is an active research area.[254]
However, despite the current popularity of new DFT techniques, they start straying away from the search for the exact functional. DFT potentials obtained with modifiable parameters are no longer 'true' DFT potentials since they are not functional derivatives of the exchange-correlation energy.[255] Moreover, even though quantum mechanical methods are incredibly accurate and robust, they are extremely computationally expensive. A method developed to use DFT with MD to reduce computational cost, is named Ab Initio Molecular Dynamics (AIMD).[232] Due to the computational cost of studying the electronic degrees of freedom, the computational load of such simulations is far loftier than classical molecular dynamics. For this reason, AIMD is typically restricted to much smaller systems and much shorter times, limiting their applicability for protein folding problems in the ergodic timescale.[230,256,257] However, AIMD methods can calculate the potential energy of a system on the go, as needed for conformations in an MD trajectory. This calculation is limited to the close neighborhood of the reaction coordinate. Although various approximations exist, these derive from theoretical concerns, not empirical fitting. AIMD calculations produce massive information not available from empirical methods, such as the density of electronic states. A significant benefit of using AIMD methods is the capacity to study reactions involving breaking or forming covalent bonds, which correspond to multiple electronic states.[258]
Moreover, AIMD methods also allow recovering effects beyond the Born–Oppenheimer approximation[259] using strategies like mixed quantum-classical dynamics.[260–262] Such methods are called mixed or hybrid quantum mechanical and molecular mechanics (hybrid QM/MM).[263–266] The most significant advantage of the hybrid QM/MM approach is the speed. The computational cost of performing classical MD in the most straightforward case scales $O(n^2)$, where n is an integer denoting the number of atoms in the system. The cost is mainly due to calculating the numerous electrostatic in-

teractions (every particle interacts with every other particle). Nevertheless, the use of cutoff radius, periodic pair-list updates, and the variations of the particle-mesh Ewald's (PME)[133,267] method have reduced the cost between $O(n)$ to $O(n^2)$. On the other hand, the most straightforward AIMD calculations typically scale $O(n^3)$ or worse.[268–270] Therefore, only a tiny fraction of the system (typically the active site of an enzyme) uses quantum-mechanical calculations, and the remaining system undergoes classical MD treatments. More sophisticated implementations of hybrid QM/MM methods exist that treat light nuclei susceptible to quantum effects (such as hydrogen nuclei) and electronic states quantum mechanically.[271–273] Such methods allow generating hydrogen (proton) wavefunctions (similar to electronic wavefunctions). Proton wavefunctions help investigate phenomena such as hydrogen tunneling.[272] One such example is the calculation of hydride transfer in the enzyme liver alcohol dehydrogenase. In this case, quantum tunneling is essential for the proton, as it determines the reaction rate of alcohol metabolism.[274]

1.1.5 Data-driven Methods for Protein Structure Prediction & de-novo Protein Design

Other than explicit *ab-initio* methods, there are several other extant models of determining a protein's secondary and tertiary structures from its sequence. Most of such methods use some sort of homology modeling,[275–279] bioinformatics approaches,[280,281] and stochastic methods such as Monte-Carlo Sampling[282–284] and machine learning.[285] Making use of such methods has also led to advances in the field of *de-novo* protein design[286–289] where given a required structure, the algorithm generates best fit sequences, optimized according to the least free energy rule of equilibrium and stability.

One such method is the Rosetta algorithm,[290] which uses protein structures from the Protein Data Bank[96,291] repository to generate a fragment library with different probabilities assigned to each fragment and structure. The algorithm then randomly samples different fragments and applies the structure to the protein sequence of in-

terest at different positions along the backbone. After that, Rosetta minimizes the energy at each location and retains the lowest energy pairing of location and fragment for further statistical analysis. Protein design[286,288,289,292] is the opposite of such a method, where the algorithm randomly samples protein sequences for which a given structure exists in the lowest energy well in the energy landscape. The experimental data stored in the PDB, drives both Rosetta and the protein design algorithms. Such data-driven models (fig 1.5) constitute one type of bioinformatics analysis called homology modeling.[275–279]

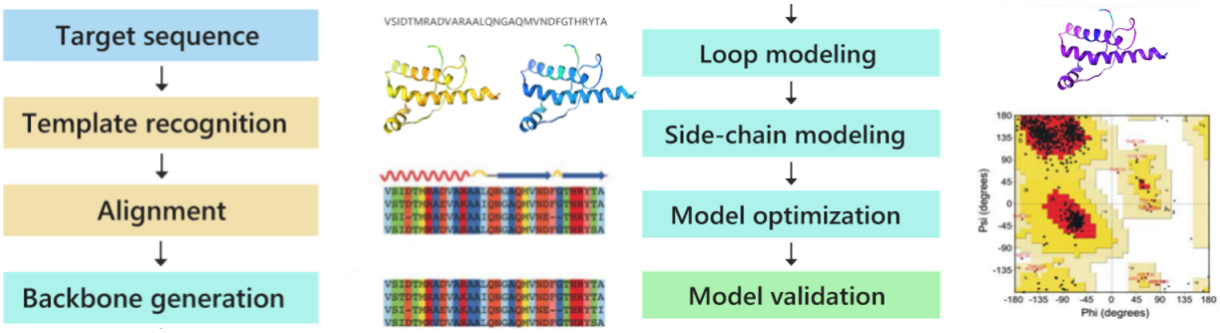


Figure 1.5: **A typical data-driven structure prediction pipeline**: The figure shows a typical bioinformatics procedure used extensively to study protein structures. Figure Source: *https://www.profacgen.com/homology-modeling.htm*

Bioinformatics

Bioinformatics is a multidisciplinary field that invents techniques and software to understand biological data, mainly when the data sets are massive and complex.[280,281] Bioinformatics performs *in silico* studies of biological questions using computational and statistical approaches.[278–281] Typical bioinformatics use-cases include identifying candidate genes and single nucleotide polymorphisms (SNPs). Often, such identification enables understanding the genetic cause of diseases, desirable properties, unique

adaptations, or disparities between different demography groups. In genetics, it assists in the sequencing and annotating of genomes and their observed mutations. It plays a role in developing biological and gene ontologies to manage and query biological data. It also helps study gene expression, regulation, and protein translation. Bioinformatics techniques also aid in comparing, analyzing, and interpreting genomic and proteomic data.[293–295] More generally, they aid in the understanding of evolutionary aspects of molecular biology. At a more molecular level, it enables us to analyze and catalog the signaling pathways and networks that are essential to systems biology. In structural biology, bioinformatics aids in the simulation and modeling of DNA, RNA, proteins, and biomolecular interactions in the proteome.[295] The proteome is the full and complete set of proteins produced or altered by an organism or system. Proteomics facilitates the identification of ever-increasing numbers of proteins due to improvements in sequencing technologies enabling deeper exploration of the sequence space.[296,297] Deep mutational scanning[298–300] is one such empirical method taking advantage of newer sequencing technology such as Next Generation Sequencing (NGS)[301–303] combined with huge libraries of protein mutations and biological variants to explore the phenotypical effects of genome level mutations in a multiplexed manner in the laboratory.

In molecular biology, Combinatorial mutagenesis[304] is a vital laboratory procedure where the researcher deliberately mutates the DNA of a model organism such as *Enterobacteria Phage M13*,[305–307] *Escherichia Coli*,[308] or *Saccharomyces Cerevisiae*[309–311] to produce libraries of mutant genes, proteins, strains, or other genetically modified organisms. The researchers mutate numerous components of a gene, its regulatory elements, and its gene products to examine the functioning of a genetic locus, process, or product in detail. The mutation produces mutant proteins with exciting properties or improved or unexplored functions that have the potential to be commercially useful. Mutant strains may also show up that have a practical applications or allow the investigation of the molecular mechanisms underlying a particular cell function.

The ML/AI algorithms use massive amounts of sequence and functionality data to predict new protein sequences, model evolutionary dynamics, and physics or predict other related functions, a type of metric learning.[312]

Directed evolution (DE)[313–315] is a technique used in protein engineering that mimics the process of natural selection in a laboratory setting using gene libraries to guide proteins or nucleic acids toward a user-defined purpose, as shown in fig 1.6. It consists of subjecting a gene to iterations of mutagenesis, selection (expressing the mutant variants and separating proteins with the desired function), and amplification (breeding a template for the next round with a new generation of organisms). It is performed *in vivo* (in living organisms) and *in vitro* (inside freely suspended cells outside a living organism or in solution).[316,317] Directed evolution is used in protein engineering as a recourse to rationally concocting altered proteins and for experimental studies of fundamental evolutionary tenets in a controlled, laboratory setting.[313]

Protein structure prediction is an important application of bioinformatics. The protein's amino acid sequence, the primary structure, can be easily translated from the gene sequence that codes for it. In most cases, this primary structure uniquely defines a 3D structure in the protein's native environment. Structural bioinformatics[318] is the branch of bioinformatics associated with analyzing and predicting the three-dimensional structure of biomacromolecules such as proteins, RNA, and DNA. It deals with abstractions and inferences about macromolecular 3D structures, such as comparisons of prevalent folds and local motifs, tenets of molecular folding, evolution, binding affinities, and structure/function relationships, operating from experimentally obtained structures and computational calculations. The primary purpose of structural bioinformatics is the design of new methods of analyzing and exploiting biological macromolecular data to solve problems in biology and generate new knowledge-base. Comparative modeling, a type of structural bioinformatics known as homology modeling,[275–279] corresponds to constructing 3-dimensional structures

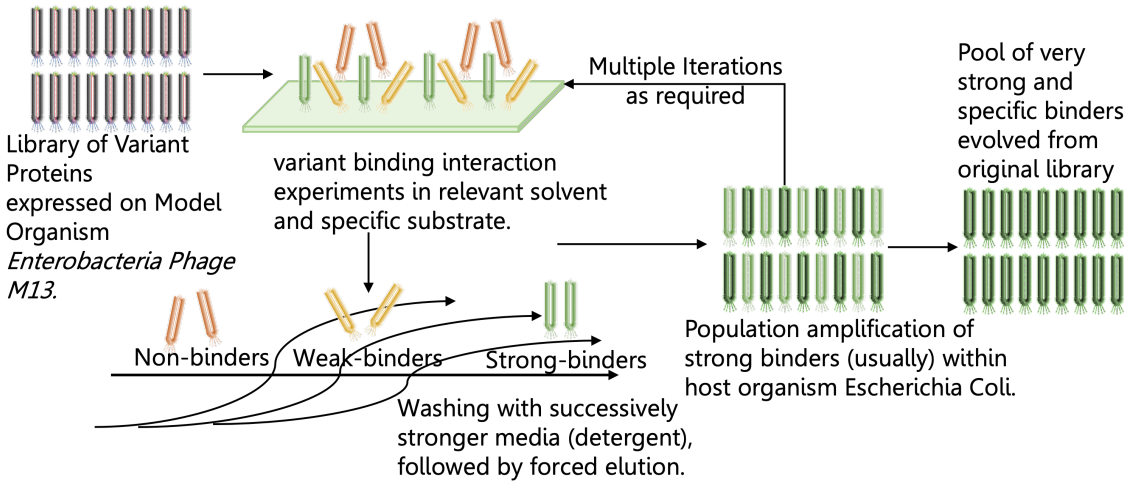


Figure 1.6: **A typical Directed Evolution through Combinatorial Mutagensis approach**: The figure shows a typical high throughput empirical procedure used extensively to study protein sequences and the effects of mutations on their functions. Deep Mutational Scanning offers a hitherto unseen glimpse into the phenotypical effects of point mutations by mutating the the genomes of model organisms such as *Enterobacteria Phage M13*.In many experiments, the model organisms have been modified to express human proteins. Such transgenic microorganisms enable deep scanning of phenotypical effects of mutations in human essential proteins such as those involved in the coagulation of blood, Alzheimer's disease, Anemia, and Diabetes.

from the sequence of a target protein and a template protein with a previously known structure. The literature agrees that evolutionarily related proteins present a conserved 3-dimensional structure.[319] In addition, sequences of distantly related proteins with sequence matching/alignment agreement of lower than 20% can present different folds. Phylogenetically and cladistically related proteins have similar sequences, and naturally occurring homologous proteins have similar 3D protein structures.[320,321] 3-dimensional protein structure is evolutionarily more consistently conserved than is theoretically expected based on sequence conservation alone.[321] The structural-homology model's quality is conditional on the sequence alignment[322–324] and quality of the template structure. The alignment gaps (commonly called insertion-deletions or indels)[325,326] that indicate a structural region in the target but not in the template

complicate the homology modeling process.[327] Structure gaps in the template arising from poor resolution in the experimental procedure (usually X-ray crystallography) make homology modeling suboptimal.[328]

Model quality diminishes with declining sequence identity from sequence alignment algorithms; a typical model has $\sim$ 1–2 Å root mean square deviation (RMSD) between the matched C_α atoms at 70% sequence identity but only $\sim$ 2–4 Å agreement at 25% sequence identity.[322,323] However, the errors are significantly more prominent in the random coil regions, where the target and template proteins' amino acid sequences may be completely different. In addition, parts of the model created without a template, usually by loop modeling, are typically much less accurate than the rest of the model.[323] Errors in the sidechain packing and position also grow with reducing sequence identity, and deviations in these packing configurations constitute a significant reason for poor model quality at low levels of sequence identity.[322] Such atomic-position errors are considerable and inhibit the application of homology models for research objectives that mandate atomic-resolution data, such as drug discovery and design and protein-protein interaction predictions; even the quaternary structure of proteins may be difficult to predict from homology models of their constituent subunits. The principal inaccuracies in homology modeling, which deteriorate even further with lower sequence identity, emanate from errors in the initial sequence alignment and unsuitable template selection.

Nonetheless, homology models help reach qualitative inferences about the biochemistry of the sequence under study, especially in developing hypotheses about why specific residues are conserved, which may, in turn, lead to empirical techniques to study those hypotheses. For example, the spatial arrangement of conserved residues may suggest whether a particular residue is conserved to stabilize the folding, to participate in binding some small molecule, or to foster association with another protein or nucleic acid.[329,330]

Informatics techniques used in structural bioinformatics are (a) Selection of Potential targets by comparing them with repositories of known structures and sequences. Targets are also selected based on their protein domain. Protein domains are building blocks that can rearrange to form new proteins; (b) Tracking X-ray crystallography trials- X-ray crystallography is used to probe the 3-dimensional structure of proteins. However, pure protein crystals must form to use X-ray to study those protein crystals, which can take many trials. Too many trials lead to a need for tracking the circumstances, processing conditions, and results of the trials. Similarly, supervised machine learning algorithms used on the stored data can identify conditions that might increase the yield of pure crystals; (c) Study of X-Ray crystallographic data- The diffraction patterns acquired from shooting X-rays onto electrons is the Fourier transform of the electron density distribution. Bioinformatics-based extrapolation methods can generate an electron density map, which uses the location of selenium atoms as a reference to determine the rest of the structure; (d) Study of NMR spectroscopy[331,332] data - Nuclear magnetic resonance spectroscopy experiments yield high dimensional data, with each peak in the spectrum corresponding to a chemical group within the sample. Optimization methods convert the spectra into 3-dimensional structures; and (e) Correlating Structural data with functional details- Structural studies can be used as a probe for structure-function relationship prediction.

Machine Learning

Besides bioinformatics, several recently developed Machine Learning (ML)[333,334] and Artificial Intelligence (AI)[335,336] techniques exist to predict the secondary and tertiary protein structures. Artificial Intelligence methods also exist that use bioinformatics principles such as sequence alignment and homology modeling to predict structure-function relationships and proteins that interact with each other. Several supervised and unsupervised ML and AI algorithms make use of massive datasets

of protein structure (PCDDB,[337], PDB, UniProt[338–340], EMBL-EBI[341,342]: European Molecular Biology Laboratory's European Bioinformatics Institute's databases), sequences in the NCBI databases,[343–346] interaction and protein expression data in the literature,[343] and other independent datasets such as the Immuno Epitope Database (IEDB).[347,348] The usual problems[324,349] include (a) de-novo prediction of secondary or tertiary structure in terms of coordinates in a pre-defined euclidean space that agrees with experimentally observed X-Ray Crystallography or NMR data; (b) Predicting sequences that bind to a particular target; (c) Prediction of binding affinities or some other quantitative metric related to protein function; (d) understand the governing physics of protein structure attainment and related biological function; and (e) quantify and model the evolutionary dynamics to design proteins with desirable functions for medical, pharmaceutical, and technological use cases. AI models for studying such problems are an active area of research that includes many examples such as DeepMHC,[350] DeepSeqPan,[351] AlphaFold,[352] AlphaFold2,[353–355] and NetMHC[356–358]. One of the significant constraints of ML and AI models is their dependence on high-quality and high-fidelity data with minimal noise. Experimentally obtained structures in the PDB database must be parsed, cleaned, and reformatted to be usable by such algorithms. Moreover, adopting new technologies such as combinatorial mutagenesis, directed evolution, and deep mutational scans that leverage high throughput methods such as next-generation sequencing to deliver billions of sequence data within a few days increased the available data to build better and more robust ML and AI models. However, the protein sequence space is vast and virtually limitless, whereas the structural experiments are cumbersome and cannot keep up with the sheer rate of sequence discovery.

Several information-theoretic approaches help untangle the output sequence data from such massively parallelized high throughput experimental setups and help clean, label, and quantify the data correctly for use with ML/AI approaches.[336,359–362] How-

ever, despite massive data sources, robust data-driven models, and extensive ongoing research in the field, there is still no widespread consensus on the best, most physically relevant representation of proteins and other molecules.[363] AI models are still primarily limited to treating sequences as binary matrices denoting the presence or absence of an amino acid or nucleic acid at a particular location in the sequence[364]; or letters, where each letter represents a type of nucleic acid or amino acid. AI models largely ignore several molecules such as peptidoglycans, peptidolipids, peptoids, and many other physiologically relevant biomolecules because of the inconsistency of representing such molecules as computer-readable inputs. Several attempts at fingerprinting[363,365–369] exist in the literature, but such abstractions lead to erroneous or uninterpretable correlations while trying to decipher the decisions made by the AI models. Such difficulties complicate the gleaning of any foundational physical principles of biomolecular structure adoption and associated functions. Molecular recognition and Signal Transduction are even more complicated problems than simply predicting a reasonably acceptable primary or secondary structure and some high-level function. The motivation to understand the fundamental basis of specificity during molecular recognition and precision during the signal transduction cascade along the entire signaling pathway necessitates using first principles in conjunction with ML/AI models to design physically relevant statistical models, called Physics-Informed models[78,370] that we describe in the following paragraphs.

Physics-Informed Data-driven Models

Physics-informed Models[371,372] are universal mathematical function estimators that can derive the information about all the dominant physical laws that govern a phenomenon described by a given dataset during the statistical learning process. the algorithms describe the phenomena in terms of partial differential equations (PDEs).[373,374] They overcome the obstacles of low and noisy data availability for some engineering or biological systems that make most advanced ML methods lack robustness, rendering

them inadequate to model such systems. The preliminary knowledge of known physical laws is depicted as regularization terms in training ML models and constrains the permissible solutions' space.[373,375] Such constraints increase the precision of the function estimation. This way, including this preliminary information into the model, improves the information content of the available data, enabling the algorithm to apprehend the right solution and generalize well even with a few training examples. Partial differential equations depict most physical laws that govern the dynamics of a system.[375] However, such equations cannot be solved precisely and analytically in > 1 dimension (which makes them an Ordinary Differential Equation), and consequently, numerical strategies must be used (such as finite differences,[376] finite elements,[377–379] and finite volumes that discretize the space[380]). These governing PDEs must be solved in such a setting while including prior assumptions (initial and boundary conditions), linearization, and sufficient time and space discretization.[379] Solving the underlying PDEs of physical phenomena employing deep learning has lately arisen as a new domain of scientific machine learning, taking advantage of the universal feature approximation and heightened expressivity of neural networks.[381,382] Deep neural networks could potentially resemble any high-dimensional function, provided that sufficient training data of good quality exists.[383] However, such networks do not necessarily consider the physical aspects underlying a real-world phenomenon, and the level of estimation accuracy supplied by the networks is still extremely conditional on careful specifications of the problem geometry and the initial and boundary conditions. Without this preliminary knowledge, the solution is not unique and loses physical interpretability[384] or agreement. Conversely, physics-informed neural networks (PINNs) leverage underlying physical equations of the system to be modeled in neural network training and hyperparameter tuning.[385–390] The training procedure design of PINNs aims to fit not only the provided training data but also the imposed governing equations. In this fashion, physical knowledge guides a neural network to model training data that do not necessarily need to be massive and complete. PINNs

can potentially find an accurate solution to partial differential equations without knowing the boundary conditions, purely from data.[371] Therefore, with some understanding of the physical aspects of the situation and some form of training data (even sparse and incomplete or noisy), PINNs can find an optimal resolution with high fidelity and robustness. PINNs enable addressing an expansive range of phenomena in computational science and are a pioneering methodology for developing new classes of autonomous numerical solvers for PDEs. Notably, the trained PINNs can predict the values on simulation meshes (like that in Finite Element Methods) of different resolutions without the need to be retrained, as interpolation is built-in.[391] Additionally, they allow for exploiting automatic differentiation (AD)[392] to compute the required gradients in the fields described by PDEs, a new category of differentiation methods widely used to derive optimal connectivity and weight matrices in deep neural networks and assessed to be superior to numerical and symbolic derivatives.[393–395]

1.1.6 Random Matrix Theory

Considering that (a)Physics-informed neural networks (PINNs), as described above, still use weight matrices with randomly populated weights and (b) Random Matrix Theory (RMT)[396,397] is a well-known branch of applied mathematics that models nucleonic and fermionic energy states and quantum chaos[398,399]; it is an admissible presumption that RMT has a hitherto unknown but critical role to play in the PINN framework.[400]

A random matrix is a random variable that takes matrix values, i.e., a matrix in which all elements are random variables. Several crucial physical phenomena are partial differential equations construed as matrix and eigenvalue problems. For example, Eugene Wigner introduced the field of random mean-field matrices to study nuclear energy states and theoretically probe the nuclei of heavy atoms.[401] **Wigner posited that the spacings between the energy spectra obtained from a heavy nu-**

cleus with many fermions resemble the trends in eigenvalue spacings of a random matrix. Furthermore, Wigner claimed that the nuclear spectral spacings are independent of the individual random variables but depend only on the symmetry class of the underlying time evolution. Wigner also postulated that the probability that eigenvalues of such a large random matrix coincide is vanishingly small.[402] This postulate came to be known as Wigner's surmise. Wigner's postulates were since proven accurate, and mean-field random matrices belonging to Gaussian ensemble classes (see fig 1.7), that follow Wigner's surmise now ubiquitously model the behavior of large and disordered Hamiltonians ($\hat{\mathbf{H}}$, The matrix valued energy operator in the time-dependent Schrödinger equation, and the more general Dirac Equation formalism of Quantum Mechanics of many body systems).

Similarly, in quantum chaos,[399,403,404] where researchers attempt to model classical Brownian systems as emergent phenomena from first principles, the Bohigas-Giannoni-Schmit (BGS) conjecture[405] posits that RMT adequately models the behavior of the energy value spacings of quantum systems that show chaotic behavior in the classical realm. RMT has applications to the chiral Dirac operator in QCD (Quantum ChromoDynamics, a study of quarks),[406,407] 2-dimensional quantum gravity hypotheses (an attempt to unify general and special relativity),[408] physics in the mesoscopic scale (a scale where biomolecules operate),[409] torque during spin transfer,[410,411] the fractional quantum hall effect,[412], high-temperature superconductors (necessary for lossless energy transfer and circuitry)[413] and quantum dots (necessary in quantum computing)[414,415]. RMT has found fame in numerical analysis to describe errors in computation during matrix multiplication operations[416,417] via logic gates, multivariate statistics in estimating covariance matrices,[418–421] optimal control theory,[246] and theoretical neuroscience[422–424] where random matrices represent the unknown synaptic weights in a connectome.

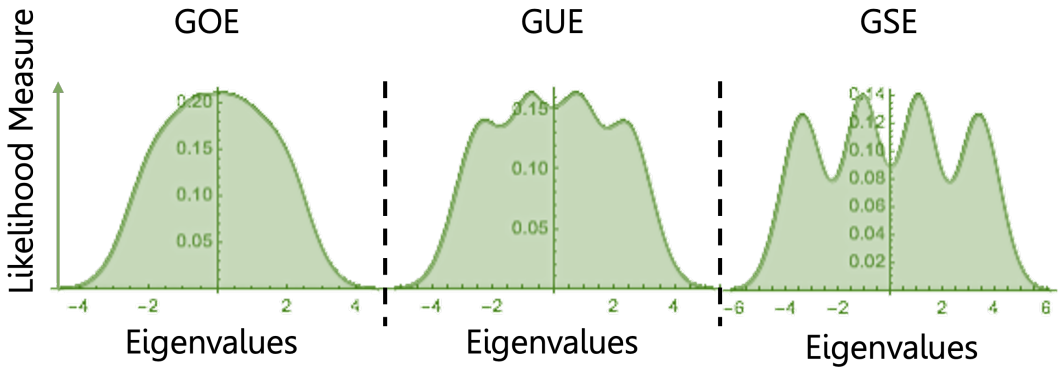


Figure 1.7: **Gaussian Ensemble symmetry classes for single point correlation distribution of eigenvalues of random matrices**: The figure shows the three types of symmetry classes under the Gaussian Ensembles. The first, Gaussian Orthogonal Ensemble (GOE) models Hamiltonians in the many-body Schrödinger equation where the particles have both rotational and time reversal symmetry. The entries of the matrix are real valued and the GOE works in Euclidean space; The Gaussian Unitary Ensemble (GUE) models Hamiltonians in the many-body Schrödinger equation where the particles do not have time reversal symmetry. The entries of the matrix are complex values of the form $a \pm ib$ and the GUE works in multidimensional Riemann space; The Gaussian Symplectic Ensemble (GSE) models Hamiltonians in the many-body Schrödinger equation where the particles have time reversal symmetry but no rotational symmetry. The entries of the matrix are quaternions of the form $t + x_1\hat{i} + x_2\hat{j} + x_3\hat{k}$. The GSE works in Minkowski spacetime (one independent time dimension and three spatial dimensions).

1.2 Assumptions and Motivation

The main interest in RMT for the current body of work is that RMT models the energy Hamiltonians $\hat{\mathbf{H}}$ in the time-dependent Schrödinger equation for many interacting particles given in equation(8) as:

$$\hat{\mathbf{H}}\psi(\mathbf{r},t) = \mathbb{E}\psi(\mathbf{r},t) \tag{8}$$

$$\hat{\mathbf{H}} = -\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{r},t) \tag{1.1}$$

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

where $V(\mathbf{r}, t)$ is the potential energy operator, and $\mathbf{r}$ is an arbitrary vector in space with dimensionality ≤ 3. ∇^2 changes its dimensions accordingly. Density Functional Theory (DFT) attempts to precisely estimate the Hamiltonian $\hat{\mathbf{H}}$ and its eigenvalues (allowed single-particle energy levels) $\mathbb{E}$ and eigenvectors (wavefunctions) $\psi(\mathbf{r}, t)$ for many body systems (it is only exactly solvable for the Hydrogen atom that consists of just one electron around one proton) by rationally assuming a periodic potential functional $V(\mathbf{r}, t)$. The obtained single-particle energy states (Eigenvalues of the Hamiltonian $\hat{\mathbf{H}}$) are then used in the Fermi-Dirac equation (4) to obtain the number of particles in each single-particle energy state at the given temperature T.

RMT comes into play by overlooking the search for the true potential functional $V(\mathbf{r}, t)$ and instead modeling the entire Hamiltonian $\hat{\mathbf{H}}$ as a random matrix. in DFT, the particles have to be non-interacting, or the exchange correlation functional has to be assumed differently as well. But in RMT, non-interacting particles v. interacting particles is simply a case of sampling elements of the random matrix with i.i.d (identically and independently distributed) normal $\mathcal{N}(0, 1)$ distribution vs. sampling each random matrix such that each of its elements are formed from dependent distributions. Wigner used a formulation where he sampled all elements of a real valued Hermitian matrix $H : H_{i,j} \sim \mathcal{N}(0, 1)$ $i.i.d$, to model the statistical behavior of nucleonic (particles such as protons and neutrons constituting the nucleus of the atoms with overall half-integer spins) energy spectra. However, a similar formulation also works for other fermions (particles such as electrons with half-integer spins that take part in the chemical behavior of atoms). Random Band Matrices (RBMs),[425] with non-zero entries around the leading diagonal and all other entries zero, have recently been used to study electron interactions in systems where valence electrons are either localized or progressively delocalize, and they found that a sharp transition occurs. after a certain point, all eigenvectors of such matrices are fully delocalized (quantum unique ergodicity)[426] and the eigenvalue distributions belong to the Gaussian Orthogonal Ensemble class (described further in chapter 3)[427]. The RBM study

makes a connection that when electrons are delocalized, or interact over large distances, their eigenvalue statistics follow the semicircle rule propounded by Wigner, while localized and independent electron wavefunctions result in a Poisson distribution of the eigenvalues of the RBM.[427] **The current body of work attempts to leverage such formalism of RBM in quantum mechanical interactions to model the electron dynamics during protein folding** because protein folding leads to conductivity changes, to inquire whether quantum mechanical phenomena can steer data-driven artificial intelligence models (a novel type of PINN) to predict protein secondary structures, protein complexes, and biomolecular interactions. The study also attempts to probe the mysteries behind the specificity of molecular recognition and the flow of information during error-free signal transduction cascades along the signaling pathway.

1.2.1 Quantum Mechanics in Protein Function

From the discussion so far, it is clear that the significant chasm between current state-of-the-art computational modeling algorithms (explicitly and implicitly physics-driven, data-driven, and physics-inspired hybrid algorithms) and the fundamental understanding of molecular recognition and signal transduction is closing at an inadequate pace because of the following:

(a) Incomplete understanding of underlying physical mechanisms,

(b) Sparse and noisy protein structure data,

(c) Unavailability of dynamic protein interaction data,

(d) The impossibility of modeling protein complexes and non-protein, non-nucleotide biomolecules simultaneously, and

(e) Intrinsic disorder in active functional proteins making the job more challenging.

Quantum mechanics, although ubiquitously at work everywhere, is not a significant focus area for researchers in molecular biology and bioinformatics. They assume that, as in the case of many relatively hot, wet, and noisy systems, quantum mechanical phenomena transition into classical dynamics. Decoherence[428] is the emergent classical behavior from underlying quantum mechanical behavior where wavefunctions collapse and subatomic particles behave more like real particles than smeared-out probability density waves in spacetime.

However, scientists discovered many examples in recent years (fig 1.8) where biological phenomena depend on the quantum mechanical behavior of electrons and nuclei and cannot be described by classical physics.[429–431] Some essential questions that the nascent field of quantum biology aims to answer are:

(a) Quantum mechanical treatment of certain biological phenomena is well-known, but how is coherence maintained in the cells?

(b) Are quantum mechanical explanations the only explanations that can fully describe said phenomena?

(c) Did such quantum mechanical phenomena proffer evolutionary advantage to organisms?

(d) Can quantum mechanics explain the origin and functions of early molecular life before the first prokaryotic cells? and ultimately,

(e) Is quantum mechanics only trivially involved in biological processes, or is there a preference for maintaining quantum coherence in specific processes that rely on the quantum mechanical nature of particles?

Several recent works in the literature have shown that coherence could be conserved even at high temperatures and aqueous gel-like environments due to structured water

around all biomolecules and the thermal noise interacting positively with the quantum waves to reinforce rather than collapse the waves.[432] Some others posit that because a chance happening at some point during evolution, a protein so folded that it structured the water around it to act as a filter for noise, conferring the ability of the protein to use quantum mechanical behavior.[433] Coincidentally, that protein must have rendered some phenotype advantages to increase the chances of survival and adaptation of the cell long enough to reproduce, to have been evolutionarily conserved. Either way, whether water acts as a filter for all proteins, or if only some proteins evolved to make use of water and their own 3D structures in that manner, especially because we still do not know the purpose of low frequency oscillations in both α-helices and β-structures, it would seem that biological processes occur somewhere at the edge where quantum phenomena transition into classical dynamics.

The literature now accepts that quantum mechanical phenomena drive photosynthesis.[434] The studies indicate that organisms have evolved to develop ways to protect quantum coherence[435] that enhances photosynthetic efficiency, which has a clear evolutionary advantage. Single-molecule spectroscopy now depicts the quantum aspects of photosynthesis without the interference of static noise,[436,437] and some studies use this approach to designate reported signatures of electronic quantum-coherence to nuclear dynamics[438] occurring in chromophores. However, analyses investigating transport dynamics suggest that interactions among electronic and vibrational excitation modes in photosynthetic protein complexes demand a semi-classical, semi-quantum rationale for exciton energy transfer.[439] While quantum-coherence dominates the short-term dynamics, a classical description accurately describes the excitons' long-term demeanor.

Similarly, proton tunneling is the mechanism behind alcohol metabolism in the liver involving alcohol dehydrogenase. Proton tunneling along the hydrogen bonds in the DNA double helix (tautomerization)[440,441], changes the complementarity of nucleotides. It is also among many proposed methods by which mutations occur in the

DNA double helix strands. Olfaction is another poorly understood biological phenomenon.[442,443] The vibration theory of olfaction[443] posits that particular molecules that vibrate with the correct frequencies trigger specific receptors that detect the vibration and send messages to the olfaction center of the organ that controls biological processes, usually the brain. The assumption is that the resonance between molecular vibrations of sensing-protein receptors and odor-molecule vibrations creates a virtual bridge for electrons to pass through and trigger a signaling cascade. Experimental results *in vivo* are mixed.[444]

Quantum mechanical aspects of magneto-reception[439,445,446] in some migratory birds and long-range cooperation between enzymes are undeniable. Empirical and theoretical studies show that cryptochrome proteins in the retina of migratory birds (especially birds that do not have magnetite crystals in their beaks), such as the European robin, are replete with the amino acid Tryptophan.[447–449] Breaking a Tryptophan residue into the indole ring and free hydrogen is not impossible in a watery environment. Indeed, because the radicals now exist in a spin triplet state (two electrons that used to connect, if separated, do not need to follow Pauli's exclusion principle and therefore exist in a superposition of states where the sum of their half-integer spins equals either 1, 0 or -1). Such a spin triplet state is susceptible to slight changes in the angle of the earth's magnetic field, as it impacts the spin, which is an angular momentum like quantity.[448,450] Any disruption of the spin-triplet state due to changing magnetic field angles triggers vision receptors in the retina, enabling the birds to see the angular change in magnetic field lines in the sky.[451] Such susceptibility to perceiving magnetic field lines guides their annual migration to and from the equator (earth's magnetic field lines are perpendicular to the surface at the poles and parallel at the equator) during winter and spring, respectively.

Meanwhile, in the case of enzyme cooperativity over long distances,[453–455] while synthesizing biomolecules and metabolites, low barrier hydrogen bonds[456] enable the

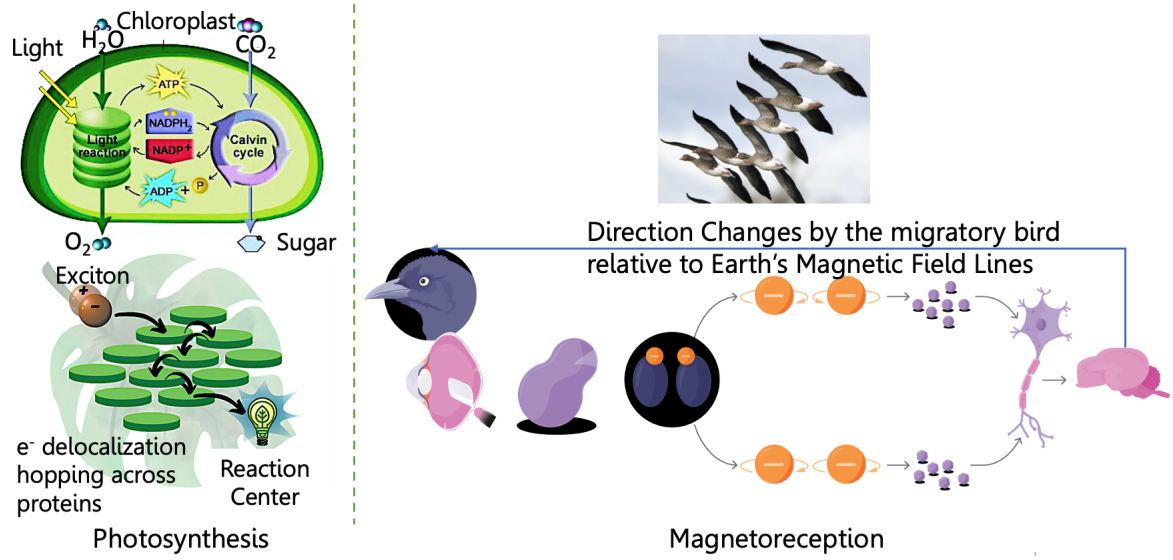


Figure 1.8: **Quantum Mechanics driven phenomena in biological processes**: The figure shows some of the most well-known biological processes that depend on the quantum mechanical nature of electrons and protons: Photosynthesis and Magnetoreception.[136,442] The former leverages electron delocalization, while the latter leverages superposition of spin states, called a spin-triplet-state. Whether such quantum mechanical processes can exist at temperatures and levels of thermal noise inside living cells is a matter of debate, the main hypotheses in favor posit that the water is structured in such a way around the proteins that they filter out most frequencies of thermal noise, and only let in frequencies that can enhance the quantum mechanical vibrations.[452] *Figure Sources: https://www.azoquantum.com/Article.aspx?ArticleID=281, https://www.scienceabc.com/pure-sciences/is-quantum-physics-applicable-in-biological-systems-too.html#photosynthesis*

flow of electrons through a network of water molecules between two or more enzymes or an enzyme and its target molecule (fig 1.9). A Low-barrier hydrogen bond (LBHB) is a particular type of symmetric hydrogen bond.[457] LBHBs can occur when the pKa of the two hetero-atoms closely match, allowing the hydrogen to be more equally shared. This hydrogen-sharing causes the formation of especially short, strong hydrogen bonds.[458] Such hydrogen bonds leverage the quantum mechanical nature of hydrogen bonds rather than the purely electrostatic description. Water, for example, shows a single-well hydrogen bond allowing the hydrogen electron to have an equal

probability of being at both the Oxygen atoms. It is impossible to say which water molecule the hydrogen atom belongs to or is covalent bonded. All the water inside our cells and extracellular space is known to be structured, not randomly positioned. Standard hydrogen bonds are usually 2.8Å long and are either symmetrical or asymmetrical. LBHBs are 2.5Å long and are symmetrical necessarily. Single-well hydrogen bonds are shorter than 2.5Å. Long-term stable LBHBs occur in the hydrophobic and water-excluding parts of proteins.[459,460] Short-lived LBHBs occur at the surface of proteins with water, while single-well hydrogen bonds exist among water molecules. Whether enzymes truly leverage such LBHBs for catalysis or cooperativity was a matter of much debate until, in 2019, scientists proved that enzymes cooperate over long distances within the cells by using the LBHBs and single-well hydrogen bonds as a sort of aqueous electric circuit.[461] Tautomerization by proton hopping is complimented by electrons using LBHBs to flit between water molecules, and enzymes, thereby getting entangled with other electrons along the way. The movement of electron waves (and probable resonance among multiple electron waves) changes the charge state of each water molecule they pass through, allowing for tautomerization.[461] LBHBs and single-well hydrogen bonds between water molecules are especially great for such processes because the barrier for electron-wave propagation is either low or absent, mirroring the barrier to proton delocalization along the hydrogen bond.

1.2.2 Hypotheses for quantum-mechanical phenomena during protein folding, molecular recognition and signal transduction

Despite theoretical and empirical advances in establishing quantum mechanics as a non-trivial player in some biological processes, it is unclear if they include protein folding, molecular recognition, and signal transduction. It is also unclear if quantum mechanics principles were a driving force for protein folding, specificity during molecular recognition, and error-free execution of signal transduction cascades along the

entire signaling pathway, how would they even be involved.

An early attempt in this area was in 1990 with the creation of the Resonant Recognition Model (RRM).[463] Researchers usually transform the sequence of amino acids and nucleic acids (in the case of RNA/DNA) into a series of numerical values, an array, representing some physical characteristic of the individual amino- or nucleic acids when they analyze proteins, DNA, or RNA. In the case of RRM, the numerical value represented the Electron-Ion Interaction Potential (EIIP) for each amino and nucleic acid in a given sequence, describing the valence-electrons' energy states. The EIIP values for amino acids, for instance, were calculated using the model of the pseudo-potential from DFT as given by:

$$\langle \overrightarrow{k+q} \mid w \mid \overrightarrow{k} \rangle = \frac{Z \times sin(1.04Z\pi)}{8\pi} \tag{1.2}$$

where q is the momentum change of the delocalized electron in the interaction with potential w, and: $Z = \sum Z_i/N$, where Z_i is the number of valence-electrons of the i^{th} component of each amino acid. N is the total number of atoms in that amino acid. Each amino acid was represented as its EIIP value, and the researchers converted a protein sequence into a numerical array of EIIP values. The next step in the RRM process was to apply a Discrete Fourier Transform to the EIIP series and normalize the length (the amino acids were assumed to be equidistant in the sequence, with a separation of 388 pm = 3.88$\mathbf{\AA}$) of the series and amplitudes of the pattern obtained.[464,465]

Discrete Fourier Transform coefficients of EIIP series representing two or more proteins were analyzed to get the cross-spectrum, a correlation matrix. After that, the main conclusion from the RRM model was that two proteins with similar functions resonate at particular frequencies obtained from their Discrete Fourier Transform coefficients as described above. As the RRM algorithm included more proteins of the same family from different sources in the fray, the number of resonant frequencies dwindled until one significant peak remained.[466] The RRM developers assigned that

resonant peak as the frequency signature for the particular function (e.g., Oxygen binding in the case where RRM was used to analyze many hemoglobin proteins obtained from several species).[467] They found that the frequencies also resonated for proteins that were counterparts in a molecular recognition event, and there was a phase shift of about 180 degrees at that frequency.[467] The observation led some researchers involved with the RRM study to surmise that molecular recognition might be mediated, at least up to some extent, by resonance between energy level frequencies along the sequence.[467] They also observed that the amino acids whose EIIP values contributed most to the resonant frequency component were clustered in and around the active sites of the proteins in 3D space, regardless of their proximity along the sequence. The second observation led the researchers to posit that electron energy level frequencies are somehow involved in the 3D organization of proteins and their folding into secondary and tertiary structures.[468,469]

Since the 1990s, other groups have added several modifications and updates to the RRM model, including employing other amino acid properties from the AA-index[470] database, with more than 500 different experimentally measured physicochemnical and structural properties of all the natural amino acids. Even so, the experimental validation as well as other theoretical studies of the role of electron energy level frequencies along the protein sequence in the 3D structure of proteins and their role in imparting specificity to molecular recognition through resonance effects with counterpart molecules is still not established. Fundamental physical explanations of electron delocalization directly impacting protein folding are lacking because explicitly simulating such a system via DFT or *ab-initio* MD are computationally prohibitive. DFT implicitly assumes the Born-Oppenheimer approximation to hold that electron motion is so fast relative to the nuclear movement that the algorithm ignores any change in the position of atoms while calculating electronic behavior. Even *ab-initio* MD has to follow the Born-Oppenheimer approximation and only advances to the next frame and calculates new atomic coordinates after all the quantum mechanical calculations

are complete in the current time frame with static atom positions.

Furthermore, the exact proofreading mechanism behind electron energy level frequencies that might impart *specificity* to molecular recognition events remains a mystery. The nature of information transfer during signal transduction is still unanswered, along with the questions surrounding the near error-free execution of the signaling cascade in trillions of cells in millions of species.

Meanwhile, Random Matrix Theory has advanced since its early days in trying to model nucleonic energy spectra. Researchers in the Courant Institute and elsewhere proved that Random Band Matrices[425](A random symmetric matrix where only a few rows and columns are non-zero around the leading diagonal, and the rest of the matrix elements are zero. The width of the band denotes the number of filled rows or columns around the leading diagonal) belong to Gaussian Orthogonal Ensembles, and the width of the band around the leading diagonal determines the locality vs. non-locality of its eigenvectors. Non-local eigenvectors are associated with eigenvalues that do not coincide, instead repel on the number line, while localized eigenvectors are associated with overlapping eigenvalues.[426] Since such matrices model valence electrons, delocalized electron wavefunctions (non-localized eigenvectors) are related to non-degenerate energy levels, i.e., spacings between energy levels increase as faraway electrons start interacting, as in a conductor. In contrast, localized electrons in an insulator are related to degenerate energy levels.

Since the conductivity of different protein structures[108] tells us that alpha helices and helices, in general, are more conductive than Beta structures (sheets and turns), and beta structures are more conductive than random coils, RMT could potentially provide a means to evolve the secondary bond analogous connectivity of the molecular graph representation of proteins. Moreover, when proteins fold spontaneously, their conformational entropy necessarily goes down. The literature believes that some other type of entropy must go up to offset such change. One assumes that either en-

thalpic changes in the surrounding solvent or greater configurational entropy of solvent molecules in the vicinity of the proteins, compensates for the loss of conformational entropy when proteins fold. Both such assumptions are on shaky footing because protein folding might not be an enthalpic process (*one school of thought says that hydrogen bonds are satisfied either with water or with the protein itself, so no source of enthalpy change exists. There are differing opinions but there is a lack of experimental evidence on either argument, so possibly there is an yet unobserved enthalpic component that drives protein folding*). The water molecules around the proteins also get more structured as the proteins fold, indeed participating non-trivially and significantly in the processes of both protein folding and molecular recognition events as described above and elsewhere in the current work. Additionally, the proof for the presence and empirical validation of LBHBs was only available in 2019 when long-range enzyme cooperativity was observed to leverage LBHB circuits with water molecules.[462] Unfortunately, current understanding is insufficient to explain entropy-driven protein folding and specificity of molecular recognition fully. The few candidate entropy measures that might compensate for the loss of conformational entropy are either incorrect or unobserved or lack sound theoretical and empirical evidence.

From the discussions above, **the current study formulates its hypotheses as enumerated below** (for all cases, we assume that sufficient time has passed for the system so that the electrons have had enough time to randomly and uniformly arrive at all coordinates as applicable in the finite space, i.e., the system of particles achieves ergodicity):

(a) Just as Random Band Matrices (RBM) belong to Gaussian Orthogonal Ensembles, Random Graph Structured matrices (RGM), by being linear transforms of Random Band Matrices, should also belong to the Gaussian orthogonal Ensemble. The information entropy of the global eigenvalue distribution of such

RGMs increases with the number of edges in the graph structure.

(b) Since an increasing number of edges between nodes in a graph embedded in 3-D Euclidean Space is illustrative of protein folding, the decrease in configurational and conformational entropy of the protein structures must be anti-correlated with the information entropy of the energy level distributions of its electrons. In addition, the energy levels lose degeneracy and become more spread out.

(c) The eigenvectors of RGM in high connectivity regime are non-local. The non-locality illustrates that electrons tend to delocalize along specific paths during protein folding. They cannot delocalize along the backbone, but LBHBs that the residues form with water molecules and the backbone itself, provides a low resistance path for electron to delocalize. Single-well hydrogen bonds in water also allow electron delocalization in the relative scarcity of dangling p-orbitals.

(d) Although restricted in their individual spatial spread, electron delocalizations along LBHBs give rise to entangled systems of electrons with overall wavefunctions spread throughout the protein along the LBHBs. The Eigenvalue statistics, both global and local, of RGM, with the physical analog of electron energy values in a protein (or any molecule), can shed light on the connection between quantum mechanics, protein folding, specificity during molecular recognition, and information transfer during signal transduction cascades.

Fig 1.10 illustrates the major questions that arise during protein folding/unfolding processes, molecular recognition and signal transduction that the current body of work tries to probe by modeling electron interactions in proteins with the help of Random Graph-Structured Matrices (RGM).

1.3 Problem Statement

Ideally, end-to-end (Sequence to function without using any intermediate non data-driven models) AI models must include fundamental physical constraints to predict protein structures and intermolecular interactions between different classes of biomolecules. The AI models must also provide interpretable insight into protein folding and molecular recognition mechanisms. In reality, even the most accurate and widely celebrated AI is uninterpretable, impossible to execute by research groups, does not provide scientific inference, and cannot establish any specific representation of proteins as the one best suited to machine learning problems for structure prediction. Such inadequacies render the billion-dollar end-to-end AI models, such as DeepMind's AlphaFold[355,471–476], a glorified data repository. Since 2020, when AlphaFold2[472,473] emerged as the state-of-the-art protein structure prediction AI tool, researchers have used it as a first-step processing method to generate static structures for simple protein systems. They must follow it up with extensive explicit computational and empirical modeling. Consequently, researchers have to rely on slow and cumbersome X-Ray Crystallography, Cryo-Electron-Microscopy (CryoEM), and Nuclear Magnetic Resonance (NMR) techniques to experimentally probe the static 3D structures of proteins (except NMR, as it helps examine dynamic structures in solution as well). Researchers then either refine the preliminary structures or corroborate them by computationally intensive explicit modeling methods such as Molecular Dynamics, *ab-initio* methods, and Density Functional Theory to gain bio-physiological insight.

The current study aims to:

- Investigate the protein folding and molecular recognition problem from first principles using Random Matrix Theory by modeling proteins as a randomly weighted graph with connectivity that reflects the electronic connectivity within the molecule;

- Obtain physical insight from modeling energy Hamiltonians of electrons in a protein as Random Graph Structured Matrices;

- Postulate possible bio-physiological mechanisms for entropic compensation during protein folding, specificity of molecular recognition, and information flow during error-free signal transduction cascades along a signaling pathway; and

- As data-driven models that are usually in the form of deep neural networks also are a series of operations performed on random matrices, the current study aims to establish metrics defined on random graph-structured matrices that scientists can use in the loss functions of molecular graph convolution networks (with any architecture) to include fundamental knowledge of protein folding into the models in the future.

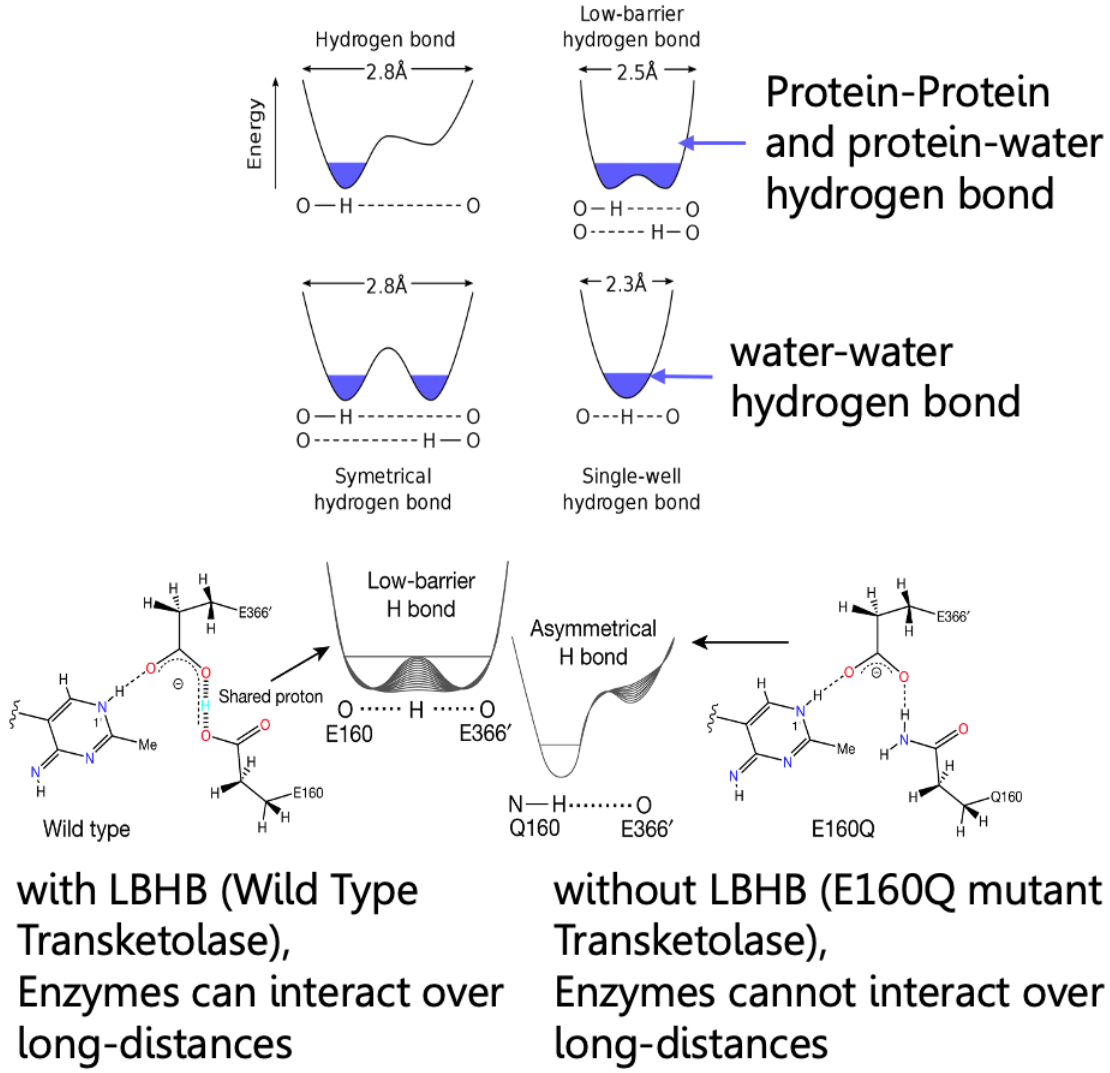


Figure 1.9: **Low Barrier Hydrogen Bonds in enzyme function**: The figure shows how Low Barrier Hydrogen Bonds (LBHBs) are beneficial when they occur both within the proteins and at their interface with the water molecules.[462] Within the protein they are stable, but are unstable and keep switching when they exist between water molecules and at the surface of water and proteins. Such a switching LBHB network acts as a circuit in long range cooperativity between enzymes such as transketolase. Two enzymes need to cooperate for thiamine synthesis. For such cooperation to happen, hydrogen bonds must be LBHBs, the water has to be structured, and thermal noise must be filtered out so that only certain frequencies exist that can enhance the quantum coherence of both protons and electrons of the hydrogen atoms. Image adapted from: *Dai Shaobo et. al.*[461]

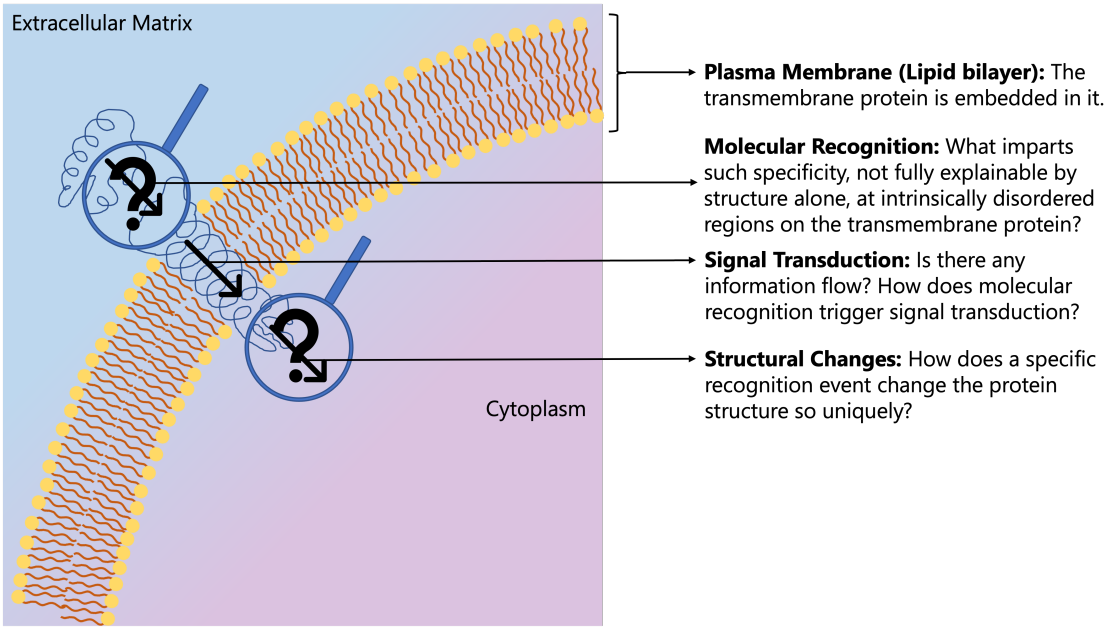


Figure 1.10: **Major knowledge gaps in protein structure-function relationships**: The figure shows the major issues and knowledge gaps that still exist in fully exploring entropic compensation and driving forces of protein structure adoption, specificity during molecular recognition that happens at intrinsically disordered regions, subsequently triggering an error free signal transduction cascade in cells.

Chapter 2

ARTIFICIAL INTELLIGENCE FOR PROTEIN STRUCTURE PREDICTION

Understanding how protein structures come about is the first step to comprehending the principles governing specificity in molecular recognition and signal transduction precision. Various protein folding theories exist, mainly from a classical mechanical and thermodynamical point of view. Several explicit computational models exist that fit the theories with varying degrees of agreement. On the other hand, experimental methods of circular dichroism, X-ray Crystallography, and Nuclear Magnetic Resonance can only tell us what the structures are but not how and why they formed nor how they guarantee the specificity and precision in the molecular recognition and signal transduction phenomena. Moreover, empirical methods of structure determination are cumbersome and painstakingly slow. They also cannot keep up with the pace at which modern sequencing technologies are discovering, identifying, and storing unmanageable amounts of the genome and proteome of many species.

Explicit computational models such as Molecular Dynamics, hybrid QM/MM, *ab-initio*-MD, and Density Functional Theory are relatively slow and computationally intensive. Although they offer physical inference to scientists ubiquitously, in academia, research labs, and industry, they do not perform accurately and quickly enough for practical purposes of structure prediction. They must be complemented by data-driven methods to keep pace with the discovery rate of new biomolecular targets and proteins.

The current chapter describes the state-of-the-art AI tool in data-driven protein structure prediction, the open-source dataset it relies upon, and a preliminary attempt to

use a deep learning technique called Variational Autoencoder(VAE)–Generative Adversarial Network (GAN)[477] for in-house *de-novo* protein structure prediction.

2.1 The PDB & AlphaFold

2.1.1 Protein data Bank

The Protein Data Bank (PDB)[96,291] is a data repository for the 3-dimensional empirical structures of biomacromolecules, such as proteins and DNA/RNA and their complexes with each other. The data commonly comes from X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy experiments. Biologists and biochemists from around the world submit data to the repository. In addition, the structures are freely available on the internet via the websites and web portals of the PDB's member organizations (PDBe[478], PDBj,[479] RCSB[291], and BMRB[480]). The Worldwide Protein Data Bank, wwPDB, oversees the PDB repository.

The PDB is critical for studying structural biology. Most prominent scientific journals and a few funding agencies now mandate scientists submit their biomolecular structural data to the PDB repository. Many other structure databases use biomolecular structures deposited in the PDB. The PDB repository and its holdings list are updated weekly (UTC+0 Wednesday). As of April 1st, 2020, the PDB comprised:

(a) About 135170 proteins, 2097 nucleic acids, and 6945 protein-nucleic acid complexes from X-Ray Crystallography;

(b) About 11337 proteins, 1325 nucleic acids, and 264 protein-nucleic acid complexes from NMR; and

(c) About 3475 proteins, 35 nucleic acids, and 1136 protein-nucleic acid complexes from cryoEM (Cryogenic Electron Microscopy) experiments.

About $134,146$ structures in the PDB repository have a structure factor file. Likewise, $10,289$ structures have an NMR restraint file, $4,814$ structures have a chemical

shifts file, and 4,718 structures in the PDB have a 3-D EM map file deposited in the Electron Microscopy (EM) Data Bank[481].

Most structures come from X-ray diffraction, but about 10% of structures are from protein NMR. When using X-ray diffraction, one obtains approximations of the coordinates of the protein atoms, whereas NMR estimates the distance between pairs of protein atoms. One solves a distance geometry problem to get the final conformation of the protein from NMR. After 2013, an increasing number of protein structures come from cryo-electron microscopy. The electron density map is viewable for PDB structures determined by X-ray diffraction with a structure factor file. The "electron density server" stores the data of such structures.

The format initially used by the PDB was called the PDB file format. The width of computer punch cards restricted the original format to 80 characters per line. wwPDB phased in the "macromolecular Crystallographic Information file" format, mmCIF,[482] which is an extension of the CIF format, in 1996. mmCIF slowly became the typical format for the PDB repository by 2014. In 2019, the wwPDB announced they would only accept depositions for crystallographic methods in mmCIF format. In 2005, an XML rendition of PDB, called PDBML,[483] was defined. These days, the structure files are accessible and downloadable in any of the three formats, though increasingly, newer structures do not fit the legacy PDB format. The database labels each structure submitted and subsequently vetted in the PDB repository with a four-character alphanumeric identifier, its PDB ID. (This is not a unique identifier for biomolecules because quite a few structures for the same biomolecule in different environments and conformations exist in the PDB with different PDB IDs.) All the AI algorithms in operation currently use the PDB as their main data source and take structure files in one of the three formats enumerated above.

2.1.2 AlphaFold

AlphaFold, an AI program from Google's DeepMind, is the state-of-the-art model for protein structure prediction.[471] It secured first place in the Critical Assessment of Techniques for Protein Structure Prediction (CASP)[360] competition in 2018 by predicting the most accurate structures for target proteins with unknown templates from all participants. A newer version of AlphaFold, AlphaFold2, again snitched first place in the CASP 2020 competition.[473] It achieved a > 90 score for about two-thirds of the target proteins in CASP's Global Distance Test (GDT), which measures the accuracy of predicted structures from empirically obtained structures of the target proteins in terms of a distance measure, usually Euclidean Distance in $\mathbb{R}^3$ or a root-mean-squared-deviation (RMSD) from the backbone conformation. A score of 100 would be an exact match with the empirical structures (backbone conformations of C_α-atoms obtained from X-Ray Crystallography or NMR experiments on purified, cooled, and crystallized proteins).

The 2018 version of AlphaFold (AlphaFold1) relied on homology modeling on evolutionarily conserved structures obtained from similarly conserved DNA and Amino-Acid sequence motifs to find residues that caused synergistic effects in changing protein structures, regardless of the proximity of the residues in the sequence. The assumption was that such residues must be close in physical 3-D space, enabling the estimation of a contact map, a sort of symmetric square matrix with binary or decimal values, relaying the presence/absence or strength of the contacts between C_α atoms of the backbone. Contact matrices are easily converted to distance matrices with some probability distributions over the matrix elements. AlphaFold1 estimated the probabilities over the distance map by minimizing potential energy obtained from the force field fit with parameters obtained from the PDB data. It treated the potential energy as a loss function (it is, therefore, a type of PINN). The network architecture was a Residual Neural Network (RNN)[484] with about 21 million parameters, taking both

1-D and 2-D inputs, including contact and distance matrices, dihedral angle tensors, evolutionary profiles, and co-evolution features from sequence alignment and homology modeling. AlphaFold1 predicts a distance matrix as a fine-grained distribution of distances over the matrix elements denoting individual C_α atoms. Additionally, AlphaFold1 also predicts a tensor with ϕ and ψ dihedral angles (ϕ^i is the dihedral torsion angle between the $C^{i-1}, N^i, C^i_\alpha, C^i$, and ψ^i is the dihedral torsion angle between the $N^i, C^i_\alpha, C^i, N^{i+1}$ functional-groups, subtended at the i^{th} residue along the backbone). Together, the distance matrix and the dihedral angle tensor help generate a 3D structure of the target protein. The AlphaFold1 trained on $29,000$ 'clean' protein structures. The DeepMind group identifies that their first approach, combining localized physics with a force field derived by parsing structural data in the PDB repository, tends to over-account for physical interactions between residues located nearby in the protein's sequence compared to interactions between residues further apart along the protein's backbone. As a result, AlphaFold1 tends to prefer models with more α-helices and β-sheets than is the case in reality. Since α and β structures are the most prevalent, they dominate the data. Others, such as M.AlQuraishi at Harvard, attempted to generate even 'cleaner' databases[370] from the PDB repository to create a better training set.

The 2020 version, AlphaFold2, is significantly different from AlphaFold1.[472] The software strategy used in AlphaFold1 comprised many modules, each trained individually, that then produce the force field that modifies the physics-based force field. AlphaFold2 substituted this with a system of smaller sub-networks associated concurrently into a solitary differentiable end-to-end model, based entirely on feature estimation, which trains comprehensively as a single entity (fig 2.1). Physics (Newtonian and thermodynamic) of the residues, in the form of energy refinement based on the AMBER[206] force field, is applied only as a final refinement step once the neural network has converged to a predicted structure with high confidence (a certainty or confidence label that the AI also produces). The refinement step only slightly ad-

justs the AI predicted structure. A crucial part of the AlphaFold2 system is two new modules, assumed to possess a "transformer" architecture[485], which progressively refines the weights on the edges of a graph representation of the input structures. The graph represents the interaction between (a) Two amino acid residues of the protein, (b) Between each amino acid location, and (c) Each of the different sequences in the input sequence alignment. Inside the architecture, these refinement transformations contain layers that use the "attention mechanism,"[354] a type of hierarchical kernel convolution for these interactions, to learn the context of the residues from training data. These transformations iterate between them with the refined residue/residue information from the first transformer network feeding into the subsequent refinement of the residue/sequence information. Then the improved residue/sequence data feeds into the next residue/residue transformer training run. The final structure prediction module, which also uses transformers, takes in the output from the previous modules to train itself iteratively.

In an example shown by the DeepMind group, the final structure prediction module achieved the correct topology for the target protein's backbone on its first iteration, with a GDT score of 78 but a large number (90%) of nonphysical bond angles and lengths.[473] With successive iterations, the number of such nonphysical features fell, and the GDT score grew. By the third iteration, the GDT score of the prediction inched closer to 90, and by the eighth iteration, the number of nonphysical features was hovering close to zero.[473]
However, despite being the most accurate model[472] for *de-novo* protein structure prediction, AlphaFold 2 suffers from the following limitations[474–476]:

(a) AlphaFold2's accuracy is not high enough for one-third of its predictions;

(b) It does not reveal the mechanism or governing physics of protein folding;

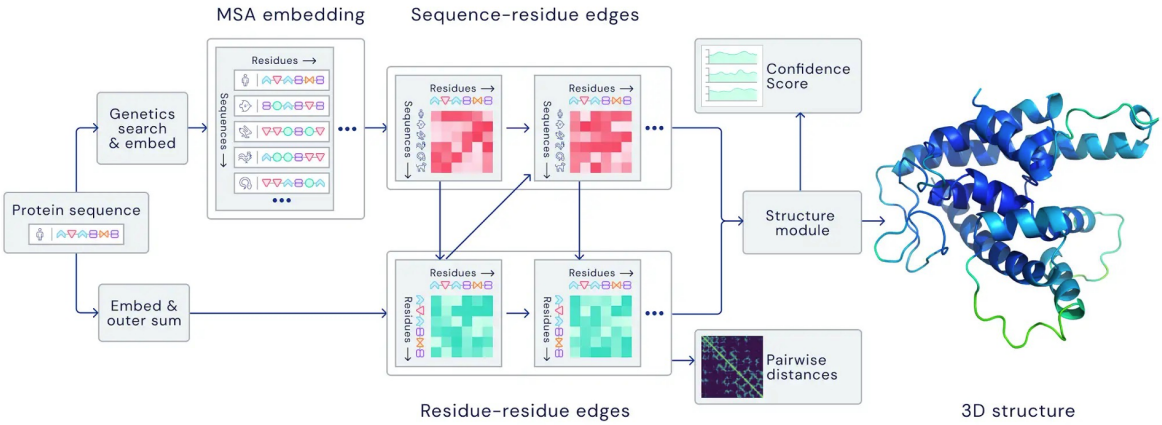


Figure 2.1: **DeepMind's AlphaFold2 network architecture**: The figure shows the multiple transformer networks employed in the AlphaFold2 algorithm. One transformer first updates its weight matrices (via backpropagation) with context information about amino acids from their location in the sequences and their evolutionary relationships through multiple sequence alignment, feeds the learned relationships iteratively as a Bayesian prior to the second transformer network that uses the information to learn more about the amino-acid context and interactions in the 3-dimensional structure. The updated weight matrices of the first and second transformer then act as a prior for the third transformer network that predicts secondary structures (which need minor minimization with the AMBER force field. The choice of AMBER was arbitrary). In the next iteration, the weight matrices of the second network act as a prior for the first, which then again updates its weights before passing it again to the second network and then to the third network. The cycle continues until the overall loss has reduced below a necessary threshold and the model has converged. *Figure Source: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology*

(c) It has minimal success in predicting structures for Intrinsically Disordered Regions at the active site of proteins, constituting about 30% of all native structures;

(d) It is not applicable for data not in the PDB database, for example, several peptides, micro proteins, protein-DNA complexes, protein-RNA complexes, protein-Glycans, protein-lipids, peptidoglycans, peptidolipids, and small molecules that are crucial for protein function. It only trained on monomeric data, i.e., single

protein tertiary structures. Therefore, it can only output one conformation for proteins with many conformations, such as short and disordered proteins[486]. However, the training database includes protein complexes after the October 2021 update (AlphaFold-Multimer)[487,488]. AlphaFold-Multimer succeeds about 70% of the time at predicting single protein-single protein interactions. Other molecules are still excluded from training data and cannot be included in any short-term updates. Performing such a feat requires a universally agreed-upon representation of biomolecules;

(e) It cannot model the interaction between proteins and other molecules because it is necessarily trained on experimental data taken from static, frozen, and purified protein crystals. Such crystals are not in their native aqueously solvated and dynamic conformations and time-dependent interactions that lead to molecular recognition and signal transduction;[489–491]

(f) It depends on the force field estimated from the PDB database, as in the case of the Rosetta Monte Carlo structure prediction method;

(g) Its training sample included 170,000 protein structures, an infinitesimal fraction compared to the sequence space of all possible naturally occurring proteins on earth;

(h) It was trained on Google's servers and took up the processing power of about 100-200 GPUs for several weeks, totaling about $10000 worth of wholesale compute time for converging on the structure of a single protein. In comparison, the University of Washington's Mox supercomputing cluster has about ten reasonably powerful GPUs distributed among all the researchers at the University, and they can only use the GPUs for a few hours to a handful of days at a time to reduce congestion;

(i) The deep neural network needed tuning of up to 21 million parameters (most deep neural networks are 'deep' because they have several network layers to operate in the double-descent regime of optimization methods like stochastic gradient descent and its plethora of variations). However, the sheer number of parameters necessarily renders the model entirely uninterpretable by scientists and humans, harming the trustworthiness of any physical inference made from the analysis of the network; and

(j) After the training process is complete, it takes several days to converge on each new structure for every new protein, a timescale similar to more rigorous computational methods such as MD or DFT.

AlphaFold1, in comparison, had even higher inconsistencies and overpredicted α and β structures due to the dataset being biased toward more prevalent structures and trained on only 29000 samples.[360] In addition, the AlphaFold1's GitHub site's ReadMe file states that the code cannot predict the structure of an arbitrary protein sequence, instead only on the sequences in the CASP-13 dataset. Notably, both versions of AlphaFold outperformed all other competitors in the CASP competitions by a significant margin. In the CASP-13 competition, AlphaFold1 gave the best-predicted structures for 25 out of 43 proteins.[360] In addition, it obtained a median score of 58.9, ahead of the two runner-up models at 52.5 and 52.4 (both deep learning models), which is a significant margin. In CASP-14, AlphaFold2 scored 92.4% on the GDT, a score almost identical to the X-Ray Crystallography data, a score that AlphFold1 only almost got in two of its 25 best predictions. Comparatively, 88% of the other competing models in the CASP-14 competition scored more than 80 on the GDT, with AlphaFold2's median score of 87 on the most challenging protein targets.[473] Furthermore, 88% of AlphaFold2's predicted structure had less than 4Å RMSD from X-Ray Crystallography structure's C_α atoms on the backbone, 76% of predicted structures had better than 3Å RMSD, and 46% better than 2Å accuracy of RMSD. How-

ever, two of the three structures that AlphaFold2 performed the worst on were NMR structures obtained from proteins solvated in water, their native dynamic state.[486,489] The third structure is a protein complex. AlphaFold2 failed in these situations as its training data did not consider such physically relevant situations. AlphaFold2 currently limits its usability to produce large libraries of predicted structures that can act as a first step for other rigorous computational models such as MD, ab-initio MD, and DFT. The limitations enumerated for AlphaFold1 and AlphaFold2 also extend to several other deep learning models, regardless of details in network architecture, training protocol, and optimization algorithms. Lastly, none of the models use strategies from quantum mechanics or first principles in their neural network architecture. In summary, there is still much to be done to realize a genuinely physics-informed and interpretable AI model for protein structure prediction to discover underlying physical mechanisms of protein folding and molecular recognition. It is not yet clear to what extent structure predictions made by AlphaFold2 will hold up for proteins bound into complexes with other proteins and other molecules, despite the creation of the AlphaFold-Multimer.[492] As a significant fraction of the essential biological machinery in a cell constitute such complexes or relates to how protein structures become modified when in contact with other molecules (molecular recognition and signal transduction), this area will continue to be the focus of considerable experimental and computational attention.

With AlphaFold2 being highly uninterpretable, it is unclear to what extent it is limited in its ability to recognize novel folds when such folds are underrepresented in the known protein structure databases. The same concern extends to intrinsically disordered structures with no fixed shape. In addition, it is also ambiguous how suggestive the frozen protein structures in crystals are of the dynamic forms found in the cells *in vivo*. AlphaFold2's difficulties with structures obtained from NMR methods[489] do not foster confidence.

In its prospect as a means for drug discovery, while the resolution of AlphaFold2's

structures may be excellent, the accuracy for active-sites (which aren't static to begin with in most cases) needs to be even higher.[491] So AlphaFold2's predictions may only be a small help in such contexts. Moreover, the prediction of small-molecule binding to protein targets is worse. Simply predicting a complementary structure isn't enough. One must glean physical inference from computational models to understand how the proteins' structure and the binding event at the active sites effects function, and how that fits within the broader scope of interconnected biological processes in organisms.

Also, since AlphaFold2 processes protein and nucleotide-only sequences by design, additional associated biomolecules are not assessed. In the absence of metallic ions, cofactors, and post-translational modifications such as protein glycosylation and phosphorylation from AlphaFold models,[493] scientists need to peruse other databases, such as UniProt-KB, for likely absent molecular segments, as they can play an essential role in the protein's folding and function. However, scientists can manually add post-translational modifications to the predicted AlphaFold2 structures and use MD or other explicit simulation methods to refine the structure.

DeepMind and the EMBL-EBI launched the AlphaFold Protein Structure Database (AlphaFold-DB)[478,494] on July 22^{nd}, 2021. The database initially contained AlphaFold-predicted structures of proteins from the complete UniProt data of the proteome of humans and 20 model organisms, totaling over 365,000 proteins. The database does not incorporate proteins with fewer than 16 (the vast majority of intrinsically disordered peptides, active sites, and micro proteins) or more than 2700 amino acid residues. The AlphaFold team intends to incorporate more predicted protein structures in the AlphaFold-DB to include most of the UniRef90 database's more than 100 million proteins. As of August 1st, 2022, all the predicted structures from UniRef90 are available on AlphaFold-DB.[495]

The AlphaFold-DB reduces the scarcity of structural data of simple proteins and enables other groups to develop data-driven models to understand the physics behind

protein folding, molecular recognition, and signal transduction. The need of the hour is to develop novel AI algorithms that can take in graph-structured protein data, provide interpretable models to glean physiological mechanisms, and relate the structures to observed functions of these complex biomolecules.

The following section describes a preliminary attempt made by the author of the current study at creating such an AI model to probe intrinsically disordered regions of proteins.

2.2 Preliminary VAE-GAN model for Predicting Intrinsically Disordered Peptide Structures

The primary emphasis of numerous recent high-throughput analyses is a molecular depiction of the entire set of functional, active sites of proteins in the cells, from protein complexes to post-translational modification sites.[493] It is becoming increasingly evident that the functional aspect of the proteins extends well beyond stable, structured domains. A considerable share of biologically relevant functional biomolecular interactions is mediated by active sites within intrinsically disordered regions of larger proteins that get recognized and post-translationally modified by complementary domains of the interacting partner.[496] Short interaction motifs (peptide motifs and post-translational modification, PTM, sites) are usually less than ten residues in length and allow both wide functional variety and density to contain polypeptide domains.[497] Moreover, on evolutionary timescales, they evolve rapidly, appearing or disappearing with equal speed, granting unprecedented evolutionary adaptability to the interactome (The parts of the proteome that interacts with one another and with other molecules).[498] In their paper in 2014,[499] P. Tompa *et. al.*, created a table that shows several types of commonly found peptide motifs in the PDB and Eukaryotic Linear Motif (ELM) repository,[500] such as classical binding, trafficking, targeting, docking, and degron motifs, moiety addition sites, cleavage, and structural modifica-

tion sites. The current study uses the identified motifs to parse through the data in the PDB database and obtain all the disordered structures from the PDB repository (the vast majority of such disordered structures are parts of larger proteins). The result was intrinsically disordered domain data of about 62 million structures belonging to about 3.7 million unique sequences. On average, each sequence has 16 different 3D structures, reinforcing the notion of intrinsic disorder.

2.2.1 Variational Autoencoders, VAE

The nature of the data for intrinsically disordered regions (each sequence has a statistical ensemble of structures it can take) mandates a variational model where a deep network attempts to encode the multidimensional inputs in terms of a few parameters that estimate the distribution from which the input samples were drawn. Variational autoencoders (VAEs)[501] permit statistical inference problems (such as estimating the distribution of one random variable from another random variable) to be typecast as statistical optimization problems (i.e., to estimate the parameters that minimize some target function). They map the input variable to a multivariate latent distribution, usually in much lower dimensions. In a VAE, the input data samples come from a parameterized distribution (the Bayesian prior), which in this case is the primary distance maps of the protein structures. The encoder and decoder are trained together such that the output minimizes the Kullback–Leibler divergence (D_{KL}, a distance measure between distributions that has connotations of mutual cross entropy, given in equation 6,7) between the parametric predicted posterior and the true posterior distribution.

From a mathematical formalism perspective, given an input dataset x characterized by an unknown probability distribution $P(x)$, the goal of the VAE-algorithm is to model or approximate the data's true distribution $P(x)$ using another parameterized distribution p_θ with parameters θ. If z is a random vector joint-distributed with x, then z represents a latent encoding of x. Marginalizing over z gives:

$$p_\theta(x) = \int_z p_\theta(x, z)dz \tag{2.1}$$

where $p_\theta(x, z)$ is the joint distribution of the input observations x and its latent space encoding z. According to the chain rule of differentiation, equation 2.1 can be written as

$$p_\theta(x) = \int_z p_\theta(x|z)p_\theta(z)dz \tag{2.2}$$

In basic VAEs,[477] z is usually a finite-dimensional vector with $z_i \in \mathcal{R}$ and $p_\theta(x|z)$ is a normal distribution. One can now define the relationships between x and z as: prior $p_\theta(z)$; likelihood $p_\theta(x|z)$; and posterior $p_\theta(z|x)$. Unfortunately, as is the case in many deep learning problems, explicitly computing $p_\theta(x)$ is prohibitive and intractable. One therefore introduces a function to approximate the posterior distribution as $q_\phi(z|x) \approx p_\theta(z|x)$ which is the encoder network with $\phi_i \in \mathcal{R}$ parameterizing q. The decoder computes the conditional likelihood distribution $p_\theta(x|z)$.

The VAE algorithm minimizes the reconstruction loss (between input and output) by optimizing the parameters θ and ϕ to make $q_\phi(z|x)$ as closely match $p_\theta(z|x)$ as possible. The loss function (re-purposing equation 6) is therefore given by:

$$\begin{aligned} D_{KL}(q_\phi(z|x)||p_\theta(z|x)) &= \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[ln\frac{q_\phi(z|x)}{p_\theta(z|x)}\right] \\ &= ln(p_\theta(x)) + \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[ln\frac{q_\phi(z|x)}{p_\theta(x,z)}\right] \end{aligned} \tag{2.3}$$

where $\mathbb{E}(\cdot)$ is the expectation. The evidence lower bound, ELBO, is given by:

$$L_{\theta,\phi}(x) := ln(p_\theta(x)) - D_{KL}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) \tag{2.4}$$

The optimization algorithm boils down to $\theta^*, \phi^* = \text{argmax}_{\theta,\phi} L_{\theta,\phi}(x)$. The algorithm maximizes the log-likelihood of the observed input data and simultaneously minimizes the divergence of the estimated posterior q from the exact posterior p_θ.

2.2.2 Generative Adversarial Networks, GAN

Generative Adversarial Networks (GAN),[502–504] on the other hand, is a class of machine learning frameworks where two networks compete in a zero-sum game. Given a training set, the generator network of the GAN learns to generate new data supposedly drawn from the same statistics as the training set. The discriminator network learns to distinguish between real inputs and generated data more rigorously with each iteration. The generator is unconstrained and its only purpose is to confound the discriminator. The basic GAN from Ian Goodfellow *et al.*,[502] is defined as:
Each probability space (Ω, μ_{ref}) defines the support of the game. There are 2 players, the generator network and the discriminator network. The strategies available to the generator is the set of all probability measures μ_G on Ω, denoted as $P(\Omega)$. The strategies available to the discriminator is defined by the set of Markovian kernels $\mu_D : \Omega \to P[0,1]$ where $P[0,1]$ is the set of probability measures possible on the support [0,1]. Since the GAN is a zero-sum game, the objective function to be optimized is given by:

$$L(\mu_G, \mu_D) := E_{x\sim\mu_{ref}, y\sim\mu_D(x)}[ln\ y] + E_{x\sim\mu_G, y\sim\mu_D(x)}[ln(1-y)] \tag{2.5}$$

where G and D stand for the generator and discriminator respectively. The generator aims to minimize the objective function while the discriminator tries to maximize it simultaneously. The task of the generator is to estimate $\mu_G \approx \mu_{ref}$, i.e., to bring its output distribution as close as feasible to the reference distribution by minimizing the Jensen-Shannon divergence which is a mutual information measure. The Jensen-Shannon divergence is a constrained case of the D_{KL}-divergence, with symmetry considerations and that it is always finite. The task of the discriminator is to output a value close to 1 when the input appears to be from μ_{ref} and 0 if it appears to be from μ_G. Practically speaking, the loss function of the discriminator is usually a cross entropy function (equation 3).

2.2.3 VAE-GAN

Since the GAN and the VAE are so uniquely compatible, **the preliminary AI model in the current study that is developed to predict secondary structures of the intrinsically disordered regions of proteins, uses a VAE-GAN architecture**[477,505] (fig 2.2). While a VAE learns to encode the given input (say, a distance map obtained from the primary structure connectivity of disordered regions) and then reconstructs it (but this time with secondary bonds incorporated) from the encoding, a GAN works to generate new data which can't be distinguished from real data, in this case secondary structures. The crucial point behind the operation of a VAE-GAN is that in case of the VAE section, we use the latent representations generated by an encoder for various tasks. The decoder of the VAE becomes the generator for the GAN in such a hybrid model. **Since the VAE-GAN is widely popular and has many statistical guarantees, the current study utilizes it to generate secondary structures for intrinsically disordered regions.**[477,505]

The main loss function for the discriminator is the binary cross entropy loss between the real/fake (1,0) labels. In information theory, the binary cross-entropy between two probability distributions p and q over the same random variables, measures the expected number of bits needed to identify an event drawn from the random variable if a coding scheme used for the random variable is optimized for an estimated probability distribution q on a sample of such events, rather than the true distribution p on the population of the events. By using equation 3, and considering the binary case, it is defined as:

$$\mathbb{H}(x)_{cross} = \sum p(x) log_b[q(x)] \tag{3}$$

$$\mathbb{H}(x)_{cross} = \sum p(x) log_2[q(x)] \tag{2.6}$$

where the base b depends on the nature of the discrete random variable x. For binary cross-entropy, b is 2.

In the current study, the VAE-GAN model is applied to intrinsically disordered proteins' structural data (architecture shown in fig 2.2). The model takes in distance maps of the primary structure of the sequence, and attempts to map them to a latent space that parameterizes the sample distribution in lower dimensions, the mean and standard deviation. Then a decoder network (which is also the generator) recreates the distance maps from the latent space, but also incorporates secondary bonds in the recreated distance map. The discriminator then tries to distinguish which of its two inputs (one from generator and one from the real secondary structures sampled from the PDB) is physically real. The algorithm is written as the following steps:

Algorithm:

- Extract sequence motifs from the paper by P. Tompa *et al.*,[499] that correspond to intrinsically disordered regions;

- Parse the PDB database and extract all structures that match the motifs obtained;

- Test to see if the extracted structures are indeed intrinsically disordered by randomly sampling from the data and visualizing using PyMol[102];

- Use known (ϕ, ψ) dihedral angle pairs corresponding to $\alpha-$helix and β-structures to identify any sequences with those structures and delete them from the dataset. We end up with about 62×10^6 unique structures with 3.7×10^6 unique sequences, an average of ~ 16 structures per sequence;

- Manually add hydrogen atoms to the N–terminus to make it be an $-NH_3^+$ functional group, and detach hydrogen atoms (if any) from the C–terminus to make it be a $-COO^-$ functional group using PyMol[102]. Then add more hydrogen atoms, if necessary, to the side chains based on their pKa values from the literature. We consider the operating pH to be 7. Any functional group

with pKa less than 7 will have an added hydrogen atom, while those above 7, will have lost a hydrogen atom;

- Represent all the cleaned structural data in the '.PDB' format for the intrinsically disordered regions, in terms of (x, y, z) coordinates with the origin at the Nitrogen at the N–terminus;

- Convert all the (x, y, z) data into an all atom distance matrix by using the Euclidean distance formula in 3-dimensions:
  $d = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$ where (x_i, y_i, z_i) & (x_j, y_j, z_j) are the coordinates of the i^{th} & j^{th} atoms respectively;

- From known sequence (contained in the '.PDB' file), select rows and columns from the distance matrix that correspond to the various known lengths of constituent covalent bonds only: $C - C, C - O, C = O, O - H, N - H, S - H$, etc., and create the 'primary structure distance matrix'. The original distance matrix is the 'secondary structure distance matrix';

- Add padding zero rows and columns to make all the data the same length. Since the longest motif was 7 amino acids long, the largest distance matrix was assumed to be of the size corresponding to a peptide with sequence PW-7: WWWWWWW, as W (Tryptophan) is the amino acid with the most number of atoms in its side-chain, thereby corresponding to the most number of rows and columns possible. All other distance matrices were padded with zero valued rows and columns to match the size of PW-7: 168×168;

- Split the data into an 80 : 20 ratio randomly. the larger chunk is kept for training while the other is kept for testing the VAE-GAN model;

- Send the primary structure distance matrix (now padded) from the training set

as iterative inputs to the encoder network of the VAE-GAN. The encoder is a standard PyTorch implemented convolutional neural network that first increases the matrix size (by sequential matrix multiplication with randomly weighted matrices of pertinent dimensions that follow the rules of matrix multiplication to determine product matrix dimensions) until 512×512 and then reduces in size to 256×256, then 64×64, 32×32, then 16×16, then 8×8 which then reduced to a 4×4 square matrix where each element, after completion of training, would represent a parameter of the statistical distribution from which the primary structure inputs were sampled;

- The Generator, which is also the decoder, then blows up the small 4×4 matrix all the way until it reaches 512×512 size, before compressing it to the original 168×168 size. This output is now considered as the 'generated secondary structure', and is sent into the Discriminator, with a 'False'(0) label as the 'known output' that the discriminator is expected to match with its own prediction;

- The real 'secondary structure distance matrix' is simultaneously sent to the Discriminator with a label 'true'(1) that the discriminator is also simultaneously supposed to match with its own prediction. The two 168×168 matrices, corresponding to the generated and real secondary structure distance matrices, are then convolved with random kernels until the size grows to $512 \times 512 \times 2$ tensor, and then falls gradually to 10×2 and finally to 1×2 before applying a sigmoid function finally to convert the final output to either 0 or 1 value, corresponding to the False or True labels respectively;

- The Discriminator error is a binary cross entropy term between the predicted [True/False] labels and actual [True/False] labels. The error gradient is then back-propagated and every matrix along the way, from the end of the discriminator to the beginning of the encoder is updated down the gradient. The

learning rate or step-size down the gradient at every matrix, is set to 0.001, to be tuned later as a hyperparameter;

- After the first iteration, we send in the next primary structure randomly sampled, and the process repeats again until the epoch error has converged and error is no longer reducing even with more and more iterations. We introduced this convergence criterion because ideally, training with 62 million samples in every single epoch, is infeasible in a university setting;

- Then the second epoch starts and the training starts again with more randomly sampled structures, until the epoch error has again converged. The entire process repeats until overall error per epoch is no longer reducing or changing, i.e., the AI model has converged;

- We plot the training error per epoch and throughout the run, and then use the learned weights to predict new secondary structures for the test sample.

2.3 Results & Limitations of the VAE-GAN & Necessity for Matrix Representation of Proteins with Tractable Metrics

The following fig 2.3 shows the discriminator error over training epochs. Clearly, the discriminator wins the zero-sum game after just 9 epochs, showing that it learns too fast compared to the generator, while the generator is either stuck in a local minimum, or a saddle point. The discriminator can almost perfectly distinguish the μ_G and the μ_{ref}, as defined in the previous section from each other, i.e., generated and real structures from each other. The generator seems to be stuck with a very high reconstruction error, no matter which direction it steps along the gradients, implying that the gradient might be vanishingly small. PyMol visualizations of predicted structures for sequences in the test set do not contain any structure that is physically real. It is apparent in the example of one intrinsically disordered structure for the

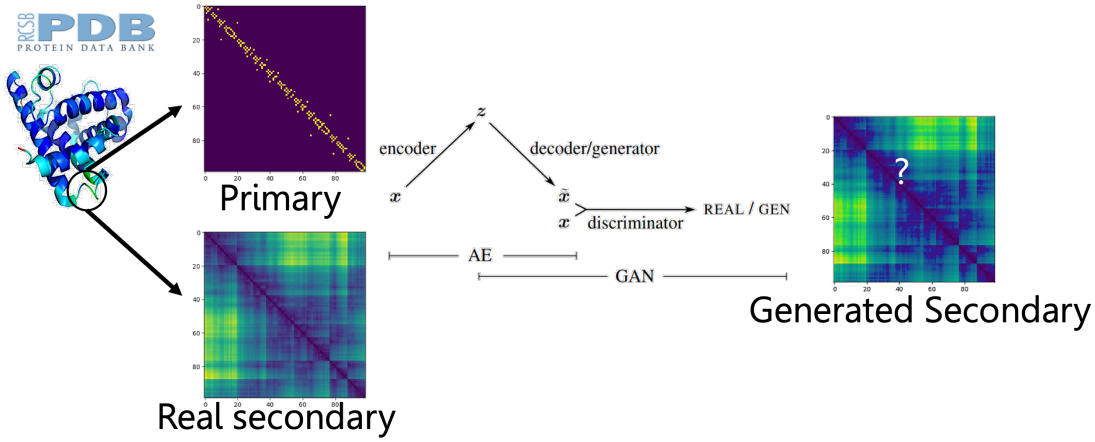


Figure 2.2: **Variational Autoencoder-Generative Adversarial Network Architecture for disordered structured prediction**: The VAE-GAN network designed for training on disordered structure data that have a $1 \rightarrow$ *many* mapping between sequence and structures. Several other models such as Long short term memory (LSTM)[506] models will not work. VAE-GAN takes in the primary structure x of a peptide as an all-atom distance map, and maps it to a latent space z. The unconstrained decoder then generates a distance map $\tilde{x}$ which is the generated secondary structure. $\tilde{x}$ is then fed into the discriminator, along with one of the real secondary-structures for the sequence. The discriminator then attempts to correctly label if the secondary structure comes from the real structural data or from the generator.

sequence FGVAAIF that the generator predicted from the sample taken from the test set, as shown in fig 2.3. The Generator has learned that there is a kink but it is still unable to learn that the backbone must be connected without breaks. This tells us that there is not much learning happening in the generator after the first few epochs, and therefore it is not generalizable in its current form, and doesn't even predict physically real structures.

One way to deal with vanishing gradients is the Wasserstein GAN.[507,508] However, we have no guarantee if the gradient is indeed vanishingly small by the time it reaches the initial layers of the generator as back-propagation proceeds, without actually monitoring the gradients at every single step for every single vector in ev-

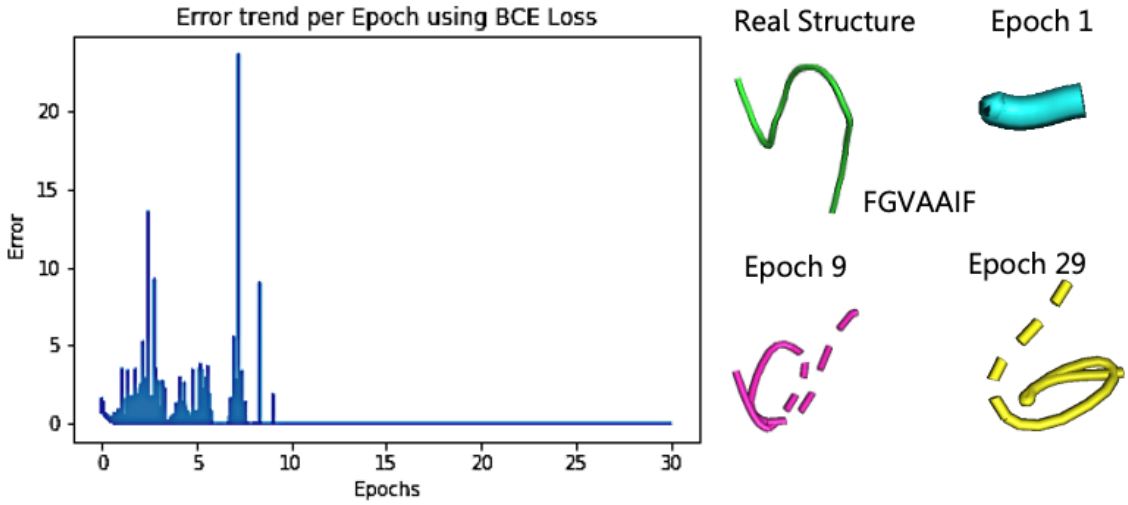


Figure 2.3: **Discriminator error over training epochs for the VAE-GAN**: The VAE-GAN trained for about 30 epochs, while the overall Binary Cross Entropy (BCE) loss has converged to zero starting at 10 epochs, and did not change. This result indicates that the gradient in the generator might be vanishingly low, so that there are no significant updates are happening to the generator network. This could also be a result of a faulty representation, and inadequate physics that cannot constrain the Generator in any meaningful way. The example peptide structure, on the sequence FGVAAIF sampled from the test set, shows visually (rendered in PyMol) that the generator has stopped learning meaningful data for the structure prediction after a few epochs at most, and the predicted structures are not physically real.

ery layer of the VAE-GAN, which is a cumbersome process. Before we even try to adopt the Wasserstein-GAN[509] formalism that simply uses an even more involved and complicated distance metric (than the Jensen-Shannon divergence described for the basic GAN in equation 2.5), it is imperative to discuss about whether there are any physics based constraints that we can put on the generator so it does not get stuck, or that the gradients do not simply go to zero. Can we glean any physical metrics from the distance matrix representation so that we can modify the error terms for the VAE-GAN, to use physics-based terms as regularizers and constraints rather than simply employing more complicated mathematical metrics? The distance matrix, that

calculates the root squared difference between (x, y, z) coordinates from PDB files, represents the proteins/peptides as a matrix where each element's row and column index is the atom label, and the values are the distance of that atom from all other atoms in that row or column. This is feasible as a representation, and thereby widely used in protein structure prediction models, but cumbersome to store, and there are no known underlying physical metrics to be applied such as a term for steric clashes or conformal entropy to minimized, or any novel physics to be imparted or learned from such a representation, without other mathematical representations of the molecular graph. There are no clear mathematical benefits or differences between using the distance matrix, or the (x, y, z) coordinate values, or some other abstraction of the protein structure. The value of the representation must therefore come from its ability to interpret or incorporate underlying physics of the system.

In summary, the take away message is that without understanding the properties of the distance maps and graphical representations of the molecules and establishing metrics that are ubiquitous in graphs that can be used to constrain the generator, the VAE-GAN approach will not work. That is the rationale for pausing on the VAE-GAN model, and changing gears towards the random matrix theory formalism in the molecular graph representation context, that forms the bulk of the current study, as described in the chapters 3, 4 & 5.

Chapter 3

RANDOM MEAN FIELD MATRICES AND RANDOM BAND MATRICES

A random matrix is a random variable that takes matrix values, i.e., a matrix in which all elements are random variables.[510] Several crucial physical phenomena are partial differential equations construed as matrix and eigenvalue problems. For example, Eugene Wigner introduced the field of random mean field matrices[511] to study nuclear energy states and theoretically probe the nuclei of heavy atoms. Mean field random matrices (all elements are non-zero) now ubiquitously model the behavior of large and disordered Hamiltonians, an energy operator in the time-dependent Schrödinger equation which for one particle takes the form:

$$\hat{\mathbf{H}}\psi(\mathbf{r},t) = \mathbb{E}\psi(\mathbf{r},t) \tag{8}$$

where $\mathbb{E}$ are the eigenvalues and $\psi(\mathbf{r},t)$ are the orthonormal eigenvectors (in this case eigenfunctions) of the Hermitian operator $\hat{\mathbf{H}} = \hat{\mathbf{E}}_{kinetic} + \hat{\mathbf{E}}_{potential}$.
$\hat{\mathbf{E}}_{kinetic}$ is the kinetic energy operator for the particles in three Euclidean dimensions (x, y, z). For a particle of mass m, it is given by:

$$\begin{aligned}\hat{\mathbf{E}}_{kinetic} &= \frac{\hat{\mathbf{p}}\cdot\hat{\mathbf{p}}}{2m} \\ &= -\frac{\hbar^2}{2m}\nabla^2\end{aligned} \tag{3.1}$$

where $\hat{\mathbf{p}}$ is the momentum operator for one particle, $\hbar$ is the Planck's constant divided by 2π, and the ∇^2 is the Laplacian operator (partial differentials in three Euclidean spatial dimensions) for one particle given by:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{8}$$

Conversely, $\hat{\mathbf{E}}_{potential}$ is denoted as $\mathbf{V}(\mathbf{r}, t)$ and does not have a well defined form, instead depends on the system in which the quantum particle exists.

For N-body systems, such a Schrödinger equation is not analytically solvable, and the potential energy terms have to be guessed by the researchers with expert knowledge of the system they are trying to study. Unlike the kinetic energy term, the potential energy depends on the spatial arrangement of the particles and cannot simply be denoted as a linear combination of individual potential energies. The Hamiltonian for an N body system of non-interacting particles as given by the following equation makes the difference clear:

$$\begin{aligned}\hat{\mathbf{H}} &= \sum_{n=1}^{N} \hat{\mathbf{E}}_{kinetic} + \hat{\mathbf{E}}_{potential} \\ &= -\frac{\hbar^2}{2} \sum_{n=1}^{N} \frac{\nabla^{\mathbf{2}}}{m} + \mathbf{V}(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N, t)\end{aligned} \tag{3.2}$$

Moreover, the kinetic energy term, although expressed as a linear sum of individual kinetic energy terms, also depends on the configuration of the N particles in spacetime to conserve total energy (it must compensate for the potential energy and vice versa). In the more realistic case, where the N particles do interact, or some of them interact in the very least, the simple linear combination of the Laplacian operators becomes moot. several Laplacian operators, each for a single particle, must be multiplied with the other interacting particles' Laplacian, leading to a convoluted kinetic energy operator, unsolvable analytically. The Hamiltonian for such systems may be written as:

$$\hat{\mathbf{H}} = -\frac{\hbar^2}{2} \sum_{n=1}^{N} \left(\prod_{n=1}^{N} \frac{\nabla^{\mathbf{2}}{}_n}{m_n} \right) + \mathbf{V}(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N, t) \tag{3.3}$$

making it matrix valued and of the Hermitian form because the eigenvalues denoting allowed energy levels are always real-valued, and the eigenvectors denoting

wavefunctions are always orthonormal.

3.1 Introduction to Random Matrices: Random Mean Field Matrices

Density Functional Theory (DFT), although cumbersome and time consuming and computationally expensive, has the appeal it does, because it is the most accurate estimator for the combined Hamiltonian of such interacting many body systems (Although many DFT approaches use non-interacting particle formalisms). Conversely, RMT can model statistical properties of the eigenstates of several many body systems without having to exactly calculate the individual kinetic or potential energy terms, and does not need expert knowledge to make a qualified guess about the kinetic and potential energy functionals. However, the usage of such RMT models have been limited to chaotic and isotropic systems so far, such as those in heavy nuclei, or plasma,[512] or in the cores of neutron stars.[513] Mesoscopic systems, especially where molecular shapes are important, and where the Born-Oppenheimer principle cannot be strictly enforced without straying from physical reality, RMT shows potential by modeling the large disordered Hamiltonian of the interacting quantum system.

In mathematics, A Hermitian matrix H is also known as a self-adjoint matrix because it is defined as $Adj(H) = H$ such that for each element in the i^{th} row and j^{th}column $H_{ij} = H^*_{ji}$ where the * denotes complex conjugate. It is defined as a complex valued square matrix that is equal to the conjugate transpose of itself. More generally, an $n \times n$ matrix H is Hermitian iff $\langle \mathbf{v}|H\mathbf{v}\rangle \in \mathcal{R} \ \forall \ \mathbf{v} \in \mathcal{C}$ where the symbol $\langle \cdot|\cdot \rangle$ denotes inner product, $\mathcal{R}$ is the set of all real numbers and $\mathcal{C}$ is the set of all complex numbers. For a real valued Hermitian H, since the complex part is zero, it should equal its transpose H^T, i.e., it is a square symmetric matrix $H : H_{ij} = H_{ji} \implies H = H^T$ with eigenvalues $\lambda \in \mathcal{R}$. All Hermitian matrices (including the RMT model of the Hamiltonian) are unitarily diagonalizable with real eigenvalues. Because eigenvalues $\lambda \in \mathcal{R} \ \forall$ square Hermitian matrices H, λ represent measurable real outcomes of the

operator H in the quantum physical context. If H is the energy operator then λ are the allowed measurable energy values, if H is the angular momentum operator then λ are the measurable angular momentum values, etc.

The most studied Random Matrix Theory (RMT) symmetry classes in the quantum physics context are the Gaussian ensembles[514]:

(a) The Gaussian Orthogonal Ensemble or $GOE(n)$ is the Gaussian measure (a generalized concept of functions) with density:

$$\frac{1}{Z_{GOE(n)}} e^{-\frac{n}{4} trH^2} \tag{3.4}$$

on the $n \times n$ real, symmetric Hermitian matrix space $H = (H_{ij})_{i,j=1}^{n}, H_{i,j} \in \mathcal{R}$. Here $Z_{GOE(n)}$ is the normalization constant to make the density (sum of all probabilities) equal to one. The distribution described here is invariant under orthogonal conjugation. GOE systems model energy Hamiltonians that have both rotational and time-reversal symmetry, such as fermions and nucleons.

(b) The Gaussian Unitary Ensemble or $GUE(n)$ is the Gaussian measure with density:

$$\frac{1}{Z_{GUE(n)}} e^{-\frac{n}{2} trH^2} \tag{3.5}$$

on the $n \times n$ complex symmetric Hermitian matrix space $H = (H_{ij})_{i,j=1}^{n}, H_{i,j} \in \mathcal{C}$. The distribution described here is invariant under unitary conjugation. GUE systems model energy Hamiltonians without time-reversal symmetry.

(c) The Gaussian Symplectic Ensemble or $GSE(n)$ is the Gaussian measure with density:

$$\frac{1}{Z_{GUE(n)}} e^{-ntrH^2} \tag{3.6}$$

on the $n \times n$ symmetric and square quaternionic matrix space $H = (H_{ij})_{i,j=1}^{n}, H_{i,j} \in \mathcal{Q}$. The distribution described here is invariant under symplectic conjugation. GSE systems model energy Hamiltonians with time-reversal symmetry but not rotational symmetry.

The joint PDF (Probability Density Function) for the eigenvalues $\lambda_H = [\lambda_1, \lambda_2, \lambda_3, ..., \lambda_n]$ of the random hermitian matrix H that belongs to one of the Gaussian Ensembles is given by:

$$jPDF(\lambda_H) = \frac{1}{Z_{\beta,n}} \prod_{k=1}^{n} e^{-\frac{\beta}{4}\lambda_k^2} \prod_{i<j} |\lambda_j - \lambda_i|^{\beta} \tag{3.7}$$

where $\beta = 1, 2, 4$ for GOE, GUE and GSE respectively. $Z_{\beta,n}$ is an explicitly computed normalization constant. For any system that belongs to the Gaussian Ensemble symmetry classes, the eigenvalues repel (as denoted by the difference term $|\lambda_j - \lambda_i|^{\beta}$) as the joint PDF has a zero of β-order when there are degenerate or coinciding eigenvalues. $\therefore \forall\ H \in GOE$ the probability $P|_{\lambda_j = \lambda_i} = 0$.
From the sequence of eigenvalues, rearranged in ascending order,$\lambda_1 (= \lambda_{min}) < \lambda_2 < ... < \lambda_m < \lambda_{m+1} < ...\lambda_n (= \lambda_{max})$, mathematicians define the normalized spacing variable $s = (\lambda_{m+1} - \lambda_m)/\langle s \rangle$, where $\langle s \rangle = \langle \lambda_{m+1} - \lambda_m \rangle$ is the mean spacing. The PDF of spacings is then approximately denoted by the equation:

$$p_{\beta}(s) \approx C_1(\beta) s^{\beta} e^{C_2(\beta) s^2} \tag{3.8}$$

for GOE, GUE, and GSE. The numerical constants $(C_1(\beta), C_2(\beta)) : \int_0^{\infty} p_{\beta}(s) ds = 1$, and $\langle s \rangle = \int_0^{\infty} s p_{\beta}(s) ds = 1$, for $\beta = 1\ \forall\ H \in\ GOE, \beta = 2\ \forall\ H \in\ GUE, \&\ \beta = 4\ \forall\ H \in GSE$, respectively.[510]

In the global regime of such matrices, the distributions of linear statistics of the form $N_{f,H} = n^{-1} trf(H)$ are of interest. The empirical spectral measure μ_H is given by:

$$\mu_H(A) = \frac{1}{n}(\lambda_H \in A) = N_{1_A,H}, \quad A \subseteq \mathcal{R} \tag{3.9}$$

The Limit:

$$\lim_{n\to\infty} \mu_H(A)$$

is a deterministic measure in most cases, usually a semicircle for Gaussian ensembles (Fig 1.7). The Cumulative Distribution Function (CDF):$\int \mu_H(A)dA$ is known as the integrated density of states $N(\lambda)$.[515] If such a CDF is differentiable, then the derivative describes the density of states $\rho(\lambda)$. General linear statistics of the form $N_{f,H} = n^{-1}\sum f(\lambda_j)$ create interest in the fluctuations about $\int f(\lambda)dN(\lambda)$. A central limit theorem

$$\frac{N_{f,H} - \int f(\lambda)dN(\lambda)}{\sigma_{f,n}} \xrightarrow{D} N(0,1) \tag{3.10}$$

exists for several random matrix classes.[516,517]

In the local regime, the spacing between consecutive eigenvalues $s = (\lambda_{m+1} - \lambda_m)/\langle s\rangle$ and joint distribution of eigenvalues $jPDF(\lambda_H)$ in an interval of length $c/n,\ c \in \mathcal{Z}^+$ are of interest. Bulk statistics pertain to such intervals in the middle of the support of the limiting measure, i.e., around the median eigenvalues, and edge statistics pertain to intervals near the boundary of the support, i.e., near λ_{max} and λ_{min}.

In bulk statistics, if we consider a constant $\lambda_0 \approx (\lambda_{max} + \lambda_{min})/2$, somewhere near the median eigenvalue, in the interior of the support of the limiting measure $N(\lambda)$, then a point process of the form

$$\Xi(\lambda_0) = \sum_j \delta\left(\cdot - n\rho(\lambda_0)(\lambda_j - \lambda_0)\right) \tag{3.11}$$

where $\lambda_j \in \lambda_H$ captures the statistical characteristics of the eigenvalues in the vicinity of λ_0. Such a point process becomes a determinantal point process with a sine kernel of the form

$$K(x,y) = \frac{sin\left(\pi(x-y)\right)}{\pi(x-y)} \tag{3.12}$$

for the GUE as $n \to \infty$. The universality principle states that the limit $\lim_{n\to\infty}\Xi(\lambda_0)$ depends only on the symmetry class the random matrix H belongs to and not on the specific model of the random matrices, the median eigenvalue, the λ_0, or the

underlying individual random variables for each element of H. Edge statistics for s, on the other hand, usually follow a Tracy-Widom distribution.[518]

Another type of random matrices are square Wishart matrices of the form $H_{n\times n} = XX^*$, where $X_{m\times n},\ (m \geq n)$ is a random matrix with i.i.d entries, and $X^* : X^*_{ij} = \bar{X_{ji}}$ is the conjugate transpose of X.[519] The method of constructing the matrix H is through the multiplication of X and X^*, which makes the entries of H dependent on one another. Therefore, Wishart matrices are essential to estimate covariance measures in multivariate statistics.[520] The Wishart distribution (the joint probability distribution of the eigenvalues of the Wishart matrices) is the conjugate prior to the inverse covariance matrix of a multivariate normal random vector in a Bayesian setting. Wishart matrices are usually used to model quantum systems with individual wavefunctions that are dependent on each other before they interact. *In the analyses in the current body of work, we do not use Wishart matrices.*

RMT's potential for modeling the Hamiltonian of mesoscopic many-body systems with extensive interaction between the particles, that are usually modeled as Brownian classical systems, is limited to isotropic quantum chaotic systems such as the free electron cloud in a metal of high conductivity. The application of RMT to especially investigate what happens when electrons are not delocalized throughout the entirety of the finite space, such as in insulators and semi conductors, gave rise to the non mean field case, the Random Band Matrix.

3.2 Random Band Matrices in the Delocalized Phase

Random Band Matrices (RBM) are symmetric square matrices that have non zero elements in rows and columns close to the leading diagonal such that the number of rows or columns with nonzero elements right near the leading diagonal denotes the width of the band W. They have been used to model quantum chaotic systems such as free electron gases in metals.[426,427,521,522] Recently, it was proven that there is a strong relationship between the width of the band around the diagonal, and the shape of the

eigenvalue distribution in the support of the limiting empirical spectral measure, and that there is a sharp transition from a Poisson point process to a Gaussian Orthogonal Ensemble type semicircle.[425] According to P. Bourgade *et. al*,[425] if we consider an $N \times N$ symmetric RBM (the band is one dimensional, its width can only change in one direction, i.e., electrons in the system being modeled can only delocalize in one spatial direction) with general ($\sim$ i.i.d $\mathcal{N}(0,1)$ distribution of the entries and the band width $W \geq N^{3/4\ +\varepsilon}$ for any $\varepsilon > 0$, then in the bulk of the spectrum (near the median eigenvalues) and as $N \to \infty$ (i.e., large N), the following results show up:

(i) The semicircle law holds up to the scale $N^{-1+\varepsilon}$ for any $\varepsilon > 0$.

(ii) The eigenvalues (λ_i) locally converge to the point process belonging to the Gaussian Orthogonal Ensemble.

(iii) All the eigenvectors are delocalized, i.e., their L^∞ norms are all simultaneously bounded by $N^{-\frac{1}{2}+\varepsilon}$(after normalization in L^2) with overwhelming probability, for any $\varepsilon > 0$.

(iv) Quantum unique ergodicity (QUE) holds in the sense that the local L^2 mass of eigenvectors becomes equidistributed with overwhelming probability.

In general, for $W >> \sqrt{N}$, delocalization, QUE and Gaussian Orthogonal Ensemble spectral statistics hold, and for $W << \sqrt{N}$, eigenstates (the superposed wavefunctions) are localized and the limiting spectral measure converges to a Poisson point process. In physics terms, when materials conduct, their electrons are delocalized throughout the material, and are therefore dependent on each other, correlated in some way. Such dependence leads to them not having the same exact energy, i.e., they exist in different non-degenerate energy levels. That is what leads to the semi circle distribution of eigenvalues (energy levels) for wide band RBMs. For insulators where electrons are localized and do not correlate with one another, their energy

values are also independent and there is considerable overlap among their energies, leading to a Poisson point process distribution of the eigenvalues of narrow band RBMs, as shown in the fig 3.1 below.

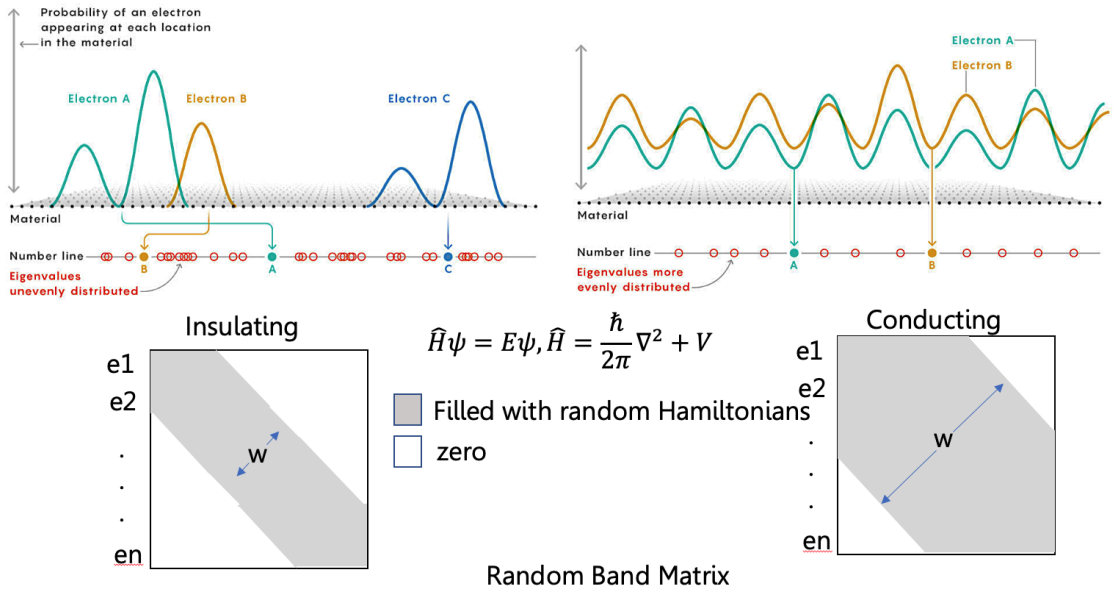


Figure 3.1: **Eigenvalue statistics and non-locality of Eigenvectors in a Random Band Matrix**: The figure depicts a one-dimensional random band matrix (RBM), it's relationship to the Schrödinger equation, the system that it models, the shapes of the eigenvectors of such RBMs and how they relate to the conductivity of a system. Non locality of eigenvectors in wide-band RBMs depicts that electrons have delocalized throughout the available space uniformly, i.e., it satisfies quantum unique ergodicity. It is also related to higher conductivity in such a system. The Eigenvalue single point correlation function for such a case follows the Semicircle rule, and falls under the Gaussian Orthogonal Ensemble (GOE) symmetry class of random matrices. In the case of narrow-band RBM models systems with localized eigenvectors, the model system is insulating, and the Eigenvalue single point correlation function follows a Poisson point process.*Figure adapted from https://www.quantamagazine.org/universal-pattern-explains-why-materials-conduct-20190506/*

It is well known that proteins by virtue of being organic molecules are themselves not conductive. However, electron transport is known to occur in several biological

processes. What remains vague is that how does this transport occur if proteins are not conductive. Soliton (single crested waves) models have previously been proposed but no proof was found.[523] It is also unclear if the 'same' electron actually moves through the protein structures, or if the potential energy moves from one end to the other, as in a series of pendulums hung together and barely touching, knocking out an electron at the other end. The latter explanation, of energy moving instead of the actual electrons, is more plausible, because proteins are not conductive. However, for such a phenomenon to occur, the electrons must somehow be entangled together, or at least their wavefunctions must interact and/or overlap, from one end of the protein to another. The question then arises, what kind of an entanglement exists and how to model it in terms of energy levels and wavefunctions (eigenvalues and eigenvectors) of the energy operator Hamiltonian.

Another well timed evidence comes to light that when proteins fold, they make stable low barrier hydrogen bonds (LBHBs) with themselves, and also low barrier but unstable hydrogen bonds with surface water.[461] Whether the folding actually happens or not depends on the sequence; whether the sequence allows for extensive LBHBs or just a few, leads to stable structures or disordered random coils respectively. α-helices are way more 'conductive' than β-structures, which are way more conductive than random coils, with the posited reason being the higher number of hydrogen bonds in each.[108] The conductivity is anisotropic. It is higher in the direction of the hydrogen bonds, and in both alpha helices and beta structures, two of the most common structures, accounting for almost 70% of protein structures, the hydrogen bonds are LBHBs.[460] It is not too much of a stretch to therefore imagine that folding into stable structures via extensive networks of LBHBs, leads to electrons farther along the sequence or the backbone to become 'entangled' via intermediary electrons delocalizing along the LBHBs. Modeling a Hamiltonian for such electron interactions, despite the electrons not being delocalized as much as in a metallic conductor, must also follow the rules laid out by RBM band width analysis. **The author in this current body**

of work therefore hypothesizes that:

- If a protein folds into a stable structure, far away electrons become dependent on one another due to individual wavefunctions overlapping in space over ergodic timelines via LBHBs or due to intermediate wavefunctions overlapping by delocalizing over shorter distances, then the eigenvalues of such a randomized Hamiltonian must repel each other, i.e., the probability of them coinciding goes down.

- Additionally, if a stable structure does not form and the proteins stay in an extended structure, then there isn't many avenues for the electrons in the protein to interact with one another, rather they interact with the solvent molecules, delocalizing along single-well hydrogen bonds in water.In such a situation, it is highly likely that the electrons of the protein are not dependent on each other, but rather with the solvent, where the LBHBs are unstable, thereby reducing delocalization of the protein's electrons over ergodic timelines.

- The random Hamiltonian matrix for such a case would obey the Poisson distribution over its eigenvalues, i.e., there is lots of degeneracy.

- From an information theoretic point of view, as a protein folds and settles into a stable structure, the global probability distribution over the eigenvalues (allowed discrete energy levels) of the random Hamiltonian representing the protein's electron interaction energies (both potential and kinetic together) becomes wider as the eigenvalues don't crowd around the same value on the number line. The uncertainty of the energy values increases.It becomes more difficult to say, what energy would an electron picked at random from the protein, would most likely possess. In other words, the Information entropy of the eigenvalues of the

Hamiltonian that models a protein structure, must increase when the protein folds.

- Conversely, if the protein does not have a stable structure, but rather remains in a disordered state, the uncertainty in its energy values is not as high. The information entropy of the eigenvalues of the Hamiltonian that models a disordered protein, therefore, is low.

The question arises, then, do these hypothesized phenomena actually take place? Can a random band matrix, optimized for systems where electrons are either totally localized, or entirely delocalized, be able to model a system where electrons interact only along certain directions? If yes, then in the absence of rigorous proofs that the band width rule applies to higher then 2 dimensions (researchers have not performed simulations or written rigorous proofs for 3 or more dimensions of quantum chaos modeled as a random band matrix), and considering that proteins are chiral anisotropic molecules, what type of random matrix would be most suitable? The random graph structured matrices (RGMs) described in the next chapter, come to the rescue.

Chapter 4

RANDOM GRAPH STRUCTURED MATRICES IN PROTEIN FOLDING

4.1 Random Graph Structured Matrices in the Delocalized Phase

Previous iterations of the Random Graph structured Matrix (RGM) in the literature include random connectivity between nodes equally distributed in several different finite spaces, including a 3-dimensional Euclidean torus.[524,525] The presence or absence of edges is tuned by a cut-off threshold ε on the Euclidean-distance measure. Nodes are are said to have edges between them if the distance between the nodes $d < \varepsilon$. It has been shown that as the dimension of the random matrix $n \to \infty$, then the limiting density of the single point correlation function of the eigenvalues of the RGM approaches the semicircle distribution, under Gaussian Orthogonal Ensemble Symmetry as ε increases.[524] The difference between the limiting density of the single point correlation function of the eigenvalues and the instantaneous density of the same for the RGM is usually estimated as a KL-Divergence. **Since it can be said that since such an RGM can be obtained by a random combination of row and column transformations of a Random Band Matrix (RBM) of the same dimensions with the same number of nonzero elements, the current body of work hypothesizes that the same statistics as RBM, including universality of Gaussian Orthogonal Ensembles and Quantum Unique Ergodicity (QUE), hold for RGM.**

To apply RGM formalism to protein structures, the current work assumes that the proteins can be thought of as nodes (atoms) in a 3-dimensional Euclidean space, connected by edges (bonds). However, since RGM are suitable to simulate large

disordered Hamiltonians that are electron energy operators (or other fermion energy operators; could also be momentum operators), and since the bonds have a one to one correspondence between electrons rather than atoms (atoms can have single, double or triple bonds, mandating a weighted adjacency matrix that throws off the requisite randomness and symmetry in Random Matrix Theory formalism), we further fine-grain the mental image and imagine proteins (and other molecules) as a bunch of mostly localized valence electrons (imagined as particles) in 3-dimensional Euclidean space, forming edges with the valence electrons of the other atoms within a threshold distance $\varepsilon \in \mathcal{R}^+$ which is the radius of the sphere in the finite cubic Euclidean space, within which all points interact to form mostly Low Barrier Hydrogen Bonds (LBHB) and $\pi - \pi$ stacking, and some other minor variations of weak secondary bonds.

For the case of numerical simulations, at first, we randomly assign (x, y, z) coordinates to n nodes, where n is an arbitrarily large positive integer (usually $n > 100$). The coordinates are identically and independently sampled from a distribution. For a uniformly distributed point cloud in 3-dimensions, we assume n points in 3-dimensional finite cubic Euclidean space of volume l^3 with coordinates $(x, y, z)\ \forall\ x, y, z \sim U(\{-(l/2), ..., (l/2)\})$ $i.i.d$, where $U(\{-(l/2), ..., (l/2)\})$ is the uniform distribution over the set $[-l/2,\ l/2]\ \forall\ l \in \mathcal{R}$. **The adjacency matrix thus obtained can now be imposed as a mask on a mean field random matrix to obtain a random graph-structured matrix, RGM**. An adjacency matrix $A_{n\times n}(\varepsilon) : A_{ij}(\varepsilon) = 0\ \forall\ D_{ij} > \varepsilon$ and $A_{ij}(\varepsilon) = 1\ \forall\ D_{ij} < \varepsilon$ where the pairwise Euclidean distance between the n points is given by the distance matrix $D_{n\times n} : D_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$. The following paragraphs describe the new math developed in the current body of work for such RGMs, as applicable to the protein's electron adjacency.

Now, let G be an $n \times n$ real-valued square matrix such that $G_{ij} \sim \mathcal{N}(0, 1)$ $i.i.d.$ with the mean of the standard normal distribution $\mu = 0$ and standard deviation $\sigma = 1$. Then let H be a square symmetric real-valued Hermitian matrix such that $H = (G + G^T)/\sqrt{2n}$, normalized to belong to the mean-field case of a random matrix

belonging to the Gaussian Orthogonal Ensemble (GOE) symmetry class. If the adjacency matrix $A(\varepsilon)$, derived from the graph with n nodes and m edges, then $A_{n\times n}(\varepsilon)$ is a symmetric matrix, with m non-zero elements either in the lower or upper triangle. We define the random matrix with imposed graph structured as a Hadamard (element wise) product between the adjacency $A(\varepsilon)$ and the symmetric matrix $H_{n\times n}$ as obtained above: $H \circ A(\varepsilon)$ ($A(\varepsilon)$ acts as a mask on H). since all the individual elements (Gaussian random variables in their own right) are independent of each other the joint probability density (jPDF) of such an RGM is given by the product of all the non zero random variables in $H \circ A(\varepsilon)$. If we would have considered the Wishart class of random matrices with $H = GG^*$, $G^*_{ij} = \bar{G_{ji}}$, then we wouldn't be able to do a simple multiplication to find the jPDF$(H \circ A(\varepsilon))$ as the elements are no longer independent of one another. However, as $n \to \infty$, the Wishart matrices are known to transition smoothly into GOE. Moreover, if $\lambda_{H\circ A(\varepsilon)}$ denotes the set of the ordered eigenvalues of the matrix $H \circ A(\varepsilon)$, then $\sum_i^n (\lambda^i_{H\circ A(\varepsilon)})^2 = Tr(H \circ A(\varepsilon))^2$. Therefore in the global regime, the single point correlation function of $\lambda_{H\circ A(\varepsilon)}$ can be estimated by the joint-PDF of the elements of the $n \times n$ dimensional RGM $H \circ A(\varepsilon)$ which is given by:

$$\begin{aligned} jPDF(H \circ A(\varepsilon)) &= \prod_{j\geq i, A_{ij}|_\varepsilon \neq 0} P((H \circ A(\varepsilon))_{ij}) \\ &= \prod_{j\geq i, A_{ij}|_\varepsilon \neq 0} C_1(A(\varepsilon)) e^{C_2(A(\varepsilon)) tr(H\circ A(\varepsilon))^2} \end{aligned} \tag{4.1}$$

where the constants $C_1(A(\varepsilon))$ and $C_2(A(\varepsilon))$ depend on the number of non-zero elements in the adjacency $A(\varepsilon)$. The values of $C_2(A(\varepsilon))$ can be obtained by calculating the number of nodes (electrons) n and the number of nonzero off-diagonal entries m in the upper triangle of $A(\varepsilon)$. for a protein analog, the number of nodes (electrons) n will of course vary for different protein sequences, but it stays the same as the protein folds. However, m will vary as the number of edges (bonds) in the protein changes as the protein folds, and for different protein sequences as well. Therefore, depending on

the values of n and m, the overall mean and standard deviation of the diagonal entries ($\mu_{diag} = 0, \sigma_{diag} = \sqrt{2/n}$) and off diagonal entries($\mu_{off-diag} = 0, \sigma_{off-diag} = 1/\sqrt{2n}$) of $H \circ A(\varepsilon)$ can be estimated and $C_2(A(\varepsilon))$ can be split into $C_2(n)$ and $C_2(m)$. One can then find $C_1(A(\varepsilon)) = C_1(n, m)$ from the properties of PDFs: the integral of the global eigenvalue PDF from equation 3.15 should be equal to 1. Therefore:

$$
\begin{aligned}
& C_1(A(\varepsilon)) \int_{-\infty}^{\infty} e^{C_2(A(\varepsilon))Tr(H\circ A(\varepsilon))^2} \prod_{j\geq i, A_{ij}|_\varepsilon \neq 0} dH_{ij} = 1 \\
\implies & C_1(A(\varepsilon)) = \left(\int_{-\infty}^{\infty} e^{C_2(A(\varepsilon))Tr(H\circ A(\varepsilon))^2} \prod_{j\geq i, A_{ij}|_\varepsilon \neq 0} dH_{ij} \right)^{-1} \\
\implies & C_1(n, m) = \left(\int_{-\infty}^{\infty} e^{C_2(n)Tr(H\circ A(\varepsilon))^2} \prod_{j=i} dH_{ii} \right)^{-1} \\
& \qquad \times \left(\int_{-\infty}^{\infty} e^{C_2(m)Tr(H\circ A(\varepsilon))^2} \prod_{j>i, A_{ij}|_\varepsilon \neq 0} dH_{ij} \right)^{-1}
\end{aligned} \tag{4.2}
$$

where the integral is the product of integrals over each individual element on the diagonal and the off-diagonal upper(or lower) triangle. From the values of $\mu_{diag}, \sigma_{diag}, \mu_{off-diag}$, & $\sigma_{off-diag}$, and by considering n and m as variables, we obtain:

$$C_1(n, m) = \left(\frac{1}{2}\right)^n \left(\sqrt{\frac{n}{\pi}}\right)^{n+m} \tag{4.3}$$

$$C_2(n) = -\frac{n^2}{4} \tag{4.4}$$

$$C_2(m) = -mn \tag{4.5}$$

In numerical simulations, we can count n and $m = m(\varepsilon)$ explicitly, and use the well-known normal distribution formula for a random variable x:

$$\mathcal{N}(\mu, \sigma) = \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Then we plug in the values of μ and σ for diagonal and off diagonal (lower or upper triangle) entries, and ultimately calculate the values of $C_1(n, m), C_2(n), \& C_2(m)$. We

can also simply estimate the probability density directly from the numerically sampled elements of the RGM in practice.

In the global regime, the single point correlation function of $\lambda_{H\circ A(\varepsilon)}$ with fixed but arbitrarily large dimension $n > 100$ approaches GOE's limiting density (semicircle) as $m(\varepsilon)$ increases with ε. To figure out if the information entropy of the single point correlation function (which is a probability distribution) of $\lambda_{H\circ A(\varepsilon)}$ increases with increasing $m(\varepsilon)$, we use the formula for continuous differential entropy of a probability distribution P given by:

$$\begin{aligned} CDE(P) &= -\int P\ln(P) \\ CDE(jPDF(H\circ A(\varepsilon))) &= -\int_{-\infty}^{\infty}\Big(C_1(n,m)e^{C_2(n,m)Tr(H\circ A(\varepsilon))^2} \\ &\quad\times\ln\Big(C_1(n,m)e^{C_2(n,m)Tr(H\circ A(\varepsilon))^2}\Big)\prod_{j\geq i, A_{ij}|_\varepsilon\neq 0} dH_{ij} \\ &= -C_1(n,m)\int_{-\infty}^{\infty}\big(C_2(n,m)Tr(H\circ A(\varepsilon))^2 \\ &\quad+\ln\left(C_1(n,m)\right)e^{C_2(n,m)Tr(H\circ A(\varepsilon))^2}\prod_{j\geq i, A_{ij}|_\varepsilon\neq 0} dH_{ij} \end{aligned} \tag{4.6}$$

By solving for the above, with $\mu_{diag}, \mu_{off-diag}$ and $\sigma_{diag}, \sigma_{off-diag}$ obtained from the transformation $H = (G+G^T)/\sqrt{2n}$ of each of the normal random variables $\mathcal{N}(0,1)$ for each element of G and G^T, and keeping n and m as variables (in the case of a protein, n will vary with protein sequence only, while m will vary with protein sequence, and the number of secondary bonds in a given protein's secondary structure as well, given a finite threshold cut-off for electron-electron interaction ε), we get:

$$\begin{aligned} CDE(jPDF(H\circ A(\varepsilon))) &= CDE(\varepsilon) = CDE(n,m) \\ &= \frac{m+n}{2}\left(1+\ln(n)-\ln(\pi)\right)+\frac{m}{2}\ln(2) \end{aligned} \tag{4.7}$$

for which a numerical value is obtained once we know n and m from $A(\varepsilon)$. The Information entropy can also be directly estimated from the numerical simulations by estimating the single point correlation function of $\lambda_{H \circ A(\varepsilon)}$ and then using the discrete version of the information entropy formula given in equation 1. The trends in $CDE(n, m)$ are visualized for m(with fixed n) and n(with fixed m) in fig 4.1.

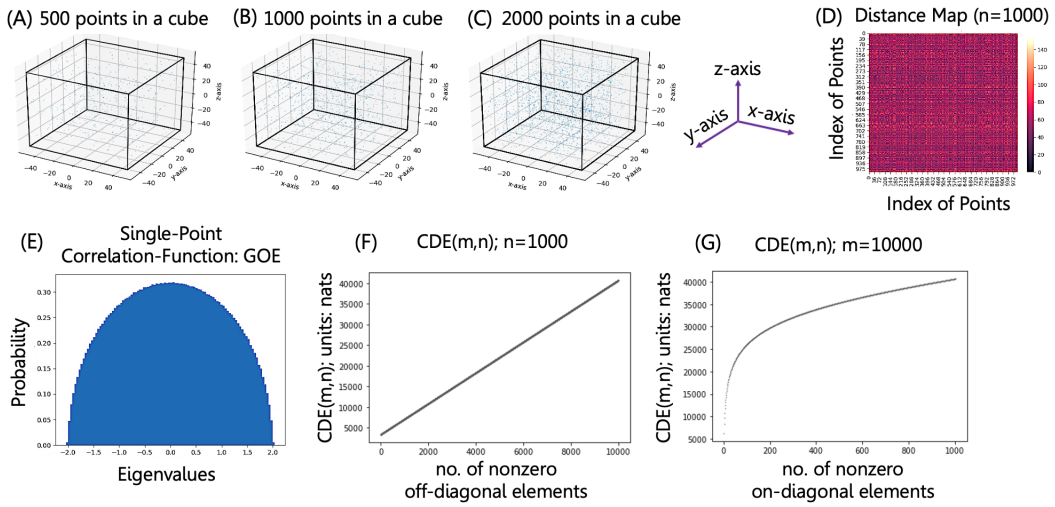


Figure 4.1: **Numerical Simulation Setup and Checks**: (A)–(C)Visual Representation of $n = 500, 1000$ and 2000 points, uniformly distributed along all 3 coordinates in a finite cubic Euclidean Space of volume 100^3; (D) The pairwise distance matrix for the $n = 1000$ case; (E) Single point correlation function of eigenvalues of a mean field matrix of dimension 1000×1000;(F) Trends in CDE(n,m) from equation 4.7 evaluated as $m \to \infty$ for fixed $n = 1000$; (G) trends in CDE(n,m) evaluated as $n \to \infty$ for fixed $m = 10000$

Next, we come up with an algorithm to numerically interrogate the properties of a Random Graph structured Matrix (RGM), as described below:

Algorithm:

- Sample n random points (nodes, electrons) uniformly in a cube (a torus or a sphere is possible as well). We chose the cube because protein simulations usually take place in a cubic unit cell. We also chose $n >> 100$, in a cube of

volume l^3, $l = 100$ in arbitrary units, origin at $(0, 0, 0)$ and $min/max = \pm 50$ for the x, y and z coordinates (See fig 4.1(A)-(C) for visualization).

- Calculate distance of every point from every other point, resulting in an $n \times n$ dimensional pairwise distance matrix D.

- Set a threshold value ε between 0 and the maximum possible distance (body diagonal of the cube of edge length l) which is $d_{max} = \sqrt{3l^2} \sim 173.20, l = 100$ arbitrary units. ε can be calculated by either partitioning d_{max} into equal length segments (equispaced) or by partitioning the distribution of the D_{ij} into equiprobable sections (See fig 4.1(D) and fig 4.2(A)-(D)).

- To the distance matrix D, Apply a threshold $\varepsilon : \forall\ D_{ij} < \varepsilon, A_{ij} = 1$ else $A_{ij} = 0$, where $A_{n\times n}(\varepsilon)$ is the adjacency matrix of the n nodes in the cube.

- Randomly sample n^2 values from $\mathcal{N}(0, 1) i.i.d.$ and reshape the array of numbers to obtain an $n \times n$ dimensional real-valued random asymmetric matrix G.

- Perform the transformation $H = (G + G^T)/\sqrt{2n}$ on the matrix G to obtain normalized symmetric random matrix with the mean $\mu_{diag} = 0$ and the standard deviation $\sigma_{diag} = \sqrt{2/n}$ for the diagonal elements, and the mean $\mu_{off-diag} = 0$ and the standard deviation $\sigma_{off-diag} = 1/\sqrt{2n}$ for the lower and upper triangular elements to satisfy the symmetry constraints for GOE. Since the standard deviation of the off diagonal elements is half that of diagonal elements, we can be rest assured that the random mean field matrix H belongs to the Gaussian orthogonal Ensemble (GOE), and that it has been properly normalized.

- Repeatedly sample more G matrices, so that we have N different H matrices per adjacency matrix $A(\varepsilon), \forall\ \varepsilon$, to probe the "certainty" of the metrics (standard

deviation about the expected value of the metrics) that we are trying to establish for any individual RGM corresponding to $A(\varepsilon)$.

- Calculate the eigenvalues λ_H^k of the mean field random matrices $H^k, k = 1, 2, ..., N$.

- Calculate the probability density function (as an array, by binning the values in bins of equal width as in a histogram) of $\lambda_H^k \ \forall \ H^k, \ k = 1, 2, ..., N$ together. That should be the limiting density case as $\lim H^k \circ A(\varepsilon) = H^k$ as $\varepsilon \to d_{max}$ (fig 4.1(E)).

- Apply the adjacency $A(\varepsilon)$ on all the matrices H^k by element-wise multiplication to obtain the RGMs of interest. Then find the eigenvalues $\lambda_{H \circ A(\varepsilon)}^k$ of the RGMs $H^k \circ A(\varepsilon), \ k = 1, 2, ..., N, \ \forall \ \varepsilon$, pool them together, and calculate their probability density function as a histogram, by equal sized binning, which is their single point correlation function, after normalization so that the area under the histogram equals 1 and by scaling so that the min-max range of the eigenvalues are comparable. Fig 4.3 shows the plot of all the single point correlation distribution for all the eigenvalues of the RGM for different values of n and ε.

- Calculate the CDE of the RGMs $H^k \circ A(\varepsilon), \ k = 1, 2, ..., N, \ \forall \ \varepsilon$ with adjacency $A(\varepsilon)$ at threshold ε, two ways: first, numerically by applying $CDE(m, n) = -\sum P \ln(P)$, where P is the numerically computed single point correlation function of the eigenvalues of the RGMs; and then from the analytical formula derived in equation 4.7. Fig 4.3 shows the plot of the trends in analytically obtained CDE of the jPDF of the RGMs from equation 4.7 and the numerically computed CDE of the single point correlation function of the eigenvalues of the RGMs for different values of n and ε.

- Obtain the median and median+1^{th} eigenvalues $(\lambda_{n/2}^k, \ \lambda_{n/2+1}^k)$ in the bulk of the support of all eigenvalues from the RGM $H^k \circ A(\varepsilon), \ k = 1, 2, ..., N, \ \forall \ \varepsilon$,

and obtain the distribution of the spacing between them. Usually, the median eigenvalues should fall in the vicinity of 0 ± 0.1 on the number line. Properly normalize the spacing values by dividing by the expected value of all consecutive eigenvalue spacings. As seen in fig 4.3, the median eigenvalue for RGM properly normalized is at 0.

- Find the mean and standard deviation of the bulk spacing distribution thus obtained $\forall\ k\ \&\ \varepsilon$.

- Find the mean and standard deviation of the distribution of all eigenvalue-pair spacings, not just in the bulk.

- Find the mean and standard deviation of the eigenvalue pair spacings at the edge of the support of the eigenvalues. Trends for all the means and standard deviations for eigenvalue pair spacings at the bulk and edge of the support of the eigenvalues for the RGMs are shown in fig 4.4, $n = 1000,\ \&\ \forall\ \varepsilon$.

- Using equation 6, calculate the KL-Divergence $D_{KL}(\rho(\lambda_{H^k \circ A(\varepsilon)})||\rho(\lambda_H)$, where $\rho(\cdot)$ is the probability distribution, $\lambda_{H^k \circ A(\varepsilon)}$ are the eigenvalues of the RGMs $H^k \circ A(\varepsilon),\ k = 1, 2, ..., N\ \forall\ \varepsilon$, and λ_H are the eigenvalues of the mean field random matrix H. The trends of this metric are shown in fig 4.4 for multiple values of n and ε

- Obtain eigenvectors of all the RGMs $H^k \circ A(\varepsilon)\ k = 1, 2, ..., N\ \forall\ \varepsilon$ and look for signatures of quantum unique ergodicity in the ratio of $||L^\infty||/||L^2||$ norms for each RGM. If the ratio is close to 1 then the eigenvectors are heavily localized, and if the ratio is close to $1/\sqrt{n}$ then the eigenvectors are heavily delocalized. The $||L^\infty||/||L^2||$ ratios of the Eigenvectors as ε values increases are shown in fig 4.4 and fig 4.5.

As seen in fig 4.1, the volume $l^3, l = 100$, of the cubic box is kept the same, but number of points increases, so we can use n as a proxy for density ($\rho = n/l^3$). We also see in fig 4.1 (F) and (G) that the $CDE(n, m)$ from equation 4.7 increases linearly with m at a fixed value of n, and increases logarithmically with n at a fixed value of m at first before becoming linear asymptotically around $n = 200$ For lower values of m, the logarithmic term in $CDE(n, m)$ from equation 4.7 will dominate so the asymptotic linearity of the $CDE(n, m)$ function will appear at greater values of n. Nevertheless, as expected, the increase in number of edges (secondary bonds in the case of a protein as it folds) increases the continuous differential entropy of the jPDF of all elements in the RGMs, which is analogous to the CDE of the single point correlation function of the eigenvalues of the RGMs.

In the protein folding scenario, the n and m translate to number of electrons and number of bonds (that increases while n remains constant during protein folding as long as the protein folds into a stable structure such as α-helix or β-sheets), the random entries of the RGM denote the random interaction energies between the electrons. It could also denote momentum, depending on the interpretation, but Eugene Wigner, as discussed before, established the correspondence of eigenvalue spacings with energy level spacings of fermions, therefore lending credibility to the idea that the Random matrices of the GOE type, are analogous to the Hamiltonian energy operator in Quantum Mechanics. Accordingly, the eigenvalue statistics inform us about the statistics of the allowed energy levels that the electrons can occupy in the protein or protein-water complex.

From fig 4.2(A),(C) it is clear that the equiprobable thresholds cluster around the mean of the pairwise distance values, and therefore add a similar number of off-diagonal entries m to the RGMs as the threshold increases, which is seen in fig 4.2(E). Similarly, we also see that equispaced thresholds (fig 4.2(B),(D)) that simply segment the maximum distance possible in the cube equally, add unequal number

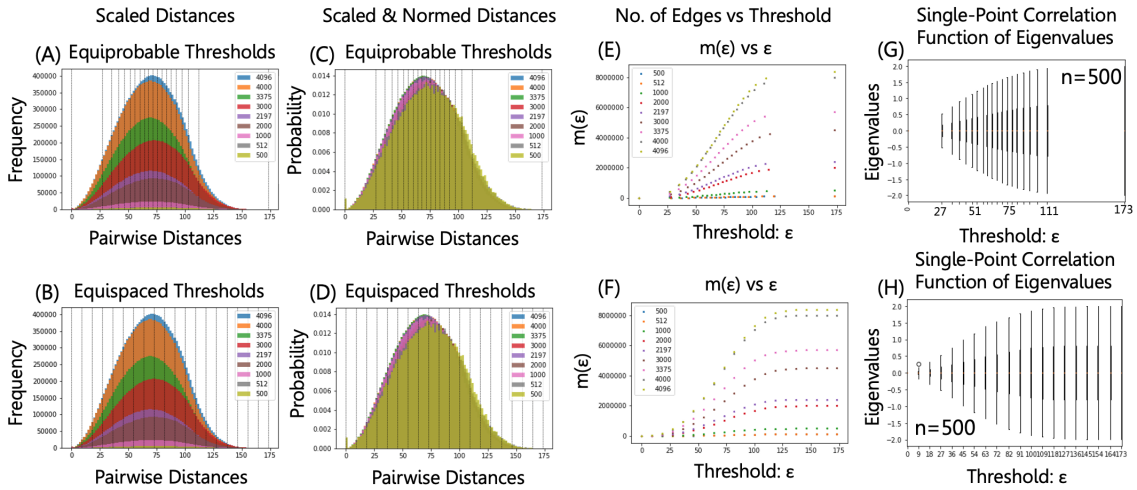


Figure 4.2: **Applying thresholds on Point Clouds**: (A)-(B) Scaled and binned pairwise distances with equiprobable and equispaced thresholds respectively. In the equiprobable thresholds case, each segment contains similar number of pairwise distances for each point cloud density n while for the equispaced case, the x axis is simply segmented into equal length sections; (C)-(D) Same as (A) and (B) respectively, except the pairwise distributions are normalized to show the probability of each distance value in the point cloud of density $n = 500, 512, 1000, 2000, 2197, 3000, 3375, 4000, 4096$ in a 100^3 cubic box;(E) Single point correlation function of eigenvalues of a mean field matrix of dimension 1000×1000;(F) Trends in CDE(n,m) from equation 4.7 evaluated as $m \to \infty$ for fixed $n = 1000$; (G) trends in CDE(n,m) evaluated as $n \to \infty$ for fixed $m = 10000$

of off-diagonal entries (fig 4.2(F)). For an example case of $n = 500$, we also see that the single point correlation functions of the eigenvalues of the RGMs show the expected trend of becoming wider as ε increases for both equiprobable (fig 4.2(G)) and equispaced (fig 4.2(H)) thresholds. The values of ε for equiprobable thresholds are $[0, 27, 35, 41, 46, 51, 55, 60, 63, 67, 71, 75, 79, 83, 87, 92, 97, 103, 111, 173]$ and the same for equispaced thresholds are $[0, 9, 18, 27, 36, 45, 54, 63, 72, 82, 100, 109, 118, 127, 136, 145, 154, 164, 173]$. However, due to the constraints of space and for clarity, in the next figure (fig 4.3) we do not place the boxplots depicting the single point correlation function of the eigenvalues at the positions mandated by the actual values of ε, but

at the index of the individual ε values (i.e., $1, 2, ..., 20$, where the first value $\varepsilon = 0$ has index 1, second value $\varepsilon = 9$(equispaced) or 27(equiprobable) has index 2, ..., and the last value $\varepsilon = 173$ has index 20).

From the point of view of protein folding, the CDE trends tell us that as the protein folds, the number of bonds it has with itself, increases, (reflected by increasing ε value), i.e., more and more electron wavefunctions overlap, and therefore get dependent on one another, leading to a reduction in degeneracy as no more than two electrons can occupy an energy state, and the electrons also have to have opposite spins. The same phenomenon is observed when covalent bonds are formed as well. The molecular orbitals are wider apart, and are of different energies than individual atomic orbitals with degeneracy.

From fig 4.3, we see that the trends for the single point correlation functions (shown as a series of boxplots) of eigenvalues of RGM with $n = 500, 1000, 2000, 4000$ as the value of ε increases (the plots only show the index of the ε values on the x-axis instead of actual values to avoid cluttering of the plots). In all the boxplots, the median eigenvalue is the mean eigenvalue, which is at 0, and is shown by the horizontal yellow line. As expected, the eigenvalues spread out more and the distribution flattens out further as $\varepsilon \to \sqrt{3l^2}$(max distance possible in a cube of volume l^3), until at $\varepsilon = 173$ the single point correlation function is a semicircle (as shown in fig4.1(E)), regardless of how the ε values were obtained (equispaced or equiprobable partitioning of the pairwise distance distribution).

Fig 4.3(I),(K), show the $CDE(n, m)$ calculated from actual values of n and m counted from the adjacency matrices $A(\varepsilon)$ computed at the threshold ε applied on the pairwise distances. As expected, the CDE(n,m) increases as $m \to \ n(n-1)/2 \ \forall \ n$. Similarly, when we obtain the single point correlation function($SPCF$) of the eigenvalues directly from the generated RGM as described in the algorithm above, and use the

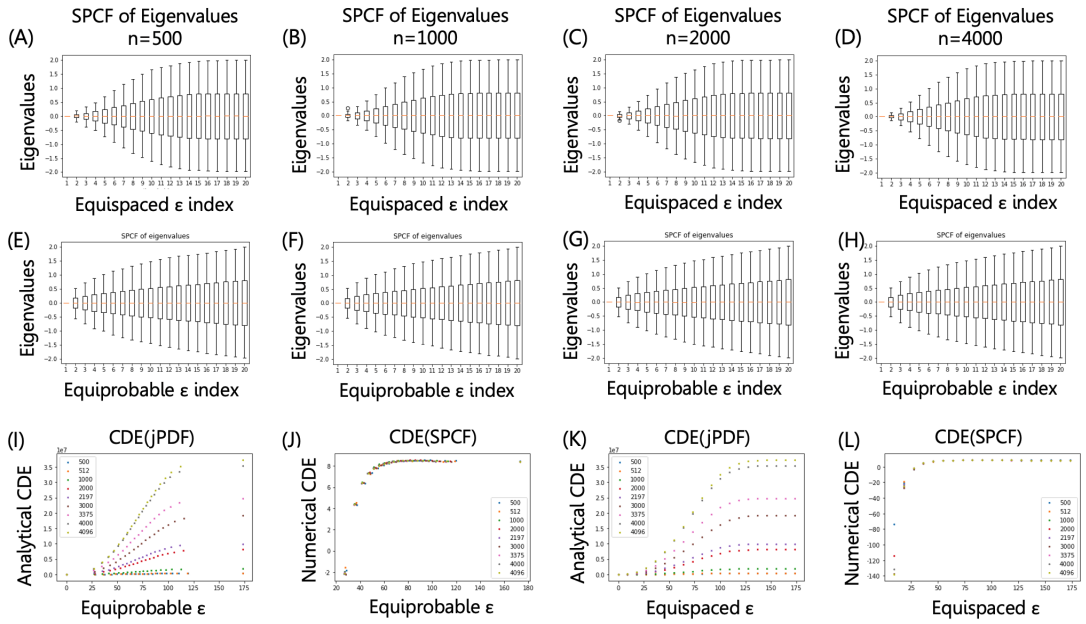


Figure 4.3: **Statistics of Single Point Correlation Function of Eigenvalues of RGM**: (A)–(D) The trends in the single point correlation function of the eigenvalues are shown for $n = 500, 1000, 2000$ and 4000 as ε increases; (E)-(H) Same as (A)-(D) respectively, except the ε values are from the equiprobable thresholding;(I) CDE(n,m) from equation 4.7 is evaluated and trends are shown $\forall n = 500, 512, 1000, 2000, 2197, 3000, 3375, 4000, 4096$, showing the trends in information entropy of the distribution of the joint PDF of all elements of the RGM as ε increases for equiprobable thresholds.(J) Trends in numerically computed CDE of the binned eigenvalue distribution (histograms) for all n as ε increases for equiprobable thresholds. (K),(L) same as (I) and (J) respectively but for equispaced thresholds.

equation

$$CDE(P) = -\sum P \ln P$$

to obtain numerical values of the CDE of the $SPCF$, and plot the trends for both equispaced and equiprobable thresholds(fig 4.3(J),(L)), we see that the CDE increases and then stabilizes as $m \rightarrow \ n(n-1)/2 \ \forall \ n$, similar to the analytical $CDE(m,n)$ derived on the $jPDF$ of the RGM entries themselves, showing that $CDE(jPDF(H \circ A(\varepsilon)) \sim CDE(SPCF \text{ of } \lambda_{H \circ A(\varepsilon)})$.

In the case of protein folding, the trends in the eigenvalue single point correlation function tells us that as number of bonds increases within the protein (more and more electrons start interacting, their wavefunctions overlapping, no longer independent), we expect to see a reduction in the degeneracy of allowed energy levels for the electrons. Since the CDE of the $SPCF$ also increases and then stabilizes, we can expect that there is a decrease in certainty (information entropy increases, so the information content, amount of surprise, has increased, but encrypted, according to the laws of information theory) among the single-point correlation function of the allowed energy levels of the electrons. The implication is that if we were to measure the energy of many electrons one by one in a folded protein, many energy levels will be measured, rather than a certain expected value. In the case of disordered proteins, however, the proteins do not fold into stable structures, so the information entropy of their energy-levels does not increase, and therefore if we measure the energy of the electrons, one by one, in such a protein, we would mostly measure a certain expected value of energy that most electrons will possess. Since ligands and other small molecules bind to proteins at the active sites, and the active sites mostly have the disordered regions, rather than fully α or β structures, it can be said that the approaching molecules make some measurement on the electrons present near the active site. The above discussion posits that since the expected value of electron energies (regardless of which electron it is) is very certain for the disordered regions than the folded regions, it is easier to make a measurement at the disordered regions and therefore bind to it. This can also serve as a template, and only the molecules that can match a similar expected energy level (frequency of wavefunctions) will bind to the disordered region of the protein-receptor. Moreover, as a ligand binds to the receptor on one end and settle into a stable structure momentarily, the other end of the receptor unfolds to present residue to other ligands to start the signaling cascade. As such, one part of the protein receptor loses certainty in expected energy, while the other end gains the certainty, showing a clear direction of information transfer,

probably mediated by the low barrier hydrogen bonds along the other fully-folded and more conductive regions of the protein.

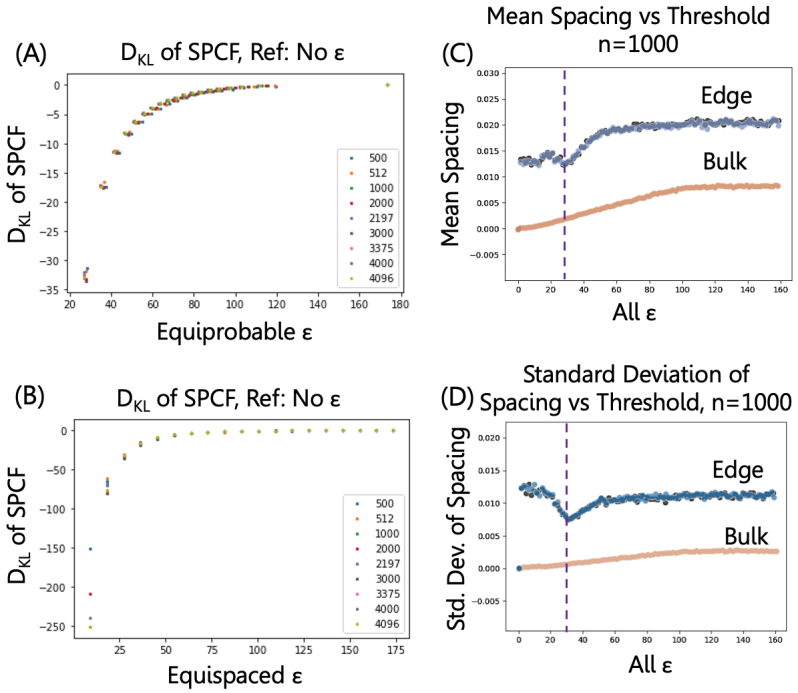


Figure 4.4: **KL-Divergence of Single Point Correlation Function of Eigenvalues and Statistics of their spacings**: (A)–(B) Trends in KL-Divergence D_{KL} as obtained by applying equation 6 on the $SPCF$ of Eigenvalues of RGMs, with the SPCF of the eigenvalues of the mean-field random matrix ($\varepsilon = 173$) as the reference distribution, for both equiprobable and equispaced ε values; (C) Mean value of spacing distributions $\forall\ \varepsilon,\ H^k \circ A(\varepsilon),\ k = 1, 2, ...1000,\ n = 1000$ calculated between the edge eigenvalues (largest and next largest; smallest and next smallest) and the bulk (median and median+$1th$) eigenvalues as described in the algorithm; (D) Standard deviation value of spacing distributions

Fig 4.4(A) and (B) show us the KL-Divergence, a measure (equation 6) of how different two distributions are, with the reference distribution being the single point correlation function of the eigenvalues, and how this KL-Divergence changes with increasing values of ε. As can be seen, as $\varepsilon \to \sqrt{3l^2}$ for $l = 100$, the KL-Divergence

smoothly reaches zero for both the equispaced and equiprobable threshold values, confirming that the eigenvalue single point correlation function tends towards GOE statistics smoothly. This was expected because the KL-Divergence is simply a more accurate measure of continuous entropy (a better analogue for Shannon's information entropy of discrete distributions) than the continuous differential entropy $CDE(m, n)$. For the Eigenvalue spacing distributions at the edge and the bulk(median and median+1 eigenvalues, which fall in the range 0 ± 0.1) of the support of eigenvalues, we see that the mean value of the spacing increases with ε at first and then stabilizes (fig 4.4(C)), confirming that not only is the $SPCF$ of eigenvalues becoming wider and more semicircular, but also that the eigenvalues are repelling each other more and more along the real number line, before maintaining a certain expected spacing between each other. However, the standard deviation of the spacing distributions fig 4.4(C)), especially for edge eigenvalue spacings, shows a strong minimum at a certain value of ε before increasing and stabilizing at a significantly higher value than the minimum. This observation is surprising, and suggests that at a certain level of connectivity among the points (pertaining to a certain threshold ε value, which is 27 for the example case of $n = 1000$ points), the eigenvalues not only repel each other more, but that we get more and more certain about the spacing that is maintained between the eigenvalues. Interestingly, this spacing that we are most certain of, isn't the maximum spacing that the eigenvalues ultimately maintain between themselves, if connectivity keeps increasing. We can therefore posit that only at the threshold where the eigenvalue spacings have the lowest standard deviation, and therefore highest certainty, the eigenvalues themselves hold their positions on the real number line, while still maintaining a non-zero spacing between them.

Extrapolating to the case of proteins, such an observation suggests that depending on density of electrons in 3-D space, as the electrons gain access to more energy levels when they overlap (as the protein folds, for example), the separation between energy levels increases, making it more difficult for electrons to gain or lose energy. However,

at a certain distance of electron-electron interaction, there comes a time when the allowed energy levels not only maintain a spacing between them, but also that the spacing becomes more certain, suggesting that the energy levels themselves become more strongly preferred, indicating the presence of longer time coherence (As long as we consider the RGM to represent the Hamiltonian energy operator) of the quantum nature of electrons. It is worth it to investigate what this certain threshold distance is that could be responsible for maintaining longer time coherence of quantum properties of electrons. However, if we consider the RGM to represent the momentum operator instead, the minimum spacing standard deviation at the certain threshold value indicates the distance at which the electron wavefunctions or eigenstates are most highly delocalized. Since the electron wavefunctions are the eigenfunctions of the momentum operator, which are the eigenvectors of the RGM, we interrogate the delocalization of the eigenvectors of the RGM in fig 4.5. In fig 4.5, we can clearly see from the ratio

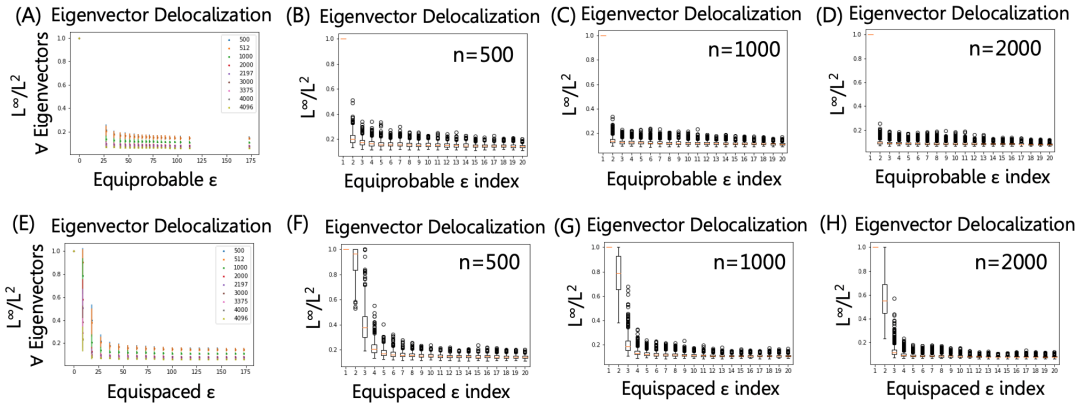


Figure 4.5: **Delocalization of Eigenvectors of RGM**: (A) Box-plots showing the distribution of $||L^\infty||/||L^2||$ norm ratio for all eigenvectors of all RGM matrices at different values of n and equiprobable thresholds ε; (B-(D)) same as (A) but only for the $n = 500, 1000, 2000$ cases respectively. the x-axes in)B)-(D) show the threshold indices rather than the actual threshold value for clarity in viewing the boxplots; (E)-(H)Same as (A)-(D) but for equispaced thresholds.

of $||L^\infty||/||L^2||$ norms for each eigenvector, corresponding to each eigenvalue of each

RGM $H^k \circ A(\varepsilon)\ \forall\ n$, and further visualized more clearly in fig 4.5(B)-(D), (F)-(H) for equiprobable and equispaced thresholds for $n = 500, 1000, 2000$ cases only. The $||L^\infty||$ norm picks out the maximum value in an array, while the $||L^2||$ norm is the root sum-squared value of all the values in an array. So for an array of size n, the norm ratio will be 1 if the maximum value dominates all the other values combined, indicating localization, whereas if the ratio approaches $1/\sqrt{n}$ then the maximum value is no more or less dominant than any other value in the array, indicating total delocalization. We can clearly see that the Eigenvectors start out very highly localized at $\varepsilon = 0$, because the norm ratio is 1, and then as ε increases, they approach their respective $1/\sqrt{n}$ values, which is $0.044, 0.032, 0.023$ for $n = 500, 1000, 2000$ respectively. The observation indicates Quantum Unique Ergodicity in the RGM pretty early on in the degree of connectivity, depending on n. In fact, the rate at which the norm ratio drops to stable low values is dependent on the point density n as we see that for $n = 500$, the ratio drops below 0.2 at $\varepsilon \sim 27$, while for $n = 1000$, it happens at $\varepsilon \sim 18$, and for $n = 2000$, it happens at $\varepsilon \sim 9$, seemingly indicating a direct relationship with the threshold at which we observe the minimum for Eigenvalue spacing standard deviations.

The implications for the protein folding scenario could be that at a given density of electrons, the distance threshold below which all the individual electron wavefunctions are interacting, via either covalent or secondary bonds, leads to a very high degree of delocalization of superposed wavefunctions (similar in principle to the mathematical graph percolation effect), accompanied with a high degree of certainty in the spacing between allowed energy levels (or momenta) and therefore the energy levels themselves, displaying a possible long term quantum coherence, as well as a high degree of certainty in expected value of measured energy of electrons if the proteins remain disordered, and high uncertainty in measured energy of electrons if the proteins fold into stable structures. At this particular distance-based interaction threshold, quantum properties are maintained, and we can begin to tease out the secrets of molecular

recognition and signal transduction from the point of view of electrons. We must therefore interrogate what such an ideal threshold could be for the protein folding case, as well as optimize the following metrics to dynamically evolve graph connectivity:

(a) Maximize the standard deviation of $SPCF$ of eigenvalues w.r.t. $m(\varepsilon)$

(b) Maximize the $CDE(n, m)$ of the RGM and the $CDE(SPCF)$ of the eigenvalues w.r.t. $m(\varepsilon)$

(c) Maximize the expected (mean) spacing between consecutive eigenvalues w.r.t. $m(\varepsilon)$

(d) Minimize the eigenvalue spacing standard deviation w.r.t. $m(\varepsilon)$

However, before one can use the RGM metrics identified above in a data driven model (such as the previously implemented VAE-GAN as described in chapter 2), we establish the veracity that indeed the information entropy of the eigenvalue single point correlation distribution measure of the RGM (that models the large disordered Hamiltonian of the valence electrons in a protein) increases as the protein folds (rather than deliberately increase m as we did before by simply increasing ε). Since the limiting density of the single point correlation function of the eigenvalues is different for different proteins (no RGM will ever reach the mean field connectivity, but rather the connectivity, and therefore the number of nonzero off diagonal elements, will depend on the specific protein sequence and its structure) and is unknowable if the protein structure is unknown (which makes the whole point of the study moot), there must exist other parameters or other metrics of the eigenvalue statistics, such as the standard deviation of the edge and bulk eigenvalue spacing distributions as detailed above, and non locality of eigenvectors of random graph matrices, that can act as an opposing phenomenon to the asymptotically linearly increasing CDE. The following

section describes the algorithm that we adopted to investigate whether CDE (of the eigenvalue single point correlation distribution of an RGM) increases as a model protein folds in an aqueous medium, as simulated by the widely accepted Molecular Dynamics (MD) simulation (a purely classical Brownian motion model of protein folding), and a control system (pure water) MD simulation. We choose Polyalanine with 21 Alanine repeating monomers as our protein. It is considered a model sequence, because it is very well established in the literature that the protein folds into a stable α-helix structure. We also test other statistics of RGM in the Polyalanine-21 and pure water model systems to tease out other fundamental physical phenomena and see if the numerically established metrics have physical relevance during protein folding in aqueous media. We are unable to use any outside MD simulations because the trajectories are not comparable due to different force fields and different simulation parameters as well as the fact that most raw trajectories are not completely uploaded to publicly accessible repositories.

4.2 Approach & Methods to model protein Folding with Random Graph Structured Matrices

After establishing some metrics such as the information entropy of the single point correlation distribution measure of the RGM type Hamiltonian of electron interaction in proteins during folding, the local spacing distribution and its properties in the bulk of the support of the spectrum of the RGM, and the overall spacing of the spectrum, the next step is to come up with a process to use such formulae and abstracted results on protein folding simulations that are well trusted in the literature, to tease out the physical relevance of the established metrics.

The step by step process is outlined below (see fig 4.6 for a visual depiction):

Algorithm:

(a) The first step is to find a model protein that takes a well-known stable secondary structure in water. One such protein is Polyalanine-21, where the sequence is

composed of just a string of 21 Alanine residues. Polyalanine has a very stable α-helical structure in solution,[115] and Brownian motion models converge on the structure in a reasonable amount of time. We chose to go with 21 repeats of Alanine, to ensure we were operating in the large degrees of freedom (large n) regime for the random matrices. Such a 21 length sequence is termed a peptide instead of a protein.

(b) The next step was to ascertain a system that has no significant change in its connectivity during a Brownian model based simulation, as a 'control' system. Water is a very obvious choice as it is also the solvent in which Polyalanine-21 is suspended.

(c) To model the Brownian dynamics of Polyalanine-21 folding in water, and the dynamics of a system of water molecules themselves, we use the CHARMM-27 force field, with default parameters, Van-der-Waals cut off distance at 10Å, the TIP3P water model, particle mesh ewald summation, Limited-memory Broyden-Fletcher-Glodfarb-Shanno (LBFGS) minimization (stores only the major dynamic modes of the inverse Hessian matrix representing the gradients) to minimize the potential energy in an NVE ensemble class (constant number of particles, constant volume, constant free energy), until convergence of the simulation. It took about 30000 time steps, spanning $300ns$ to converge the peptide folding simulation of Polyalanine-21. The water simulation used explicit solvent molecules placed inside the periodic simulation box by the MD algorithm (GROMACS) in such a way as to minimize steric clashes. All the MD simulation itself did was evolve the system over time. The system already began with as little potential energy as possible, so no major changes in hydrogen bond numbers, or structuring of water is expected without the involvement of the protein.

The system was simulated for 2500 time steps spanning $25ns$. Because classical molecular dynamics only implicitly use quantum mechanics in the parameters of their Newtonian force fields, and do not explicitly include electronic behavior, the success of RGMs in showing statistics of electron dynamics that arise from such a classical trajectory would be attributable to the nature of quantum mechanical interactions, and not to any artifact of molecular dynamics simulations. RGM would also provide a simpler way to analyze MD trajectories, as well as to tease out fundamental physics from a meso-scale Brownian dynamical model.

(d) The trajectories of both simulations in spacetime, contain as many frames as time-steps, obtained as a series of files in the PDB format, listing their (x, y, z) coordinates, with the Nitrogen at the N–terminus being the origin with coordinates (0,0,0). We first convert the trajectories into an all atom distance matrix D_{ij} as before. The row and column indices of distance matrices represent atom indices. We start numbering from the hydrogen on the N–terminus of Polyalanine-21, all the way until the last hydrogen on the C–terminus. After we get to the C_α atom, we label the next few indices with atoms from the residue attached to the C_α atom, then move to the next carboxyl group and so on. We do not include solvent water in the peptide adjacency but it can be easily incorporated in future versions. For the water simulation, we start the indexing of the distance matrix at the first atom at an arbitrary corner of the simulation box (we keep the indices attached to the same atoms throughout the simulation runs for consistency).

(e) In the next step, we apply a distance cut off ε to obtain the adjacency matrix A in terms of 0 and 1. For distances in 3-D space $d \leq \varepsilon$, where $\varepsilon \, \epsilon \, \{1\mathbf{\mathring{A}}, 2\mathbf{\mathring{A}}, 3\mathbf{\mathring{A}}, ..., 10\mathbf{\mathring{A}}\}$. we repeat the following steps for each value of ε.

(f) For each adjacency matrix A (corresponding to each time step and each ε), which is an atom to atom adjacency, indexed as described in part(d) above, same as the distance matrix, we want to obtain an electronic adjacency. To obtain such an adjacency, we repeat the row and column representing each atom in the peptide, or water, as many times consecutively as the number of their valence electrons. Hydrogen gets no repeats as it has one electron in its valence shell. Each row/column pertaining to a Carbon atom repeats 4 times, each row/column corresponding to a Nitrogen atom repeats 5 times, and each row/column identified as an Oxygen atom repeats 6 times, to obtain the electronic adjacency matrix. We consider that all valence electrons in the atoms interact with their covalent counterparts because they are indistinguishable, and the valence orbitals hybridize (for example, sp^2, sp^3 hybridization of valence shell s and $p_{x,y,z}$ orbitals in Carbon) we cannot confidently claim which electrons will form the covalent bonds with the electrons of other atoms. A covalent bond between atoms from a quantum mechanical viewpoint, is simply a combined wavefunction of all electrons participating in the bond formation, with the combined probability of locating the electrons equal at both atoms' nuclei. In a network of covalent bonded atoms, such as in a protein, the distinctions between which electron belongs to which atom is moot, and so we do not distinguish that in the matrix, leading to the formation of a band around the leading diagonal of the adjacency matrix, with sparsely filled lower and upper triangles. As the time steps of the simulations progresses, the adjacency matrices get more and more filled out. We consider secondary bonding to be only hydrogen bonding in this case as conveniently, both Polyalanine and water can only form hydrogen bonds. Van-der-Waals interactions can happen between any atom but they are much weaker compared to LBHBs. We also limit the

creation of hydrogen bonds to any atom within the distance threshold, and impose a second restriction that only Hydrogen–Nitrogen and Hydrogen–Oxygen pairs, regardless of where the hydrogen is coming from (side chain or backbone, or another water molecule).

(g) For each adjacency corresponding to each threshold cutoff distance, and time step for both Polyalanine-21 and water simulations, we count the number of rows denoted as n (corresponding to number of valence electrons in the system), number of non-zero elements in the upper triangle (or lower triangle, but not both) corresponding to the number of bonds (some bonds never changed, those are the covalent bonds, while some changed over time, and were the hydrogen bonds) denoted as m.

(h) We then generate N random matrices $G_i, i = 1, 2, 3, ..., N,\ N > 40$ and then apply the transformation $H_i = (G_i + G_i^T)/\sqrt{2n}$ generating N random hermitian matrices, normalized (normalization factor $\sqrt{2n}$ so that the standard deviation of the off-diagonal elements is half of the standard deviation of the diagonal elements (a property of the Gaussian Orthogonal Ensemble class).

(i) Next, we apply the adjacency matrices as a mask over the Random matrices. We get N different Random Graph Matrices per time frame of the simulation, and for each threshold value ε. In total, for the Polyalanine-21 simulation, considering just the peptide and not its surrounding water, we obtain a $N = 1000$ (number of samples of the random matrix)$\times$10 (ten different ε values)$\times$30000 (number of timesteps)$= 3 \times 10^8$ random graph structured matrices (three hundred million).

(j) For each RGM, we calculated the eigenvalues, the eigenvectors, the local spacings, the global spacing distribution, the continuous differential entropy, the spacing distribution between the median and median+1 indexed eigenvalues, and analyze the results. The values for n do not change with time, as number of electrons is constant, but the values for m changes with time, and with size of the system (depending on how near or far away is the value for ε. We compare the trends in CDE, standard deviations, spacing distributions, eigenvectors, and global empirical spectral measures for various threshold cutoffs and time step in the simulation, along a decreasing potential energy and decreasing conformal entropy gradient. Fig 4.6

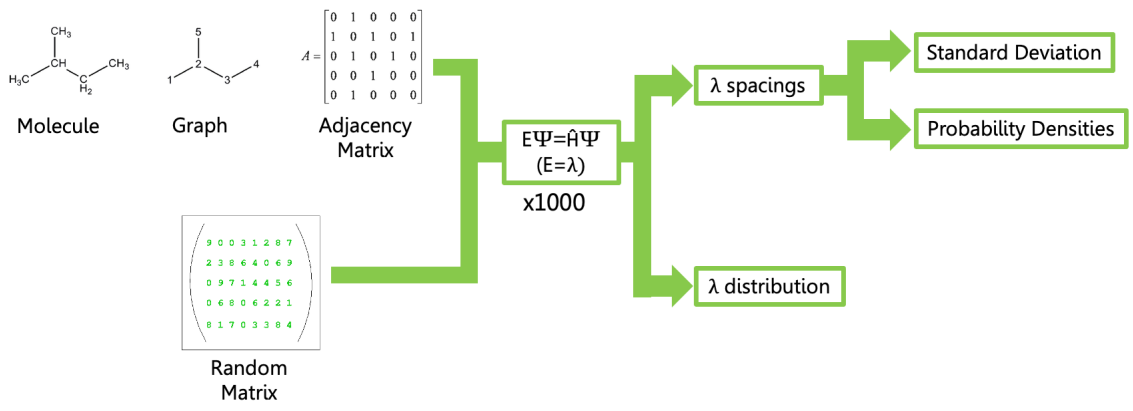


Figure 4.6: **Flowchart describing random matrix theory approach to observing energy state behavior**: The steps shown here correspond to after the MD simulation has ended, from obtaining a molecule as an atomic graph, to converting it to an electron adjacency, and subsequent analysis via random Graph structured matrix modeling.

4.3 Spectral and Eigenvector Statistics of electrons during Hydrogen Bonding in Water

We begin the application of random graph structured matrices (RGM) to physical systems, starting with a simple 1-Dimensional static water network as the simplest scenario, to observe the trends in the global and local spectral statistics of the RGM that models the Hamiltonian of such systems. We then move on to a 3-Dimensional time evolving dynamic water system, simulated via explicit solvent molecular dynamics software, GROMACS,[526] with TIP3P[208] water model (each water molecule is a triangle with polarity at the vertices corresponding to hydrogen atoms and oxygen atom, with a fixed angular separation of 104.5° subtended at the center of the oxygen atom), CHARMM-36m force field,[289] Van-der-Waals cut off distance of 10Å, 295K room temperature and other parameters as described in section 3.3. The next system to model is the Polyalanine-21 peptide solvated with TIP3P water molecules, with the classical dynamics of the peptide folding simulated in GROMACS with the same force field and other parameters.

Classical MD simulations work by initializing the molecular system of concern in an arbitrary configuration. The software then calculates the free energy of the system and the forces that each atom is subject to (gradient of potential energy), changing the configuration of the system in the next time step, (fixed increment of time, usually in picoseconds) by updating the coordinates of the atoms so that the total energy is lessened, i.e., the simulation takes a small step along the energy gradient. The energy and forces are then recalculated, and positions of atoms updated once more along a slightly new but still decreasing gradient, cycling through an iterative process until the global energy of the system has converged upon some minimum in the energy landscape. In some circumstances, a simulation can converge upon a local rather than a global energy minimum, or even a saddle point. MD Metadynamics and other advanced sampling techniques can be used to avoid this issue, albeit at a

higher computational cost.

Additionally, because the Gibbs Free Energy is minimized in GROMACS Molecular Dynamics, we should see the free energy asymptotically approach a minimal value as the system reaches energetic equilibrium[527].

GROMACS MD is designed to minimize energy using classical physics methods, so using a diffusion equation to represent the flow of energy between the lower-energy solvent and higher-energy peptide, we can expect the the free energy of the peptide to follow exponential decay ($G = -\beta_2 e^{\gamma_2 - t} + G_0$), where G_0 is the minimum possible energy of a given system. We assume $\gamma_1, \gamma_2, \beta_1, \beta_2 \in \mathcal{R}$, each constant dictated by initial conditions and constituent molecules of a given system.

Although satisfactory and feasible for use for short time scales, MD simulations present many problems when considered for implementation in practical purposes. The smallest proteins have been found to fold over microsecond timescales[528], 6 orders of magnitude larger than the picosecond (ps) time increments necessary for accurate MD simulations (larger time increments will result in longer steps along the potential energy gradient, thereby sampling the energy landscape in a shallow manner, and the simulation may never find the global energy minimum), often rendering MD simulations inaccurate for simulating large or disordered protein folding. MD simulations already present a relatively high computational cost, which has been improved from $O(N^2)$ to $O(Nlog(N))$ complexity for an N-body systems by introducing GPUs and other more modern methods of computing. Moreover, these issues are also continually being improved by improving the force fields used to calculate the forces on the system[529]. However, because the simulation treats molecules classically, as described before, the atoms are considered as point-like hard spheres utilizing classical thermodynamic principles to obtain order and structure from random oscillations (Brownian models), these simulations cannot account for quantum mechanical effects, which may have a role to play in molecular recognition and signal transduction. Further, we know during protein folding, the conformational entropy of the system decreases.

However, as previously stated, conformational and configurational entropy is only one measure of the potential entropy available for a system to possess, so other measures of entropy must increase to counter the decrease in configurational entropy, although the exact nature of such entropy measures is so far under intense debate. Random Graph Matrices, and the entropy of the distribution of their eigenvalues may provide one such entropic compensation mechanism driving protein folding, specificity during molecular recognition, and signal transduction.

4.3.1 1-Dimensional time invariant water network

To start the analysis of the 1-Dimensional water network energy states using random graph structured Hamiltonians, we created two types of 1-dimensional water networks (indicating only one degree of freedom, along an axis that coincides with the center of the oxygen atoms involved in the chain of hydrogen bonding, see fig 4.7). These 1D water networks consisted exclusively of mathematically generated water molecules connected via hydrogen bonding (no molecular dynamics, but a simple network, mathematically designed, as shown in fig 4.7). Since the system is water, the hydrogen bond is single-well type, i.e., one cannot distinguish whether the hydrogen is covalent or weakly bonded to the neighboring oxygen atoms. The hydrogen nucleus is exactly in between the two oxygen nuclei, and its electron has an equal likelihood of being located around both oxygen nuclei. The electrons of oxygen atoms therefore cannot distinguish between the hydrogen they are covalent bonded with, and the hydrogen participating in the hydrogen bond. Therefore, as long as the water is still (i.e., no bulk flow) they readily delocalize along the entire water network. We designed such networks with an intention to validate the spread of the energy states modeled with a random graph structured Hamiltonian, with increasing number of hydrogen bonds in a simple scenario without time evolution. The behavior of the energy states are analyzed by varying the interaction cutoff distance ε that varies the number of hy-

drogen bonds in the network. We intentionally never break any covalent bonds.
The first type of network named "Network Type A," is a chain of water molecules, where both the water molecules are so configured that both of the hydrogen atoms of the i^{th} water molecule are hydrogen bonded to each of the two lone pairs on the next adjacent $i + 1^{th}$water molecule, forming a chain connected via hydrogen bonds, as shown in fig 4.7(top). Conversely, the second type of network, referred to as "Network Type B," is a linear chain of water molecules so configured that one of the hydrogen atoms in the i^{th} water molecule is hydrogen bonded to one of the two lone pairs on the next adjacent $i + 1^{th}$ Oxygen atom along the linear axis, and the other hydrogen atom of the i^{th} water molecule is hydrogen bonded to a lone pair on the previous adjacent $i - 1^{th}$ water molecule, as shown in fig 4.7(bottom). Both networks, with the same number of water molecules, resulted in the same number of electrons ($n = 350$) and hydrogen bonds. The axis through the oxygen atoms is one dimensional ($\updownarrow$, z-axis, $\leftrightarrow$ is the x-axis) for both networks.

At the beginning, we tried to answer the question whether the configuration of the hydrogen bonds or other long range interactions due to electron delocalization along single-well hydrogen bonds, which is analogous to the location of the non-zero off-diagonal elements in the random graph structured Hamiltonian, impacts any statistics of the global or local eigenvalues. We sought to determine if there was any effect from the position of the bonds, or if the energy state behavior simply depended upon the number of bonds. We expected no change, since the equations for the probability density and continuous differential entropy in chapter 3 only concern itself with the number of on and off diagonal entries in the random graph matrix n, m, and not on which off-diagonal elements were non-zero. To evaluate whether such an expectation truly holds true, we created the two networks A and B as described above and their adjacency by imposing a threshold. We kept all the hydrogen bonds with their next adjacent water molecules intact throughout the simulation and the threshold only changed the distance to which electrons are delocalizing along the single-well hydro-

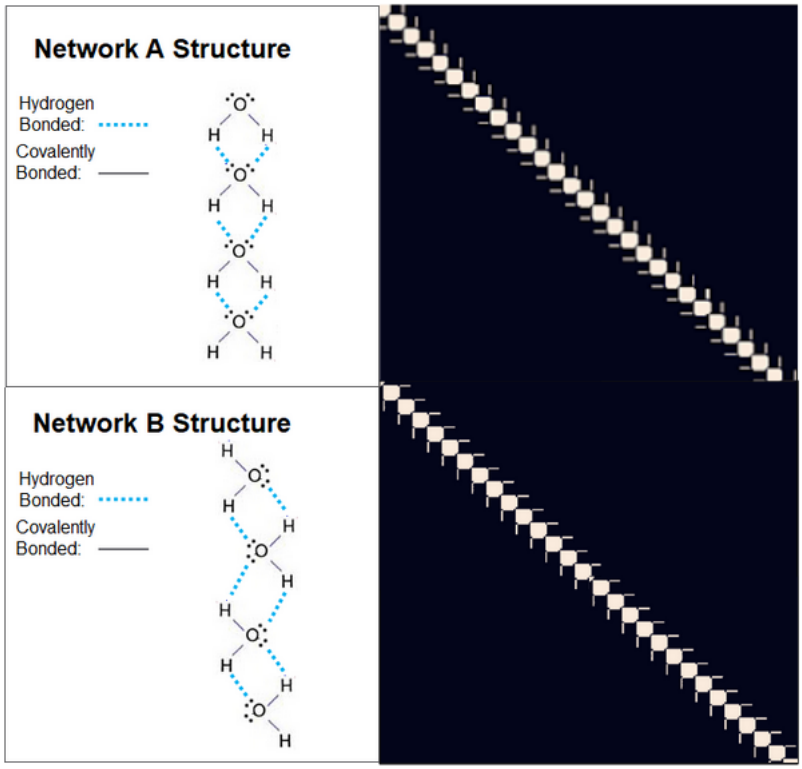


Figure 4.7: **Diagram displaying bond structure and corresponding heat maps in contrived water Network Type A (top) and Network Type B (bottom)**

gen bonds to interact with farther water molecules, i.e., the threshold $\varepsilon \geq 2.5\mathring{A}$, the hydrogen bond distance in water. Such a threshold does not change the number of hydrogen bonds that will 'definitely' be there in a water network such as this, but still adds off-diagonal elements to the random graph structured Hamiltonian, as a way of depicting long range interaction energies. We next sampled 1000 random matrices of size $n \times n, n = 350$ where n is the number of electrons in the network A or B. Since there are same number of water molecules, and the only difference is in the position of hydrogen bonds, not their number, the size of random matrices remained the same for both networks. We then transformed each random matrix by the linear transform $H_k = (G_k + G_k^T)/\sqrt{2n}, k \in \{1, 2, ..., 1000\}$, and then masked the adjacency of the water networks to obtain two sets of random graph structured Hamiltonians,

$(H_k \circ A_A), k \in \{1, 2, ..., 1000\}$ for network type A and $(H_k \circ A_B), k \in \{1, 2, ..., 1000\}$ for network type B. The eigenvalues for network type A were then obtained for each random graph matrix $(H_k \circ A_A)$ and their distribution was plotted. Similarly, the process was repeated for network type B. The superimposed plots of distributions of eigenvalue single point correlation shown in fig4.8 of the eigenvalues (energy states) for each network were identical to each other, suggesting the position of the hydrogen bonds has no bearing on the eigenvalue spectra, but their number. We also chose two eigenvalues in the bulk of the support of the spectrum (the median, and median+1 eigenvalues) and calculated their spacings for each of the 1000 RGMs for Network Type A and Type B separately. We found that the median spacing distributions were also identical and independent of the position of the bonds (i.e., which off-diagonal elements were non-zero), but simply dependent upon the number of electron interactions (either hydrogen bonds, or long range interactions), as the energy state spacing distributions were indistinguishable. This is displayed below in fig 4.8. The result is in agreement with our expectation because, as seen in the adjacency matrix in fig 4.7 between network type A and B, they are simply a linear variation of each other. Therefore, we expect to observe the same statistical behavior between the two networks with increasing value if the threshold variable ε. From this point forward, we arbitrarily chose to use Network Type A to analyze spectral statistics behavior for the RGM with linear water network adjacency.

After choosing to work exclusively using Network Type A, we then moved to observe the energy level (eigenvalue) $SPCF$ upon allowing electrons to resonate at various distances, based on the value of ε. Resonance here pertains to the phenomenon where the electron wave delocalization along the single well hydrogen bonds results in interference with electrons from other atoms and molecules, resulting in resonance between all the waves. A similar but slightly different phenomenon exists in chemistry, where resonating chemical structures of molecules are observed when the electrons are delocalized, and one cannot affix a covalent or other type of bond at one location. For

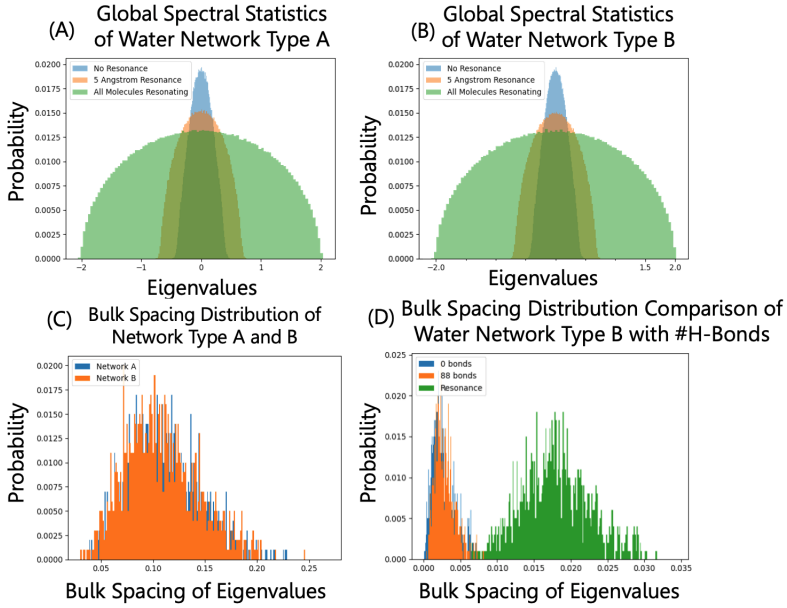


Figure 4.8: **Distribution of eigenvalue single point correlation and median eigenvalue spacing distribution for linear water network A and B**: (A) The Distribution of eigenvalue single point correlation of Network type A with increasing number of hydrogen bonds, denoted by increasing threshold distance, (B) The Distribution of eigenvalue single point correlation of Network type B with increasing number of hydrogen bonds,denoted by increasing threshold distance, (C) The median (bulk) spacing distribution for both network types are identical, indicating that the configuration of the hydrogen bonds or long range interactions do not affect the statistics, but rather the number of such bonds and interactions.(D) The bulk spacings for Network type B with increasing number of hydrogen bonds, corresponding to 0.5Å and maximum distance thresholds.

example, within the benzene rings, which have alternating double and single bonds among the carbon atoms, due to electron delocalization along dangling p orbitals, the carbon atoms in the ring each is said to have 1.5 bonds, as it is impossible to confidently say which carbon has the double bond and single bond in which orientation. We use the threshold variable ε to change the length scale over which electrons can delocalize along the single-well hydrogen bonds, thereby limiting the distance at which resonance occurs. Three such distance threshold were compared: no inter-molecular

resonance (i.e., electrons only in covalent bonds can interact), resonance within a 5Å cutoff (hydrogen binds and resonance can happen only within adjacent molecules), and resonance throughout the entire network (all electrons are interacting). As expected, more bonds (larger resonance distance) corresponded to a more semicircular distribution (wider distribution radius), validating Wigner's Semicircle Rule in the mean field case where all electrons are interacting, as shown in fig 4.9(A). such a case might not be feasible in ambient water bodies with flowing molecules, but it is within the realm of possibility in structured systems where water molecules are connected by strong hydrogen bonds and do not move around, as in cellular systems and vesicles in cells, as well as extracellular fluid in the tissues. Even blood vessels become increasingly constricted as they enmesh into the tissues, as capillaries, with extremely thin apertures of just $8 - 10\mu$m diameter and endothelial cell lining, with red blood cells passing through in a single file.

We next compared the probability distributions of median eigenvalue spacings for various numbers of bonds (fig 4.9(B)), corresponding to interaction distance cutoffs of zero non-covalent interactions, interactions only between adjacent molecules, and all molecules interacting. As expected from the Wigner Surmise, we observed a decrease in degeneracy of energy states (the mean spacing between median and median+1 eigenvalues increased) as interaction distance increased, with a gradual increase from single to double molecule resonance (0 to 88 hydrogen bonds) and an obvious jump from the 88 bond distribution to the distribution for all molecule resonance. We observed that the standard deviations of the spacings seemed to increase from a visual inspection of the spacing plots and as the electron delocalization and interaction distance increases. Expecting this may simply be a visual artifact, we calculated the standard deviation of the spacing distributions at various interaction cutoff distances ε. For each threshold value, we counted the number of hydrogen bonds, and plotted the standard deviation of median spacing distributions with increasing number of hydrogen bonds. Unexpectedly, we found that an increase in interaction distance,

resulting in more number of secondary interactions including hydrogen bonds and farther electron delocalization, corresponded to an overall decrease in standard deviation (the trend is not smooth, but a clear trend is still observable in fig 4.9(C)). The jumps in between are significantly large, so we do not draw any conclusions from it at this point. Instead, we want to see if the decreasing trend comes up again in a larger system and if the trend is more significant than the noise in the observed standard deviations, to rule out that the decreasing trend might be an artifact. If the trend holds significantly for other systems, it can be said that the spacing between eigenvalues in the bulk of the spectrum stabilizes as electrons delocalize further, and if more hydrogen bonds are formed, leading to long range interactions between electrons.
Interestingly, when we analyzed the standard deviation of the distribution of spacing between different adjacent eigenvalue pairs (not exclusively just the median and median+1 index eigenvalues), we observed that the decrease in standard deviation with number of bonds is almost invisible when compared to change in trends between edge and bulk statistics (fig 4.9(D)). Bulk eigenvalue pairs had a much lower standard deviation of spacing distribution throughout, than spacing distribution of pairs of eigenvalues at the edges (near index 0 or 350, with the 0^{th} index eigenvalue being the smallest eigenvalue, and 350^{th} being the largest). Physically speaking, the smallest and largest eigenvalues, or the lowest and highest energy levels, are respectively closest to the nuclei or the surroundings more than the eigenvalues in the bulk, so their behavior is expected to be different. No conclusions can be drawn from their behavior w.r.t. the standard deviation between eigenvalue spacings at the edge or in the bulk of the support of the eigenvalues for the RGM, without further investigation with more intensive numerical simulations, and larger matrices.

4.3.2 3-Dimensional time evolving water network

We simulate 348 TIP3P water molecules in GROMACS (the choice for number of water molecules was taken by the GROMACS software while solvating a simulation

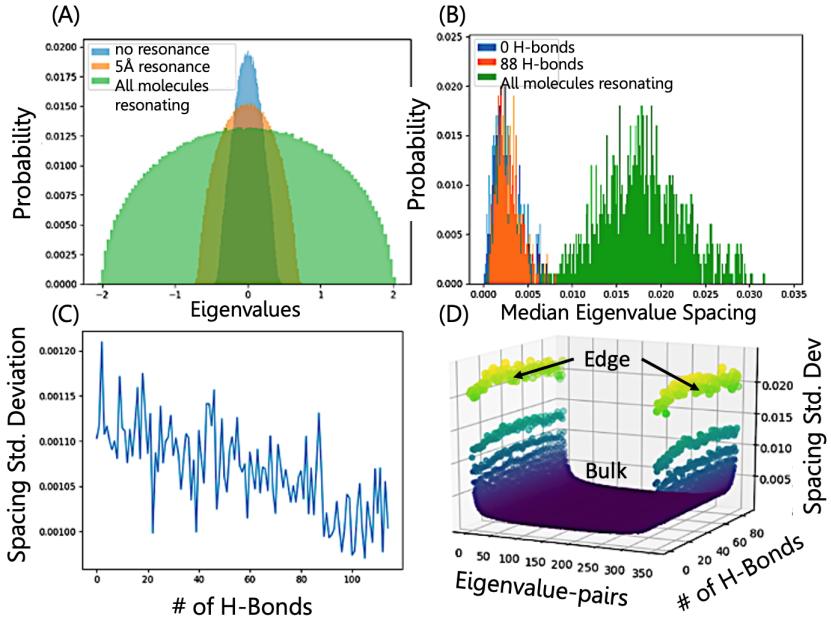


Figure 4.9: **Global and local statistics of RGMs corresponding to linear water network A**:(A)Visualizations of GOE spectral statistics for 2D Type A water network, including overlaid Probability density plot comparing distributions of energy states (Eigenvalue SPCF) with varied interaction cutoff distances, (B) Overlaid probability density plot comparing distributions of spacings of energy states at varied resonance distances, (C) Standard deviation of median eigenvalue spacing distributions varying with number of electron interactions, including hydrogen bonds, in the water network A, and (D) 3D plot of standard deviation of eigenvalue spacings for all consecutive eigenvalue pairs and various number of electron-electron interactions corresponding to changing delocalization distance cutoffs.

volume of $2nm \times 2nm \times 2nm = 8nm^3$). The number of valence electrons per water molecule is 8 (6 from oxygen and 1 each from the two hydrogen atoms), so the size of random matrices sampled for this system was 2784×2784, which is much larger than the 350×350 random graph matrices analyzed for the linear water networks. The water molecules were also initialized by the system automatically, and already placed at ideal locations. The simulation was conducted as an NVE ensemble (number of particles N, volume of system V and total free energy E remains constant) and potential energy was minimized, as is the case for any molecular dynamics simulation.

The simulation ran for 25ns, with 10ps intervals. We ended up with 2500 snapshots of the water dynamics along the trajectory.

We aimed to analyze the global eigenvalue distribution, the local bulk spacing distributions, the entropy of the eigenvalue distribution, as well as the standard deviation of the bulk spacing distribution with time, and increasing delocalization distance of individual electrons. The MD simulation has no explicit electrons, and is purely classical, so any trends we observe, must necessarily be related to trends of metrics that are defined over the Random Graph Structured Hamiltonians, and unlikely to be caused due to MD artifacts. For each of the 2500 time-frames, we imposed a delocalization distance cut off ε on the all atom distance matrix, ranging from 1Å to 10Å, the latter of which is the MD cutoff distance parameter for any secondary inter-molecular interactions. We then converted the all atom adjacency to valence electron adjacency by not repeating the hydrogen rows and columns, and repeating the oxygen rows and columns six times each. Therefore, we obtained 10 different adjacency matrices for each time-frame, where electrons can delocalize up to different distances, thereby able to interact with other electrons farther along in the volume. We next sampled 1000 random matrices of size $n \times n, n = 2784$ where n is the number of electrons in the system of 348 water molecules distributed in $8nm^3$ volume. We then transformed each random matrix by the linear transform $H_k = (G_k + G_k^T)/\sqrt{2n}, k \in \{1, 2, ..., 1000\}$, and applied the ten adjacency matrices per time-frame (total 25000 different adjacency matrices) as masks for the random matrices, to obtain 25000 random graph structured Hamiltonians representing the chaotic electron interaction energies of 2500 time-frames and 10 different delocalization lengths in 3D water networks. We then proceed to analyze the eigenvalue distributions, their continuous differential entropy (CDE), local bulk spacing distributions, their standard deviation, distribution of all eigenvalue pair spacing, and its standard distribution over time, and over different delocalization distances. We also obtain the eigenvectors of the RGM and observe the

locality or non locality of such eigenvectors (corresponding to the largest 5, middle 5, and smallest 5 eigenvalues) over time, for three chosen time frames. We expected that since it is an NVE simulation where free energy is kept constant, and since the software already placed the TIP3P water molecules at almost ideal locations by implicitly minimizing potential energy in the solvation process, the time simulation dynamics should not drastically increase the number of hydrogen bonds. We should not observe any trends over time, but rather over delocalization distance cutoff.

We began by observing the distribution of single point correlation of the eigenvalues (we have a total of 2784000 eigenvalues per adjacency matrix) which are analogous to the allowed energy states of the system. As depicted in fig 4.10(A) below, the distribution of eigenvalue single point correlation grows wider with increasing interaction cutoff distance for electron delocalization along single-well hydrogen bonds and long range interaction (corresponding to a greater standard deviation and greater continuous differential entropy). It can also be said that as the threshold cut off increases, far away electrons become more dependent on each other, i.e., they 'interact' in some way, becoming more and more entangled with each other. From this behavior, we can infer that an increase in the number of pairwise electron interactions, either as hydrogen bonds, or longer range delocalization dependent interactions, correlates with an increase in the CDE of the system. Recognizing that the CDE of the system should show no time-dependence because the molecules are already placed at locations corresponding to the minimal free energy state during solvation of the simulation box, and the micro-canonical ensemble for the MD simulation is NVE. This was validated in fig 4.10(D), as there was no significant trend in the CDE (equation 4.7) of the energy state distribution over all frames throughout the entire simulation. The CDE jumped around, but that is to be expected because MD is a stochastic simulation, that results in noise, and the single-well bonds in water keep breaking and reforming as the water molecules freely rotate in the simulation box without the presence of a

protein to structure them around its conformation.

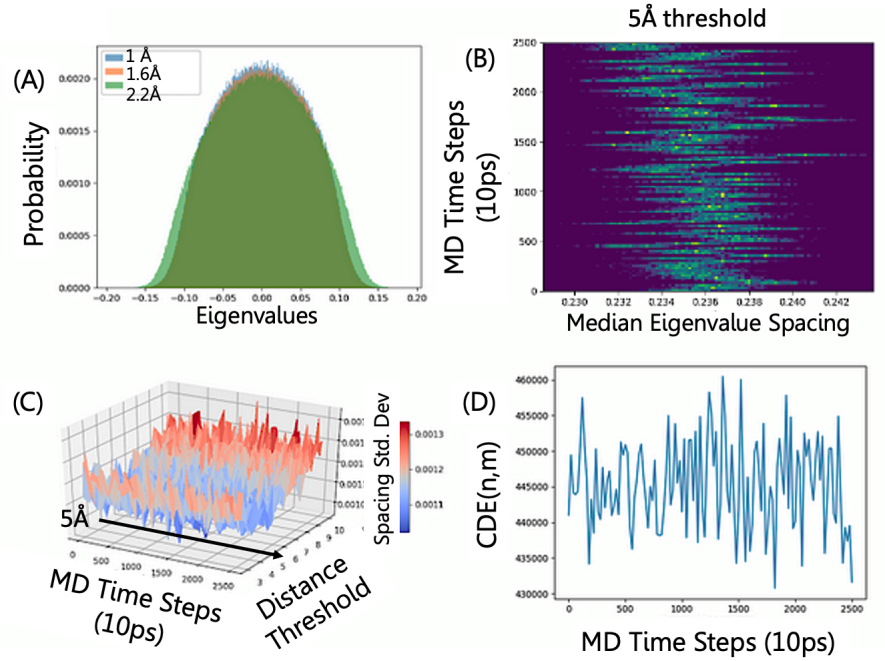


Figure 4.10: **Spectral statistics for 3D time evolving dynamic water network**: Visualizations of spectral statistics for 3D MD simulated water network, including (A) Distribution of eigenvalue single point correlation (energy states/eigenvalue SPCF) for varied interaction /delocalization distances (ε) demonstrating widening distribution of eigenvalue single point correlation with increasing delocalization distance. (B) Bivariate probability distribution plot of energy states/eigenvalue spacing with time evolution, at a fixed interaction threshold distance of 5Å (all valence electrons within that distance interact with one another), (C) Standard deviation of median eigenvalue spacing varying with the time-frame index of the simulation along the y-axis, the interaction distance threshold along the x-axis, and (D) $CDE(n, m)$ time evolution with fixed cutoff distance of 5Å, spanning the entire simulation from 0 to 25ns.

Moreover, as shown in fig 4.11(A)-(D), as anticipated, the mean of the spacing distribution moved further right as the delocalization cut off distance (threshold ε)increased, indicating an increase in spacing between eigenvalues in the bulk of the spectrum. Furthermore, in concurrence with the behavior of CDE in fig 4.10(D),

no noticeable change in the mean of the median eigenvalue spacing distribution was observed as the simulation progressed, indicating yet again, no overall change in the system time for this simulation, as expected. However, as shown in fig 4.11(A)-(D), the median eigenvalue spacing distributions clearly became more erratic and less tightly distributed as the delocalization cut off distance increased, especially past 3Å. This indicates that there is a transition in the otherwise well behaved median eigenvalue spacing distribution beyond the threshold delocalization/interaction distance of 3Å for the 3D water system. We suspected, from the linear water network that standard deviation of median spacing might decrease with increasing delocalization cut off distance, but it seems that for a cut off threshold > 5Å, the standard deviation increases, and there is increased noise in the mean of the median spacing distribution, as well as a noisy shift towards bimodal or multimodal distributions (fig 4.11(A)-(D)). We conjecture that farther than 4-5Å delocalization of individual electrons results in destabilization of the system, and so it might impose an upper bound on how far can electrons remain delocalized in a water system at room temperature.

We then analyzed the spread of the spacings by creating a surface plot (fig 4.10(C))for the median eigenvalue spacing distribution's standard deviation by varying the cutoff distance (analogous to number of pairwise electron interactions) and the time-frame of the simulation. Again, as expected, we found no time-dependence of the standard deviation for this particular system. Interestingly though, we did identify a local minimum of standard deviation at a particular delocalization/interaction threshold distance ($\varepsilon = 4 - 5$Å) across all the time-frames, notwithstanding any difference in positions of the water molecules (which shows up as noise along the frame index direction). The standard deviation of median eigenvalue spacings grows with an obvious significant trend (the trend is larger than noise) as the delocalization/interaction cut off distance increases beyond 5Å. Interestingly, since the O–H bond length is about 0.97Å, and the hydrogen bond length is about 2.5Å, a 4-5Å delocalization distance corresponds to about the length from the covalent bonded hydrogen atom of one wa-

ter molecule in a pair of hydrogen-bonded water molecules, through the oxygen, the hydrogen involved in the hydrogen bond, the other oxygen and to the covalent bonded hydrogen atom of the other water molecule, which, is well within feasibility range for electron delocalization distances in still water.

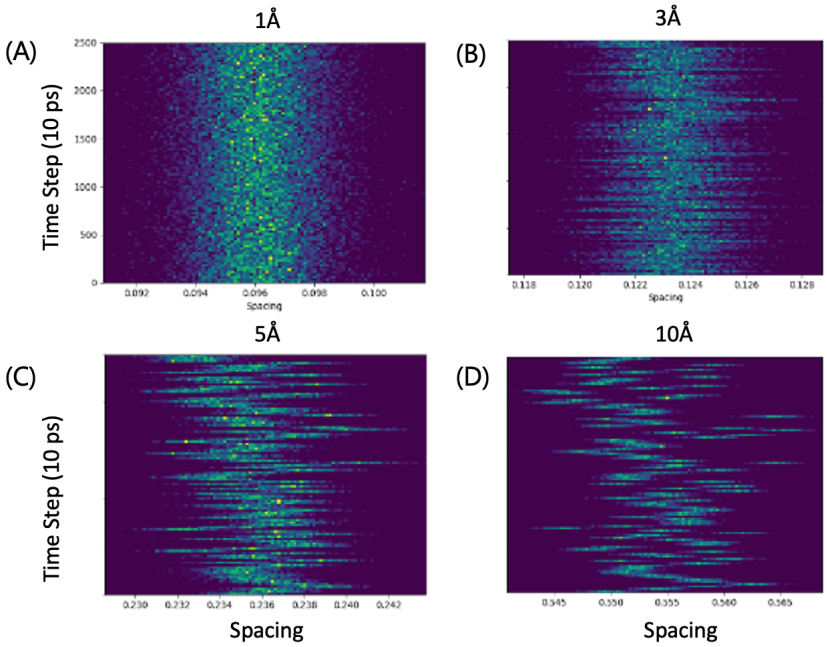


Figure 4.11: **Heatmaps depicting the trends in median eigenvalue spacing distribution**: The figure shows the trends in the median eigenvalue spacing distribution with time for four different sample delocalization distance thresholds of (A) 1Å, (B) 3Å, (C) 5Å and (D) 10Å

Next, we take the spacing distribution of all eigenvalue pairs (fig 4.12), for each of the 1000 random matrices, for three sample time-frames (0, 1250, and 2500) at 5Å threshold cut off. We observe that there is no change (the distributions are identical) in the spacing distribution of all eigenvalues with time, as expected from this water only simulation.

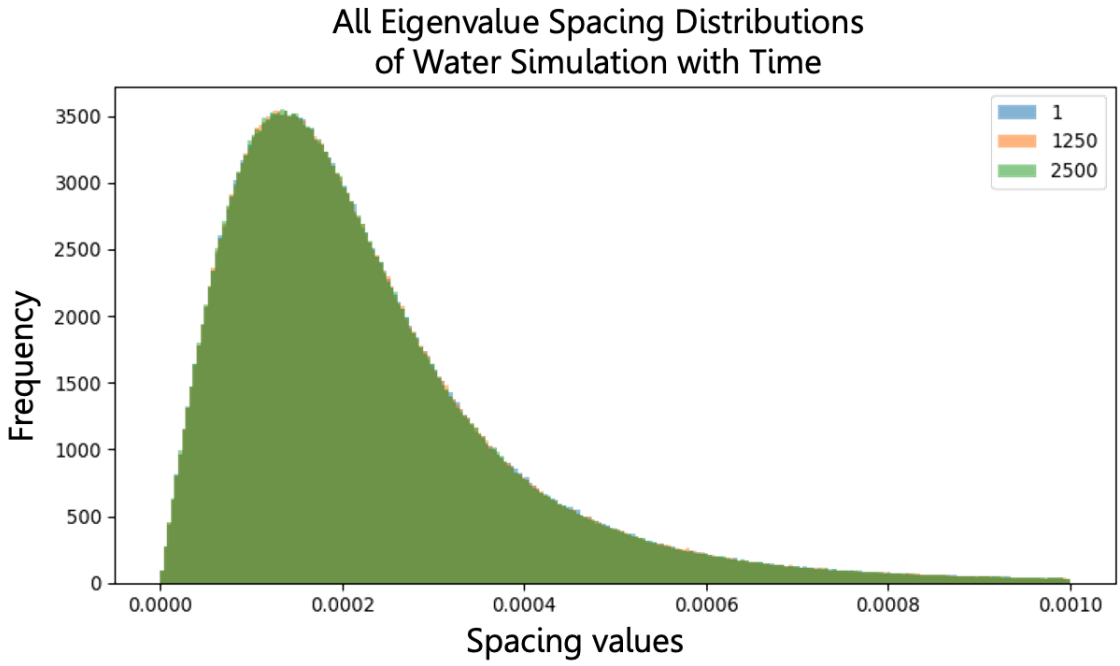


Figure 4.12: **All eigenvalue-pair spacing distribution of the RGM representation of 3D water network**: The figure shows the spacing between all the eigenvalue pairs of all the 1000 random matrices that model the Hamiltonian for each adjacency matrix (threshold distance 5Å of the 3D water system at the 0_{th},1250^{th} and 2500^{th} time-frame of the MD simulation. X-axis is the value of the spacing, and y-axis is the number of eigenvalue pairs within a small spacing bin (frequency). All the distributions overlap, and are identical, as expected for the water network simulation with NVE micro-canonical ensemble.

We also take the largest five, middle five, and lowest five eigenvalues for one of the 1000 random graph Hamiltonians, corresponding to the adjacency matrix obtained from the 0_{th},1250^{th} and 2500^{th} time-frame of the MD simulation at 5Å threshold cut off. For each of these eigenvalues, we obtain the eigenvectors. The plots of the eigenvectors are shown in fig 4.13(a) corresponding to the smallest, (b) to the middle, and (c) to the largest eigenvalues, for each of the three time frames. We observe that even though, individual electrons interaction distance is just 5Å, and no more, the eigenvectors for the largest, middle as well as the smallest eigenvalues, for all the

three time-frames, are non localized. Moreover, the adjacency matrix at 5Å cut off is very sparse, and yet eigenvectors are delocalized throughout, indicating that far away electrons might become entangled even with short delocalization distances through intermediate electrons, leading to ergodic behavior of eigenvectors.

In the next section, we repeat the exact same process, but this time for a system with Polyalanine-21 peptide folding in TIP3P water as a solvent because polyalanine has a very stable alpha helix structure in water.[115] We do not include the water itself in the analyses, rather just the peptide, so we ignore any secondary bonds with surrounding water, as Polyalanine-21 is mildly hydrophobic. One could analyze the MD trajectory with as much of the solvent water as one needs by simply including their coordinates from the output file in the all-atom distance map, before building the adjacency matrices.

4.4 Spectral and Eigenvector Statistics of electrons during Polyalanine Folding in Water

We simulate one Polyalanine-21 peptide that is folding in $\sim$ 40000 TIP3P water molecules in GROMACS (the choice for number of water molecules was taken by the GROMACS software while solvating a simulation volume of $10nm \times 10nm \times 10nm = 1000nm^3$). Since we just work with the Polyalanine-21 peptide, and ignore the solvent interactions in this particular work (future work should include solvent water), the only atoms we need to consider are carbon (4 valence electrons), hydrogen (1 valence electron), nitrogen (5 valence electrons), and oxygen (6 valence electrons). There are 63 carbon atoms, 107 hydrogen atoms, 21 nitrogen atoms, and 21 oxygen atoms, so the size of random matrices sampled for this system is 590×590. The simulation was conducted as an NVE ensemble (number of particles N, volume of system V and total free energy E remains constant) and potential energy was minimized, as is the

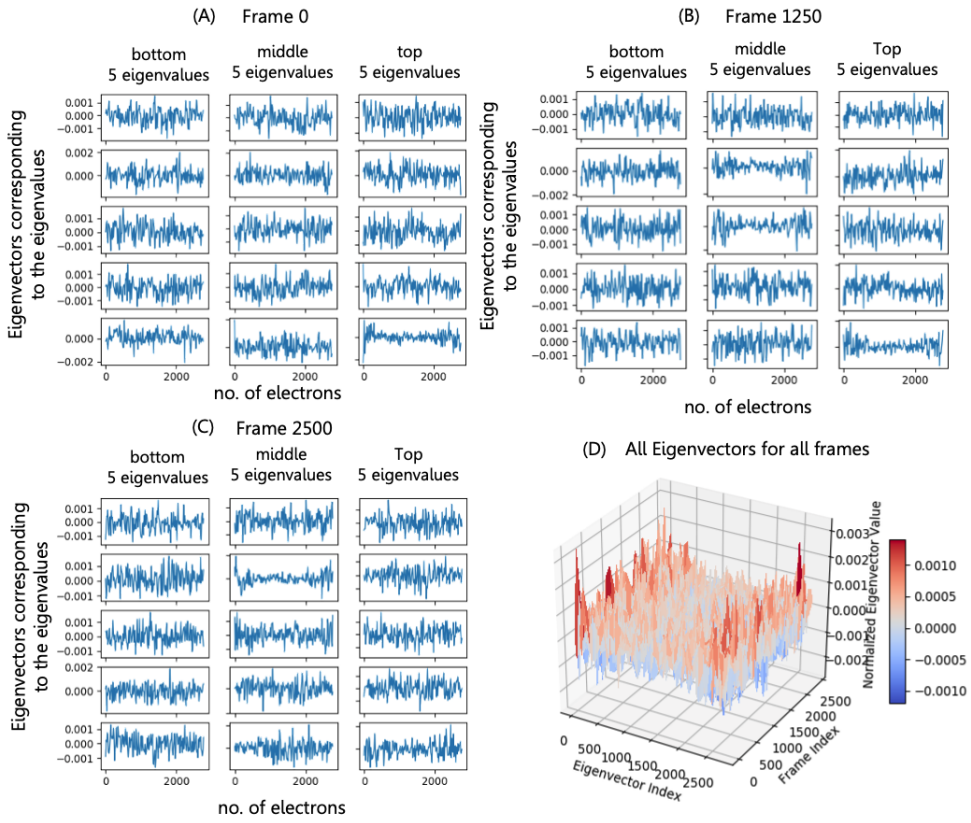


Figure 4.13: **Eigenvectors of the RGM representation of 3D water network**: (A) Eigenvectors corresponding to the largest, middle and smallest 5 eigenvalues of the RGMs constructed with adjacency matrix from frame 0 of the 3D water system simulation, (B) Eigenvectors corresponding to the largest, middle and smallest 5 eigenvalues of the random graph matrix constructed with adjacency matrix from frame 1250 of the 3D water system simulation, (C) Eigenvectors corresponding to the largest, middle and smallest 5 eigenvalues of the random graph matrix constructed with adjacency matrix from frame 2500 of the 3D water system simulation, and (D) The eigenvectors for all time frames from 3D water system simulation. Quantum unique ergodicity is displayed.

case for any molecular dynamics simulation. The simulation ran for $300ns$, with 10ps intervals. We ended up with 30000 snapshots of the Polyalanine-21 folding dynamics along the trajectory, from an extended structure in the first time frame, to a fully folded alpha helix towards the end.

We aimed to analyze the distribution of eigenvalue single point correlation, the local bulk spacing distributions, the entropy of the eigenvalue distribution, as well as the standard deviation of the bulk spacing distribution of a large disordered RGM type Hamiltonian operator with increasing time and increasing delocalization distance of individual electrons. We also wanted to investigate the eigenvectors, as well as the spacing distribution of all eigenvalue pairs of the random graph matrices that model the Hamiltonian during Polyalanine-21 folding, with protein-only adjacency matrices. As usual, the MD simulation has no explicit electrons, and is purely classical, so any trends we observe, must necessarily be related to trends of metrics that are defined over the Random Graph Structured Hamiltonians, and unlikely to be caused due to MD artifacts. For each of the 30000 time-frames, we imposed a delocalization distance cut off ε on the all atom distance matrix, ranging from 1Å to 10Å, the latter of which is the MD cutoff distance parameter for any secondary inter-molecular interactions. We then converted the all atom adjacency of size 212×212 to valence electron adjacency by not repeating the hydrogen rows and columns, and repeating the oxygen rows and columns six times each, carbon rows and columns four times each, and nitrogen rows and columns five times each, for a total of 590 rows and columns. Therefore, we obtained 10 different adjacency matrices for each time-frame, where electrons can delocalize up to different distances, thereby able to interact with other electrons farther along in the volume. We next sampled 1000 random matrices of size $n \times n, n = 590$ where n is the number of electrons in the system of Polyalanine-21 folding by itself in a $1000nm^3$ volume simulation box surrounded by $\sim$50000 water molecules. We then transformed each random matrix by the linear

transform $H_k = (G_k + G_k^T)/\sqrt{2n}, k \in \{1, 2, ..., 1000\}$, and applied the ten adjacency matrices per time-frame (total 300000 different adjacency matrices) as masks for the random matrices, to obtain 300000 random graph structured Hamiltonians representing the chaotic electron interaction energies of 30000 time-frames and 10 different delocalization lengths in 3D water solvated Polyalanine-21 peptide. We then proceed to analyze the eigenvalue distributions, the distribution of eigenvalue single point correlation and their continuous differential entropy (CDE), local bulk spacing distributions, their standard deviation, distribution of all eigenvalue pair spacing, and its standard distribution over time, and over different delocalization distances. We also obtain the eigenvectors of the RGM and observe the locality or non locality of such eigenvectors (corresponding to the largest 5, middle 5, and smallest 5 eigenvalues) over time, for three chosen time frames. We expect to see that as the peptide folds into a stable alpha helix, the global eigenvalue distribution of the random graph matrices that model the electron interaction Hamiltonian during the event will become more spread out, increasing the continuous differential information entropy (CDE) of the distribution of its eigenvalue single point correlation. We should correspondingly see an increase in the mean of the median eigenvalue spacing distribution, leading to lower and lower degeneracy among the energy levels of the system. We expect to also see that the standard deviation of the median eigenvalue spacing distribution should decrease significantly with increasing electron delocalization distance, and then increase after a certain threshold due to destabilization of the system, just like we observed in the 3D water simulation. However, we are still unsure of the reason behind such a trend. We further expect the all eigenvalue pair spacing to show a trend in this simulation, as opposed to no trend in the previous simulation of just water. We expect that the eigenvectors in this system will also display ergodic behavior, even with sparse connectivity.

As displayed below (fig 4.14), for the Polyalanine-21 folding system, we observed that the mean of the median eigenvalue spacing distribution did increase with delocaliza-

tion/interaction cut off/threshold distance ε.

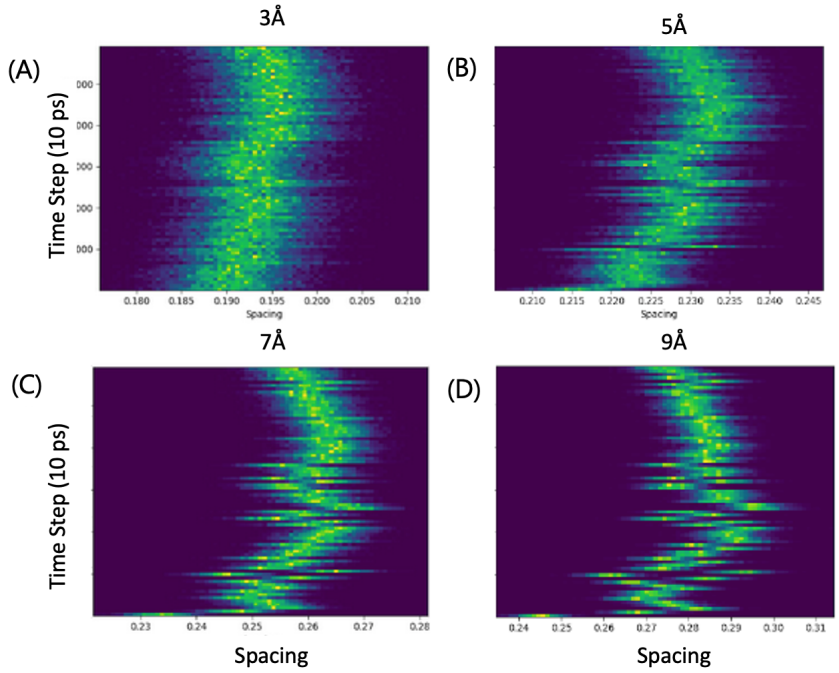


Figure 4.14: **Median eigenvalue spacing distribution during Polyalanine-21 folding at various threshold distances**:Bivariate heat maps depicting probability distribution of median eigenvalue spacings (along x-axis) and time-frame in MD simulation (along y-axis), with the map evolving with varied interaction cutoff distances for Polyalanine-21 folding. Included threshold distances are (A) 3Å, (B) 5Å, (C) 7Å and (D) 9Å

As shown in fig 4.15(A), the behavior of a Polyalanine-21 peptide folding simulation followed the expectation that the distribution of eigenvalue single point correlation of the random graph structured matrices that model electron interaction dynamics during protein folding, widens as the peptide folds as simulation progresses, increasing the number of hydrogen bonds, and pairwise electron interactions. It can also be said that as the peptide folds, far away electrons become more dependent on each other, i.e., they 'interact' in some way, becoming more and more entangled with each other, some delocalizing along the low barrier hydrogen bonds formed within the α-helix,

which may explain the increased conductivity of the helical structure. From this behavior, we can infer that an increase in the number of pairwise electron interactions, either as hydrogen bonds, or longer range delocalization and interaction along LBHBs, correlates with an increase in the CDE of the system. True to expectation, the CDE in fig 4.15(D) increased as time passed in the simulation, and stabilized around 200ns when the Polyalanine-21 peptide formed its alpha helical structure. The CDE jumped around, but that is to be expected because MD is a stochastic simulation, that results in noise. However, the trend is pretty obvious. Moreover, Varying interaction distance threshold showed that the standard deviation of median eigenvalue spacing distribution first decreases then increases with increasing delocalization distance threshold, showing a strong minimum along all time frames in the MD simulation at 5Å (fig 4.15(C)). The standard deviation of median eigenvalue spacing distribution also decreases as the system evolves with time, but the change is not as strong as with increasing delocalization distance. Mean spacing of median eigenvalue spacing distribution increases with time as protein folds, until roughly 200 ns, stabilizing thereafter when the peptide assumes a stable helical structure (fig 4.15). The CDE for Polyalanine-21 exhibited logarithmic behavior with respect to time as it asymptotically approached its stable value in its folded α-helical form, while GROMACS reduces the potential energy of the peptide exponentially (overlaid theoretical exponential decay in fig 4.15(D)).

We recognize that the standard deviation of the median eigenvalue spacings significantly increases with interaction cutoffs greater than 6Å and less than 5Å, indicating less stable energy spacings for these respective delocalization threshold distances. This is in agreement with the trend previously observed for the bi-variate energy state spacing heat maps in fig 4.14. A minimum standard deviation was identified between 5Å-6Å, which is interestingly the distance at which low barrier hydrogen bonds exist within α-helical peptide structures, the helical pitch being equal to 5.4Å. It is also the distance between the extreme end hydrogen atoms between two hydrogen bonded

water molecules, and in case of peptides, it is the distance between the amino group nitrogen and the carboxyl group oxygen that take part in hydrogen bond formation in an α-helix.

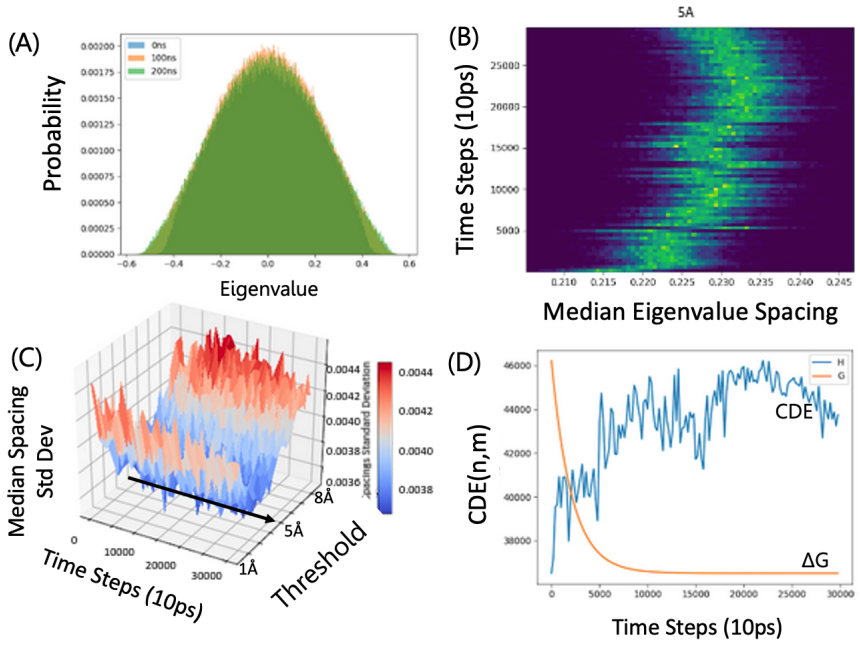


Figure 4.15: **Spectral statistics of RGM representation of peptide folding**:Visualization of spectral statistics for 3D MD simulated Polyalanine-21 peptide, including (A) Distribution of eigenvalue single point correlation for sample MD-simulation time-frames in the simulation, (B) Bi-variate probability distribution plot of median energy states/eigenvalue spacing with time-frames, at a fixed delocalization/interaction threshold distance of 5Å,(C) Standard deviation of energy states varying with the frame index of the simulation along the y-axis, the interaction distance cutoff along the x-axis, and (D) CDE time evolution with fixed cutoff distance of 5 angstroms spanning the time domain of the entire simulation, with theoretical MD energy graph superimposed (bottom right)

Next, we take the spacing distribution of all eigenvalue pairs (fig 4.16), for each of the 1000 random matrices, for three sample time-frames (0, 20000, and 30000) at 5Å threshold cut off. We observe that there is a small change in the spacing distribution

of all eigenvalue pairs with time. Indeed, the standard deviations of the distributions increase with time.

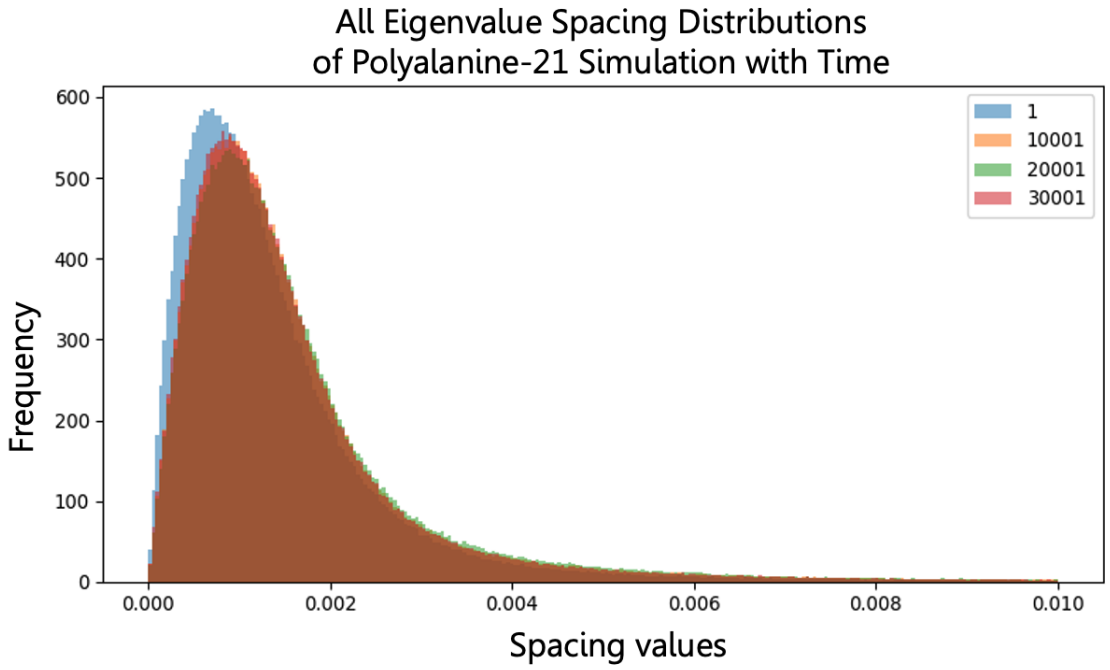


Figure 4.16: **All eigenvalue-pair spacing distribution of the RGM representation of Polyalanine-21 folding**: The figure shows the spacing between all the eigenvalue pairs of all the 1000 random matrices that model the Hamiltonian for each adjacency matrix of a Polyalanine-21 peptide at the 0_{th},20000^{th} and 30000^{th} time-frame of the MD simulation, at 5Å threshold. X-axis is the value of the spacing, and y-axis is the number of eigenvalue pairs within a small spacing bin (frequency). Mean of the distributions increase, showing repulsion between eigenvalues as peptide folds (note that we have not included the water molecules, which we propose to include in future case studies. Inclusion of water molecules will impact the trends observed) into an α-helical structure

We also take the largest five, middle five, and lowest five eigenvalues for each of the 1000 random graph Hamiltonians, corresponding to the adjacency matrix obtained from the 0_{th},20000^{th} and 30000^{th} time-frame of the MD simulation at 5Å threshold cut off. For each of these eigenvalues, we obtain the eigenvectors. The plots of the

eigenvectors are shown in fig 4.13(a) corresponding to the smallest, (b) to the middle, and (c) to the largest eigenvalues, for each of the three time frames. We observe that even though, individual electrons interaction distance is just 5Å, and no more, the eigenvectors for the largest, middle as well as the smallest eigenvalues, for all the three time-frames, are non localized. Moreover, the adjacency matrix at 5Å cut off is very sparse, and yet eigenvectors are delocalized throughout, indicating that far away electrons might become entangled even with short delocalization distances through intermediate electrons, leading to ergodic behavior of eigenvectors.

In the next chapter, we discuss the results some more and the repercussions of the results and relate it with driving forces of protein folding as regards a candidate for the compensating entropy term, specificity during molecular recognition and information flow during signaling cascades.

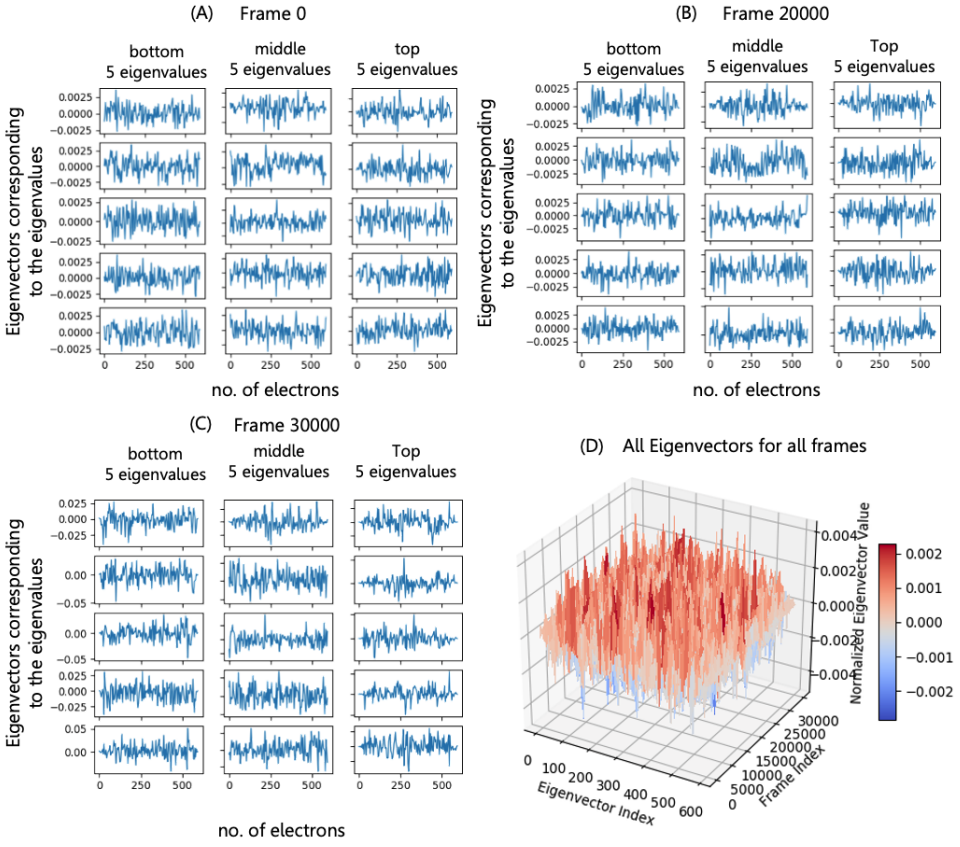


Figure 4.17: **Eigenvectors of the RGM representation of Polyalanine-21 peptide folding**:(A) Eigenvectors corresponding to the largest, middle and smallest 5 eigenvalues of the RGMs constructed with adjacency matrix from frame 0 of the Polyalanine-21 system simulation, (B) Eigenvectors corresponding to the largest, middle and smallest 5 eigenvalues of the random graph matrix constructed with adjacency matrix from frame 20000 of the Polyalanine-21 system simulation, (C) Eigenvectors corresponding to the largest, middle and smallest 5 eigenvalues of the random graph matrix constructed with adjacency matrix from frame 30000 of the Polyalanine-21 system simulation, and (D) all the eigenvectors for all time frames from the Polyalanine-21 system simulation.

Chapter 5

CONCLUSIONS & IMPACT

From Chapter 2, we see that although the distance maps are ubiquitous in deep-learning representation of proteins and other biomolecules, not much is included about the properties of the distance maps and the molecular graphs themselves. Without including inherent physics of the molecules that could be gleaned by analyzing the molecular graphs, it is hard to say what metrics should be optimized so that a VAE-GAN as described in Chapter 2 could be successful, and would work for predicting protein structures and functions end to end, without involving conventional molecular modeling tools. The work on RGM presented in the current work is a preliminary attempt at delineating which metrics obtained from molecular graphs that are originally obtained by thresholding distance maps can be used to optimize within deep-learning algorithms for protein structure-function prediction.

For both the simulated water and Polyalanine-21 systems, we observe that the key tenets of random matrix theory for graph structured matrices are valid. For the water network, we observe that the cutoff distance threshold ε at which the standard deviation minimum of energy state (Eigenvalue) spacing distributions lies between 3-5 Å. This is also the known maximum length of two water molecules connected by a hydrogen bond. Our analysis of the Polyalanine-21 peptide varied interaction distance cutoff from 1-10Å that produced median energy state spacings with minimum standard deviation at a cutoff distance of 5-6Å, which is the corresponding maximum distance between the C_α atom covalently bonded to the nitrogen (of the –NH– group) and the C_α atom covalently bonded to the oxygen (of the –C[=O]O– group) involved in the formation of low barrier hydrogen bonds that stabilize and are a renowned

feature of the α-helical structure in proteins.

It is a consequential result that such a minimum would occur at such important lengths, across all the time frames in the MD simulation (regardless of configuration of water molecules), and that the interaction cut off distance (Lennard Jones Potential) assigned by the MD simulation was 10 Å, two times larger than the 5-6Å threshold distance. The fact that individual eigenvalue spacings in the bulk of the spectrum stabilize the most at this interaction length as the Polyalanine-21 peptide folds and interacts with water, leads us to conjecture that the energy eigenvalues of the proteins become asymptotically stable as the protein folds, with stable spacings between them, and that the distances up to which electrons can easily delocalize in the structure and with the solvent, which is **unique** to water as the solvent, leads to stabilization of the energy levels. Any longer lengths of delocalization or interaction of individual electrons with other electrons, or more asymmetric/longer hydrogen bonds would not allow for such stabilization of the energy levels, according to our observations. Moreover, we see that the spacings between the energy levels, although stabilize, the likelihood that two pairs of energy levels will have the same spacing between them, goes down. It indicates that as proteins fold, energy levels stabilize at values that are less and less regularly spaced. If we use Wigner's idea that the random matrices model the Hamiltonian operator, then the stabilization of eigenvalues have implications for sustained quantum coherence in water, and uniquely in water due to the happenstance that secondary bond-lengths in water are about 5Å. What would be needed to prove the conjecture about the importance of water to maintain quantum coherence, or to measure delocalization of electrons in water-peptide systems would be to computationally investigate the systems with conventional Density Functional Theory. Although experimental measurements would be ideal, no experiment currently exists that can measure quantum properties of a water-protein system at ambient conditions.

The certainty and uncertainty in measurement of energy levels of the protein receptors by other molecules and ligands, as described previously in chapter 4, section 4.1 in the context of fig 4.3, indicates that a biomolecule that would favorably interact with such a protein, must therefore possess the exact complement to the pattern of energy level spacings and matching wavefunction frequencies of its electrons so that electrons between the two systems can interact and overlap without a net change in energy, which is the case during molecular recognition. Such a phenomenon might indicate an electron energy level spacing patterns in the protein systems can be a candidate for the fuzzy proofreading mechanism during molecular recognition at well-structured sites. However, molecular recognition happens more often at locations in the protein that are intrinsically disordered, i.e., that do not possess a stable structure, rather keep flitting between different random coil conformations. Such disordered regions, for example, are found at the top and bottom ends of trans-membrane proteins that are exposed to the extracellular matrix and the intracellular cytoplasm, while the well structured α-helical regions are used to embed the protein tightly in the plasma membrane (lipid bilayer). Such an observation raises an interesting conjecture. Since disordered proteins do not have stable structures, whichever candidate provides the easiest way for electrons to delocalize for about 5Å, might be the candidate that the protein recognizes. It is well known that at the instant of molecular recognition, the disordered region on one side (for example the extracellular side in transmembrane protreins) instantly stabilize into conformations, triggering the signal transduction cascade on the other end (for example the cytoplasm side) of the protein where the structure unravels for a short while. It subsequently releases the bonded molecule and goes back to its initial state after the short while, ready for another recognition event. The events do not last long, agreeing well with very short delocalization times of electrons in the extracellular matrix with ions floating around that disrupt delocalization.

Throughout the protein folding process, a solvent (water in the usual case) also

becomes more structured than in its isolated form. If sufficient structure were to be maintained in this gel-like condition, quantum coherence could be preserved on the inter-protein scale for longer times. Furthermore, the previously discussed quantum coherent states could be actually bolstered by external thermodynamic noise[432,530]. Water is also widely recognized as a key driver of protein folding into its native state. Structured water is also a candidate medium for information transfer via the experimentally observed low barrier hydrogen bond switching[531]. Such a mechanism exclusively allowed by quantum mechanics could allow for immediate information transfer from the disordered active site throughout the rest of the well-ordered peptide via a cascading of signals and electronic interactions. This offers an explanation for the long-range, rapid interactions observed in inter-peptide communication but not explained by classical physics[460,531]. Such electronic switching has also been observed in Geobacter, where electron transfer has been identified as a necessity for reliably producing nanowires[532].However, the nature of the information transfer is still somewhat mysterious. The potential explanation that the specificity during molecular recognition and driving force of protein folding is mediated by the structured water network, mandates several further investigations. Chiefly, the question of how far this structured water network could radially extend from the backbone of a peptide must be explored. The answer to this is crucial to answering whether or not a peptide could utilize delocalized electrons and protons to relay information. Such structure has been observed both experimentally and through the use of MD simulations in water-peptide systems for up to 10 Å away from the peptide backbone. For larger proteins, that are not isolated in the cytoplasm, the thickness can reach even further. Going hand-in-hand with the distance water networks can be structured, many scientists have also recently deduced that life likely did not actually originate in the deep volcanic vents under the ocean, instead originating in shallow pools[533–535] near the banks of freshwater or saltwater bodies, with repeated cyclical periods of hydration and drying. The important distinction of this discovery is that while oceans would be

sparsely populated with organic polymers (relative to the amount of water), drying shallow pools would allow for a critical density of organic structures to be achieved, coated wioth structured water. The aspect of this goes hand in hand with the structure of solvent networks because with a critical density of proteins and other organic molecules in a network, the structure could theoretically extend throughout the entire solvent network, not allowing thermodynamic noise to perturb the delicate quantum effects which inter-molecule communication necessitates.

Further, one of the major results of this work, is that as conformal entropy of the protein goes down when it folds, uncertainty in the electronic energy distribution increases (increase in continuous differential entropy of global eigenvalue distribution), contributing to an increase in information entropy. Such a clear anti-correlation between decrease in conformal entropy and increase in information entropy (of the energy level distribution), posits a candidate that can act as a compensatory term in the entropy driven protein folding phenomenon. Such an increase in entropy as the protein folds is likely counterbalanced by steric clashes, and the optimum electron interaction distance of $\sim$ 5Å at which the energy levels will stabilize. We truly are operating at the edge of classical and quantum mechanical behavior, necessitating future work into coherent states which are defined as eigenstates of the Hamiltonians that behave classically, or display classical oscillations. Quantum harmonic oscillators are an ideal version if such systems[536]. Connecting to previous observations, quantum coherent states have also been experimentally observed in crystallized peptides (Lysozyme). These coherent states were observed to be involved in capturing energy added to the system, where they condense into the lowest-frequency vibrational mode of the system, exerting large influence on peptide function, including structural changes[537].

Additionally, the information entropy in terms of uncertainty of electron energies, goes up when the proteins attain stable structures. Going back to the scenario of a transmembrane protein, when the disordered region that is exposed to the extracellular matrix molecularly recognizes its counterpart, it attains a stable structure for

a short while. For that short while, information entropy of the energy distribution of electrons at that site increases. Simultaneously, on the end of the protein, which is exposed to the cytoplasm, the previously hidden structured regions are suddenly unraveled and exposed, leading to a decrease in information entropy of the electrons' energy states. Therefore, there exists a very clear direction of information flow, which could be using the electrons that have delocalized for a short distance and now are entangled with the far away electrons via stable LBHBs in the intervening α-helical regions of the protein. It can be thought of as analogous to a system of hanging pendulums in a row just barely touching each other where any external force on one end would ideally immediately knock out the pendulum at the other end. RGM treatment of protein folding therefore provides a candidate mechanism to understand information flow during signal transduction cascades triggered at every step by specific molecular recognition events.

Moving forward, these methods can be used to characterize the distance at which quantum effects are preserved. Namely, we can use this method to convert a solvated peptide into a graph, and vary how many edges away from the peptide backbone the structured water bonds and interactions extend. At whatever degree of separation the continuous differential entropy trends begins to plateau as the protein folds, we can deem this the maximal distance from the protein until which there is an overall reduction in conformal entropy and increase in information entropy simultaneously. The distance might also indicate the preferable length until which quantum coherence could be preserved.

Equipped with the ability to characterize proteins in terms of their Continuous differential entropy of global spectral distributions, standard deviation trends of spacing distributions in the bulk of the spectrum, and concurrently, the standard deviation trends of distribution of spacing between all eigenvalue pairs as a protein folds, we now have another powerful descriptor of the system which is also depicted as a molecular

graph. However, unlike previous graphical representations of molecules in data-driven structure prediction models, the metrics we established and our entropic characterization is agnostic towards the specific constituent molecules of a system. It also provides an interpretable quantifier of the disorder of the system, as well as tractable metrics to optimize a graph upon. Moving forward with this interpretable quantity, we posit that such metrics and their formulations in chapter 3 should be incorporated in molecular graph AI systems for both structure prediction, and interaction prediction purposes. Such networks might actually help glean unknown mechanisms of protein folding, molecular recognition and signal transduction in the living world.

To summarize the major conclusions and impact of this work (fig 5.1), we list the following:

(a) Distance matrices alone, although informative representations of molecules, cannot guide deep-learning models statistically in an interpretable fashion, without incorporating physically relevant metrics that can be obtained from properties of the distance matrices and graphs created from them.

(b) Eigenvalues of Random Graph structured Hermitian matrices (RGM) that model electron dynamics during protein folding, indicate that as the edges between nodes in the graph increases, the global spectral distribution flattens out until it asymptotically approaches the semicircle density, as in a Gaussian Orthogonal Ensemble symmetry class of mean-field random matrices.

(c) Such spreading out of the eigenvalues indicates that as more and more edges form between the nodes, analogous to more interactions between electrons in a molecule, the energy levels become less and less degenerate, agreeing with how molecular orbitals form upon covalent bonding.

(d) The continuous differential entropy of the global spectral distribution increases

as more edges form, indicating that as more electrons interact and bonds form, as when a protein folds, information entropy of the goes up, potentially compensating for the decrease in conformational entropy, driving protein folding spontaneously.

(e) Eigenvalues of the RGM settle into a stable value as the number of edges in the graph increases, but only until a certain distance threshold, beyond which the eigenvalues start becoming more unstable. The result indicates that as a biomolecule or protein folds and electrons delocalize along the newly formed Low Barrier Hydrogen Bonds, the energy levels gradually settle into particular values and don't change much. Granted that electrons can only delocalize for 5Å along the LBHBs, if we had any other solvent other than water, or if the structure were any different, stretching the hydrogen bonds, energy levels wouldn't be able to stabilize.

(f) Eigenvalues of the RGM are more irregularly spaced at higher levels of graph connectivity, also their spacings become more significant and more different with increasing graph connectivity. Such observations indicate that when the proteins settles into stable structures, the energy levels form a specific stable pattern that might act as a proofreading mechanism to ensure specificity during molecular recognition.

(g) From the point of view of Heisenberg's uncertainty principle, if a measurement fixes the location of the electron, then necessarily its momentum (thereby its energy) is uncertain by the standards of that same measurement in the same frame of reference. The uncertainty principle is a fundamental property of waves, rather than a limitation of technological prowess, or observer. It suggests that upon molecular recognition, when two complementary molecules bind favorably, or when two disordered regions on two proteins fall into a stable struc-

ture together, a measurement happens.

(h) Such a measurement restricts the location of the electrons to within 5Å and only along specific LBHBs or single-welled hydrogen bonds, or dangling p orbitals, necessarily increasing the uncertainty in electron energies, leading to a wider energy eigenvalue distribution, corresponding to an increase in information entropy of the energy distribution and an overall loss of information. As discussed above, interconnected networks of electrons delocalizing along LBHB s through a protein, can relay energetic changes from one end to another almost instantly. The molecular recognition event on one end of the protein resulting in increasing information entropy of electron energies on that end, is quickly relayed to the other end of the protein where previously hidden and structured regions unfold momentarily, decreasing information entropy on that end, leading to a gain in information. The deduction might suggest a possible mechanism to investigate the nature of information flow in signal transduction cascades as a direct result of the Heisenberg Uncertainty Principle.(fig 5.1)

(i) Since the quantum mechanical properties of electrons, such as delocalization far enough to become entangled with other electrons in the molecule and the solvent, resulting in an entropic driving force for protein folding that compensates for reduction in conformal entropy; and a simultaneous stabilization of energy levels into irregular patterns, as well as enabling an information flow arrow across proteins participating in signal transduction cascades; depends entirely on and mediated by structured water networks forming low barrier hydrogen bonds with the proteins and single-well hydrogen bonds among the water molecules, it provides a candidate explanation for the unique necessity of water for life on earth.

Unfortunately none of the conjectures above have been experimentally tested in a

laboratory, though all of them are direct results from the random graph structured matrix treatment of protein folding, bolstered by common sense deductions from several peripheral experiments. The conjectures are simply a result of theoretical calculations of some metrics obtained from the random graph structured matrix representation of protein networks at the electronic scale. Quantum properties are notorious for their difficulty in being observed experimentally, and so we must rely on theoretical considerations until experimental observations can confidently and rigorously deny or corroborate the conjectures above.

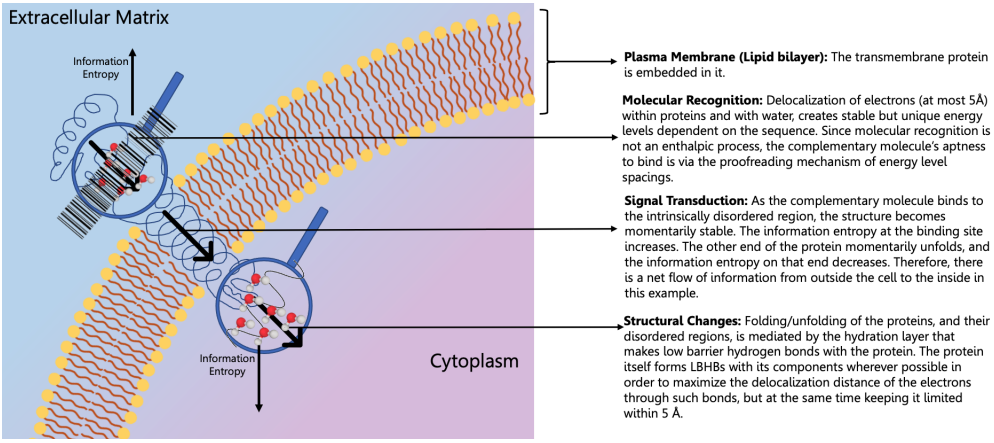


Figure 5.1: **Main physical inferences and conjectures from the RGM modeling of protein folding**:The figure shows the major conclusions and physical inferences that we obtain from the random graph matrix modeling of electron dynamics during protein folding. RGM modeling of the Hamiltonian of such interacting electrons offers several conjectures that might explain the entropy compensation during protein folding, specificity of molecular recognition events, and triggering of signal transduction cascades with clear information flow. The text within the figure reads as follows: (A) **Plasma Membrane (Lipid bilayer)**: The transmembrane protein is embedded in it. (b) **Structural Changes**: Folding/unfolding of the proteins, and their disordered regions, is mediated by the hydration layer that makes low barrier hydrogen bonds with the protein. The protein itself forms LBHBs with its components wherever possible in order to maximize the delocalization distance of the electrons through such bonds, but at the same time keeping it limited within 5Å. (c) **Molecular Recognition**: Delocalization of electrons (at most 5Å) within proteins and with water, creates stable but unique energy levels dependent on the sequence. Since molecular recognition is not an enthalpic process, the complementary molecule's aptness to bind is via the proofreading mechanism of energy level spacings.(d) **Signal Transduction**: As the complementary molecule binds to the intrinsically disordered region, the structure becomes momentarily stable. The information entropy at the binding site increases. The other end of the protein momentarily unfolds, and the information entropy on that end decreases. Therefore, there is a net flow of information from outside the cell to the inside in this example.

BIBLIOGRAPHY

[1] Shimin Zhao, Wei Xu, Wenqing Jiang, Wei Yu, Yan Lin, Tengfei Zhang, Jun Yao, Li Zhou, Yaxue Zeng, Hong Li, Yixue Li, andWenlin An Jiong Shi, Susan M Hancock, Fuchu He, Lunxiu Qin, Jason Chin, Pengyuan Yang, Xian Chen, Qunying Lei, Yue Xiong, and Kun-Liang Guan. Regulation of cellular metabolism by protein lysine acetylation. *Science*, 327, 2010.

[2] The role of cellular prion protein in lipid metabolism in the liver. *Prion*, 14, 2020.

[3] Rune Kleppe, Aurora Martinez, Stein Ove Døskeland, and Jan Haavik. The 14-3-3 proteins in regulation of cellular metabolism. *Seminars in Cell and Developmental Biology*, 22, 2011.

[4] Mitchell D. Knutson. Iron transport proteins: Gateways of cellular and systemic iron homeostasis. *Journal of Biological Chemistry*, 292, 2017.

[5] Laura Urrea, Miriam Segura-Feliu, Masami Masuda-Suzukake, Arnau Hervera, Lucas Pedraz, José Manuel García Aznar, Miquel Vila, Josep Samitier, Eduard Torrents, Isidro Ferrer, Rosalina Gavín, Masato Hagesawa, and José Antonio del Río. Involvement of cellular prion protein in -synuclein transport in neurons. *Molecular Neurobiology*, 55, 2018.

[6] Joseph G.S. Tsun, Susan Yung, Mel K.M. Chau, Sammy W.M. Shiu, Tak Mao Chan, and Kathryn C.B. Tan. Cellular cholesterol transport proteins in diabetic nephropathy. *PLoS ONE*, 9, 2014.

[7] Nobutaka Hirokawa and Yosuke Tanaka. Kinesin superfamily proteins (kifs): Various functions and their relevance for important phenomena in life and diseases. *Experimental Cell Research*, 334, 2015.

[8] Thomas A Guilliam and Joseph TP Yeeles. The eukaryotic replisome tolerates leading-strand base damage by replicase switching. *The EMBO Journal*, 40, 2021.

[9] Aaron J. Oakley. A structural view of bacterial dna replication. *Protein Science*, 28, 2019.

[10] Jeffrey L. Hansen, Alexander M. Long, and Steve C. Schultz. Structure of the rna-dependent rna polymerase of poliovirus. *Structure*, 5, 1997.

[11] Aaron Johnson and Mike O'Donnell. Cellular dna replicases: Components and dynamics at the replication fork. *Annual Review of Biochemistry*, 74, 2005.

[12] F. William Studier and Barbara A. Moffatt. Use of bacteriophage t7 rna polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology*, 189, 1986.

[13] Claudio Hetz and Feroz R. Papa. The unfolded protein response and cell fate control. *Molecular Cell*, 69, 2018.

[14] Stephen Jun Fei Chong, Saverio Marchi, Giulia Petroni, Guido Kroemer, Lorenzo Galluzzi, and Shazib Pervaiz. Noncanonical cell fate regulation by bcl-2 proteins. *Trends in Cell Biology*, 30, 2020.

[15] Asmat Ullah Khan, Rongmei Qu, Jun Ouyang, and Jingxing Dai. Role of nucleoporins and transport receptors in cell differentiation. *Frontiers in Physiology*, 11, 2020.

[16] Sangbin Lim, Joshua B. Phillips, Luciana Madeira Da Silva, Ming Zhou, Oystein Fodstad, Laurie B. Owen, and Ming Tan. Interplay between immune checkpoint proteins and cellular metabolism. *Cancer Research*, 77, 2017.

[17] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293, 2001.

[18] Drew M. Pardoll. The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer*, 12, 2012.

[19] Susan Elmore. Apoptosis: A review of programmed cell death. *Toxicologic Pathology*, 35, 2007.

[20] David Ron and Peter Walter. Signal integration in the endoplasmic reticulum unfolded protein response. *Nature Reviews Molecular Cell Biology*, 8, 2007.

[21] J. Craig Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Yuan Wang, A. Wang, X. Wang, J. Wang, M. H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. Lai Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Yu H.

Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. Ni Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291, 2001.

[22] Tsai Wen Chen, Trevor J. Wardill, Yi Sun, Stefan R. Pulver, Sabine L. Renninger, Amy Baohan, Eric R. Schreiter, Rex A. Kerr, Michael B. Orger, Vivek Jayaraman, Loren L. Looger, Karel Svoboda, and Douglas S. Kim. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499, 2013.

[23] D. Grahame Hardie, Fiona A. Ross, and Simon A. Hawley. Ampk: A nutrient and energy sensor that maintains energy homeostasis. *Nature Reviews Molecular Cell Biology*, 13, 2012.

[24] Osamu Takeuchi and Shizuo Akira. Pattern recognition receptors and inflammation. *Cell*, 140, 2010.

[25] Kai Simons and Derek Toomre. Lipid rafts and signal transduction. *Nature Reviews Molecular Cell Biology*, 1, 2000.

[26] Shinsuke Uda and Shinya Kuroda. Analysis of cellular signal transduction from an information theoretic approach. *Seminars in Cell and Developmental Biology*, 51, 2016.

[27] Stephen D. Roper. Signal transduction and information processing in mammalian taste buds. *Pflugers Archiv European Journal of Physiology*, 454, 2007.

[28] Gerald M. Cohen. Caspases: The executioners of apoptosis. *Biochemical Journal*, 326, 1997.

[29] L. Chang and M. Karin. Mammalian map kinase signalling cascades. *Nature*, 410, 2001.

[30] Yigong Shi and Joan Massagué. Mechanisms of tgf- signaling from cell membrane to the nucleus. *Cell*, 113, 2003.

[31] Tatsuaki Tsuruyama. Information thermodynamics of the cell signal transduction as a szilard engine. *Entropy*, 20, 2018.

[32] Tatsuaki Tsuruyama. The conservation of average entropy production rate in a model of signal transduction: Information thermodynamics based on the fluctuation theorem. *Entropy*, 20, 2018.

[33] Siobhan S. Mc Mahon, Aaron Sim, Sarah Filippi, Robert Johnson, Juliane Liepe, Dominic Smith, and Michael P.H. Stumpf. *Information theory and signal transduction systems: From molecular information processing to network inference*, volume 35. 2014.

[34] Boris N. Kholodenko, Jan B. Hoek, Hans V. Westerhoff, and Guy C. Brown. Quantification of information transfer via cellular signal transduction pathways. *FEBS Letters*, 414, 1997.

[35] Jean-Marie -M Lehn. Perspectives in supramolecular chemistry—from molecular recognition towards molecular information processing and self-organization. *Angewandte Chemie International Edition in English*, 29, 1990.

[36] Pedro C. Marijuán and Jorge Navarro. From molecular recognition to the "ve-

hicles" of evolutionary complexity: An informational approach. *International Journal of Molecular Sciences*, 22, 2021.

[37] Pedro C. Marijuán and Jorge Navarro. The biological information flow: From cell theory to a new evolutionary synthesis. *BioSystems*, 213, 2022.

[38] Brigitte L. Kieffer. Recent advances in molecular recognition and signal transduction of active peptides: Receptors for opioid peptides. *Cellular and Molecular Neurobiology*, 15, 1995.

[39] Crystal Nguyen, Takeshi Yamazaki, Andriy Kovalenko, David A. Case, Michael K. Gilson, Tom Kurtzman, and Tyler Luchko. A molecular reconstruction approach to site-based 3d-rism and comparison to gist hydration thermodynamic maps in an enzyme active site. *PLoS ONE*, 14, 2019.

[40] Jean Marie Lehn. From supramolecular chemistry towards constitutional dynamic chemistry and adaptive chemistry. *Chemical Society Reviews*, 36, 2007.

[41] Yaakov Levy and José N. Onuchic. Water mediation in protein folding and molecular recognition. *Annual Review of Biophysics and Biomolecular Structure*, 35, 2006.

[42] Thomas G.W. Edwardson, Kai Lin Lau, Danny Bousmail, Christopher J. Serpell, and Hanadi F. Sleiman. Transfer of molecular recognition information from dna nanostructures to gold nanoparticles. *Nature Chemistry*, 8, 2016.

[43] Georg Kustatscher, Piotr Grabowski, Tina A. Schrader, Josiah B. Passmore, Michael Schrader, and Juri Rappsilber. Co-regulation map of the human proteome enables identification of protein functions. *Nature Biotechnology*, 37, 2019.

[44] Hongbo Xie, Slobodan Vucetic, Lilia M. Iakoucheva, Christopher J. Oldfield, A. Keith Dunker, Zoran Obradovic, and Vladimir N. Uversky. Functional anthology of intrinsic disorder. 3. ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *Journal of Proteome Research*, 6, 2007.

[45] Slobodan Vucetic, Hongbo Xie, Lilia M. Iakoucheva, Christopher J. Oldfield, A. Keith Dunker, Zoran Obradovic, and Vladimir N. Uversky. Functional anthology of intrinsic disorder. 2. cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *Journal of Proteome Research*, 6, 2007.

[46] Andrew P. Marsden, Jeffrey J. Hollins, Charles O'Neill, Pavel Ryzhov, Sally Higson, Carolina A.T.F. Mendonça, Tristan O. Kwan, Lee Gyan Kwa, Annette Steward, and Jane Clarke. Investigating the effect of chain connectivity on the folding of a beta-sheet protein on and off the ribosome. *Journal of Molecular Biology*, 430, 2018.

[47] Hao Wang, David P. Fewer, Liisa Holm, Leo Rouhiainen, and Kaarina Sivonen. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 2014.

[48] IUPAC-IUB Commission on Biochemical Nomenclature. *Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains: Tentative Rules (1969)*, volume 17. 1970.

[49] IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. *Journal of Biological Chemistry*, 245, 1970.

[50] Maciej Stawikowski and Gregg B Fields. Introduction to peptide synthesis. current protocols in protein science. *Curr. Protoc. Protein Sci.*, 26, 2002.

[51] Giulietta Smulevich and Filomena Sica. Biopolymers - peptide science: Introduction. *Biopolymers - Peptide Science Section*, 91, 2009.

[52] Vladimir N. Uversky, Christopher J. Oldfield, and A. Keith Dunker. Showing your id: Intrinsic disorder as an id for recognition, regulation and cell signaling. *Journal of Molecular Recognition*, 18, 2005.

[53] Ramsy Agha, Samuel Cirés, Lars Wörmer, and Antonio Quesada. Limited

stability of microcystins in oligopeptide compositions of microcystis aeruginosa (cyanobacteria): Implications in the definition of chemotypes. *Toxins*, 5, 2013.

[54] Patrick Argos. An investigation of oligopeptides linking domains in protein tertiary structures and possible candidates for general gene fusion. *Journal of Molecular Biology*, 211, 1990.

[55] Ali Adem Bahar and Dacheng Ren. Antimicrobial peptides. *Pharmaceuticals*, 6, 2013.

[56] Mariam Rima, Mohamad Rima, Ziad Fajloun, Jean Marc Sabatier, Burkhard Bechinger, and Thierry Naas. Antimicrobial peptides: A potent alternative to antibiotics. *Antibiotics*, 10, 2021.

[57] Maurice R. Elphick, Olivier Mirabeau, and Dan Larhammar. Evolution of neuropeptide signalling systems. *Journal of Experimental Biology*, 221, 2018.

[58] Yanan Zheng, Linlin Zhang, Junxia Xie, and Limin Shi. The emerging role of neuropeptides in parkinson's disease. *Frontiers in Aging Neuroscience*, 13, 2021.

[59] Xin Yi Chen, Yi Feng Du, and Lei Chen. Neuropeptides exert neuroprotective effects in alzheimer's disease. *Frontiers in Molecular Neuroscience*, 11, 2019.

[60] Pingli Wei, Caitlin Keller, and Lingjun Li. Neuropeptides in gut-brain axis and their influence on host immunity and stress. *Computational and Structural Biotechnology Journal*, 18, 2020.

[61] Ian W. Hamley. Lipopeptides: From self-assembly to bioactivity. *Chemical Communications*, 51, 2015.

[62] Beth A. Lazazzera. The intracellular function of extracellular signaling peptides. *Peptides*, 22, 2001.

[63] Robert C. Gensure, Thomas J. Gardella, and Harald Jüppner. Parathyroid hormone and parathyroid hormone-related peptide, and their receptors. *Biochemical and Biophysical Research Communications*, 328, 2005.

[64] Richard J. Lewis and Maria L. Garcia. Therapeutic potential of venom peptides. *Nature Reviews Drug Discovery*, 2, 2003.

[65] Slavica Krantic. Peptides as regulators of the immune system: Emphasis on somatostatin. *Peptides*, 21, 2000.

[66] D. W. Green, G. Gomez, and G. H. Greeley. Gastrointestinal peptides. *Gastroenterology Clinics of North America*, 18, 1989.

[67] J. M. Polak and S. R. Bloom. Regulatory peptides of the gastrointestinal and respiratory tracts. *Archives Internationales de Pharmacodynamie et de Therapie*, 280, 1986.

[68] Michael S. Simonson. Endothelins: Multifunctional renal peptides. *Physiological Reviews*, 73, 1993.

[69] T. Simonson, D. Perahia, and A. T. Brünger. Microscopic theory of the dielectric properties of proteins. *Biophysical Journal*, 59, 1991.

[70] Sebastiaan van Heesch, Franziska Witte, Valentin Schneider-Lunitz, Jana F. Schulz, Eleonora Adami, Allison B. Faber, Marieluise Kirchner, Henrike Maatz, Susanne Blachut, Clara Louisa Sandmann, Masatoshi Kanda, Catherine L. Worth, Sebastian Schafer, Lorenzo Calviello, Rhys Merriott, Giannino Patone, Oliver Hummel, Emanuel Wyler, Benedikt Obermayer, Michael B. Mücke, Eric L. Lindberg, Franziska Trnka, Sebastian Memczak, Marcel Schilling, Leanne E. Felkin, Paul J.R. Barton, Nicholas M. Quaife, Konstantinos Vanezis, Sebastian Diecke, Masaya Mukai, Nancy Mah, Su Jun Oh, Andreas Kurtz, Christoph Schramm, Dorothee Schwinge, Marcial Sebode, Magdalena Harakalova, Folkert W. Asselbergs, Aryan Vink, Roel A. de Weger, Sivakumar Viswanathan, Anissa A. Widjaja, Anna Gärtner-Rommel, Hendrik Milting, Cris dos Remedios, Christoph Knosalla, Philipp Mertins, Markus Landthaler, Martin Vingron, Wolfgang A. Linke, Jonathan G. Seidman, Christine E. Seidman, Nikolaus Rajewsky, Uwe Ohler, Stuart A. Cook, and Norbert Hubner. The translational landscape of the human heart. *Cell*, 178, 2019.

[71] Nicole A. Brooks, Dodie S. Pouniotis, Choon Kit Tang, Vasso Apostolopoulos, and Geoffrey A. Pietersz. Cell-penetrating peptides: Application in vaccine delivery. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1805, 2010.

[72] Tomas Ganz and Robert I. Lehrer. Antibiotic peptides from higher eukaryotes: Biology and applications. *Molecular Medicine Today*, 5, 1999.

[73] Annica Carolin Staudt and Stephan Wenkel. Regulation of protein function by 'microproteins'. *EMBO Reports*, 12, 2011.

[74] Qingqing Wu, Shangwei Zhong, and Hui Shi. Microproteins: Dynamic and accurate regulation of protein activity. *Journal of Integrative Plant Biology*, 64, 2022.

[75] Kaushal Kumar Bhati, Valdeko Kruusvee, Daniel Straub, Anil Kumar Nalini Chandran, Ki Hong Jung, and Stephan Wenkel. Global analysis of cereal microproteins suggests diverse roles in crop development and environmental adaptation. *G3: Genes, Genomes, Genetics*, 10, 2020.

[76] M Feughelman, D Lyman, E Menefee, and B Willis. The orientation of the alpha-helices in alpha-keratin fibres. *International Journal of Biological Macromolecules*, 33, 2003.

[77] Catherine A. Makarewich. The hidden world of membrane microproteins. *Experimental Cell Research*, 388, 2020.

[78] Debby D. Wang, Moon Tong Chan, and Hong Yan. Structure-based protein–ligand interaction fingerprints for binding affinity prediction. *Computational and Structural Biotechnology Journal*, 19, 2021.

[79] Jason E. Hein and Donna G. Blackmond. On the origin of single chirality of amino acids and sugars in biogenesis. *Accounts of Chemical Research*, 45, 2012.

[80] Jumpei Sasabe and Masataka Suzuki. Distinctive roles of d-amino acids in the homochiral world: Chirality of amino acids modulates mammalian physiology and pathology. *Keio Journal of Medicine*, 68, 2019.

[81] Manibarathi Vaithiyanathan, Hannah C. Hymel, Nora Safa, Olivia M. Sanchez,

Jacob H. Pettigrew, Cole S. Kirkpatrick, Ted J. Gauthier, and Adam T. Melvin. Kinetic analysis of cellular internalization and expulsion of unstructured d-chirality cell penetrating peptides. *AIChE Journal*, 67, 2021.

[82] Marie Claire Bellissent-Funel, Ali Hassanali, Martina Havenith, Richard Henchman, Peter Pohl, Fabio Sterpone, David Van Der Spoel, Yao Xu, and Angel E. Garcia. Water determines the structure and dynamics of proteins. *Chemical Reviews*, 116, 2016.

[83] Garegin A. Papoian, Johan Ulander, Michael P. Eastwood, Zaida Luthey-Schulten, and Peter G. Wolynes. Water in protein structure prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 2004.

[84] Matthias Heyden. Heterogeneity of water structure and dynamics at the protein-water interface. *Journal of Chemical Physics*, 150, 2019.

[85] Bijan Ranjbar and Pooria Gill. Circular dichroism techniques: Biomolecular and nanostructural analyses- a review. *Chemical Biology and Drug Design*, 74, 2009.

[86] Sharon M. Kelly, Thomas J. Jess, and Nicholas C. Price. How to study proteins by circular dichroism. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1751, 2005.

[87] Lee Whitmore and B. A. Wallace. Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. *Biopolymers*, 89, 2008.

[88] Brent L. Nannenga, Matthew G. Iadanza, Breanna S. Vollmar, and Tamir Gonen. Overview of electron crystallography of membrane proteins: Crystallization and screening strategies using negative stain electron microscopy. *Current Protocols in Protein Science*, 2013.

[89] Ali A. Kermani. A guide to membrane protein x-ray crystallography. *FEBS Journal*, 288, 2021.

[90] Laurent Maveyraud and Lionel Mourey. Protein x-ray crystallography and drug discovery. *Molecules*, 25, 2020.

[91] Ashutosh Srivastava, Tetsuro Nagai, Arpita Srivastava, Osamu Miyashita, and Florence Tama. Role of computational methods in going beyond x-ray crystallography to explore protein structure and dynamics. *International Journal of Molecular Sciences*, 19, 2018.

[92] R. Andrew Atkinson. Nmr of proteins and nucleic acids. *Nuclear Magnetic Resonance*, 46, 2021.

[93] Nicholas J. Fowler, Adnan Sljoka, and Mike P. Williamson. A method for validating the accuracy of nmr protein structures. *Nature Communications*, 11, 2020.

[94] Amit Kumar, Lars T. Kuhn, and Jochen Balbach. In-cell nmr: Analysis of protein–small molecule interactions, metabolic processes, and protein phosphorylation. *International Journal of Molecular Sciences*, 20, 2019.

[95] Yunfei Hu, Kai Cheng, Lichun He, Xu Zhang, Bin Jiang, Ling Jiang, Conggang Li, Guan Wang, Yunhuang Yang, and Maili Liu. Nmr-based methods for protein analysis. *Analytical Chemistry*, 93, 2021.

[96] Helen M. Berman, Gerard J. Kleywegt, Haruki Nakamura, and John L. Markley. The protein data bank archive as an open data resource. *Journal of Computer-Aided Molecular Design*, 28, 2014.

[97] Naomi Zurgil, Elena Afrimzon, Assaf Deutsch, Yaniv Namer, Yana Shafran, Maria Sobolev, Yishay Tauber, Orit Ravid-Hermesh, Mordechai Deutsch, Nicholas a Zumwalde, Eisuke Domae, Matthew F Mescher, Yoji Shimizu, Linda a Zuckerman, Lara Pullen, Jim John H Jennifer Miller, Weiping Zou, Anton Zilman, Vitaly V Ganusov, Alan S. Perelson, Christina E Zielinski, Davide Corti, Federico Mele, Dora Pinto, Antonio Lanzavecchia, Federica Sallusto, Jinfang Zhu, Hidehiro Yamane, William E Paul, He Zhu, Gulnaz Stybayeva, Monica Macal, Erlan Ramanculov, Michael D George, Satya Dan-

dekar, Alexander Revzin, Liang Zhou, Ivaylo I Ivanov, Rosanne Spolski, Roy Min, Kevin Shenderov, Takeshi Egawa, David E Levy, Warren J. Leonard, Dan R Littman, Mark M W Chong, Dan R Littman, Fang Zhao, Jennifer L. Cannons, Mala Dutta, Gillian M. Griffiths, Pamela L. Schwartzberg, Xinmin Zhang, S Sun, I Hwang, David F Tough, Jonathan Sprent, Qianqian Zhang, Fadi G. Lakkis, Ming Zeng, Mirko Paiardini, Jessica C Engram, Greg J Beilman, Jeffrey G Chipman, Timothy W Schacker, Guido Silvestri, Ashley T Haase, Lauren a Zenewicz, Andrey Antov, Richard a Flavell, Irina Zaretsky, Michal Polonsky, Eric Shifrut, Shlomit Reich-Zeliger, Yaron E Antebi, Guy Aidelberg, Nir Waysbort, Nir Friedman, Mary a Yui, Leslie L Sharp, Wendy L Havran, Ellen V Rothenberg, G Hernández-Hoyos, Ellen V Rothenberg, Ben a. Youngblood, J. Scott Hale, Haydn T. Kissick, Eunseon Ahn, Xiaojin Xu, Andreas Wieland, Koichi Araki, Erin E. West, Hazem E. Ghoneim, Yiping Fan, Pranay Dogra, Carl W. Davis, Bogumila T. Konieczny, Rustom Antia, Xiaodong Cheng, Rafi Ahmed, Nir Yosef, Alex K Shalek, Jellert T Gaublomme, Hyun Tak Hulin Jin, Youjin Lee, Amit Awasthi, Catherine J. Chuan Wu, Katarzyna Karwacz, Sheng Xiao, Marsela Jorgolli, Dave David Gennert, Rahul Satija, Arvind Shakya, Diana Y Lu, John J Trombetta, Meenu R Pillai, Peter J Ratcliffe, Mathew L Coleman, Mark Bix, Dean Tantin, Hongkun Park, Vijay K Kuchroo, Aviv Regev, Andrew Yates, Robin E Callard, Jaroslav Stark, Martin L Yarmush, Kevin R King, Li Yang, Yi Zhang, Feng Yang, Katrina M Waters, Jim John H Jennifer Miller, Marina A Gritsenko, Rui Zhao, Xiuxia Du, Eric A Livesay, Samuel O Purvine, Matthew E Monroe, Yingchun Wang, David G Camp, Rex-Neal Richard D Smith, David L Stenoien, Yvonne J. Yamanaka, Christoph T. Berger, Magdalena Sips, Patrick C. Cheney, Galit Alter, J. Christopher Love, Shohei Yamamura, Hiroyuki Kishi, Yoshiharu Tokimitsu, Sachiko Kondo, Ritsu Honda, Sathuluri Ramachandra Rao, Masahiro Omori, Eiichi Tamiya, Atsushi Muraguchi, Satoshi Yamada, Jun Tsukada, Akihiko

Yoshimura, Masato Kubo, , Nikolai A. Raffler Klaus Ley Xitong Dang, Jianming Xie, Cristina M. Tato, Mark M. Davis, Eilon Woolf, Irina Grigorova, Adi Sagiv, Valentin Grabovsky, Sara W Feigelson, Ziv Shulman, Tanja Hartmann, Michael Sixt, Jason G Cyster, Ronen Alon, Linda Wooldridge, Julia Ekeruche-Makinde, Hugo a. van den Berg, Anna Skowera, John J. Miles, Mai Ping Tan, Garry Dolton, Mathew Clement, Sian Llewellyn-Lacey, David a. Price, Mark Peakman, Andrew K. Sewell, Yochai Wolf, Anat Shemer, Michal Polonsky, Mor Gross, Alexander Mildner, Simon Yona, Eyal David, Ki-Wook Karen S. Kyu Tae Kim, Tobias Goldmann, Ido Amit, Mathias Heikenwalder, Sergei Nedospasov, Marco Prinz, Nir Friedman, Steffen Jung, Donald Wlodkowic, Joanna Skommer, Dagmara McGuinness, Shannon Faley, Walter Kolch, Zbigniew Darzynkiewicz, Jonathan M. Cooper, Michele Zagnoni, John P Wikswo, Jonathan M. Cooper, Nicholas J Wilson, Katia Boniface, Jason R Chan, Brent S McKenzie, Wendy M Blumenschein, Jeanine D Mattson, Beth Basham, Kathleen Karen M Kendall A Smith, Taiying Chen, Franck Morel, Jean-Claude Lecron, Robert a Kastelein, Daniel J Cua, Terrill K McClanahan, Edward P Bowman, Rene de Waal Malefyt, M. a. Williams, Michael J. Bevan, George M Whitesides, E Ostuni, S Takayama, X Jiang, D E Ingber, T. L. Whiteside, Cameron J. Wellard, Gabrielle T Belz, Andrew M Lew, Mark R. Dowling, Huibin Wei, Bor han Chueh, Huiling Wu, Eric W Hall, Cheuk wing Li, Romana Schirhagl, Jin-Ming Jian-Xin Xin Lin, Richard N Zare, K. Scott Weber, Q.-J. Qingsheng Li, Stephen P. Persaud, J. D. Campbell, Mark M. Davis, Paul M. Allen, Jason M Weaver, Francisco a Chaves, Andrea J Sant, Nir Waysbort, Dor Russ, Benjamin M Chain, Nir Friedman, Jennifer C Waters, Ruoning Wang, Douglas R. Green, Ninghai Wang, Marton Keszei, Peter J. Halibozek, Burcu Yigit, Pablo Engel, Cox Terhorst, Hui Hongyu Zhao, Michael S. O'Keeffe, Peter T. Sage, Arlene H. Sharpe, Cox Terhorst, Yisong Y Wan, Richard a Flavell, Lucy S K Walker, Abul K. Abbas, Dana Vuzman, Michal

Polonsky, Yaakov Levy, Guillaume Voisinne, Briana G Nixon, Anna Melbinger, Georg Gasteiger, Massimo Vergassola, Grégoire Gregoire Altan-Bonnet, David Voehringer, Kanade Shinkai, Richard M. Locksley, San Francisco, Dario a a Vignali, Lauren W Collison, Creg J Workman, Fischer Verlag, N Varadarajan, B Julg, Yvonne J. Yamanaka, H Chen, Adebola O Ogunniyi, E McAndrew, L C Porter, A Piechocka-Trocha, B J Hill, Daniel C. Douek, F Pereyra, B D Walker, J. Christopher Love, Klaas P J M van Gisbergen, Paul L. Klarenbeek, Natasja a M Kragten, Peter Paul a Unger, Marieke B B Nieuwenhuis, Felix M. Wensveen, Anja ten Brinke, Paul P. Tak, Eric Eldering, Martijn a. Nolte, Rene a W van Lier, Henk-Jan van den Ham, Rob J de Boer, Rieneke van de Ven, Mariska C de Jong, Anneke W Reurs, Antoinet J N Schoonderwoerd, Gerrit Jansen, Jan H Hooijberg, George L Scheffer, Tanja D de Gruijl, Rik J Scheper, V L Tybulewicz, C E Crawford, P K Jackson, R T Bronson, R C Mulligan, Noah J. Tubo, Antonio J. Pagán, Justin J. Taylor, Ryan W. Nelson, Jonathan L. Linehan, James M. Ertelt, Eric S. Huseby, Sing Sing Way, Marc K. K Jenkins, Parul Tripathi, Naresh Sahoo, Ubaid Ullah, Henna Kallionpää, Amita Suneja, Riitta Lahesmaa, Kanury V S Rao, Bettina Trinschek, Felix Lüssi, Jürgen Haas, Brigitte Wildemann, Frauke Zipp, Heinz Wiendl, Christian Becker, Helmut Jonuleit, Cole Trapnell, Lior Pachter, Steven L. Salzberg, Yoshiharu Tokimitsu, Hiroyuki Kishi, Sachiko Kondo, Ritsu Honda, Kazuto Tajiri, Kazumi Motoki, Tatsuhiko Ozawa, Shinichi Kadowaki, Tsutomu Obata, Satoshi Fujiki, Chise Tateno, Hideki Takaishi, Kazuaki Chayama, Katsutoshi Yoshizato, Eiichi Tamiya, Toshiro Sugiyama, Atsushi Muraguchi, Karen E Tkach, Debashis Barik, Guillaume Voisinne, Nicole Malandro, Matthew M Hathorn, Jesse W Cotari, Robert Vogel, Taha Merghoub, Jedd D. Wolchok, Oleg Krichevsky, Grégoire Gregoire Altan-Bonnet, Kevin Thurley, Daniel Gerecht, Elfriede Friedmann, Thomas Höfer, Tadatsugu Taniguchi, Yasuhiro Minami, Shinya Tanaka, Yasutaka Motomura, Yoshie Suzuki, Ryoji Yagi, Hiromasa Inoue, Shoichiro

Miyatake, Masato Kubo, Yodai Takei, Sheel Shah, Sho Harvey, Lei S. Qi, Long Cai, K. Takeda, T. Kaisho, Nobuya Yoshida, J. Takeda, K. M Kishimoto, S. Akira, Madhusudhanan Sukumar, Jie Liu, Yun Ji, Murugan Subramanian, Joseph G Crompton, Zhiya Yu, Rahul Roychoudhuri, Douglas C Palmer, Pawel Muranski, Edward D Karoly, Robert P Mohney, Christopher A Klebanoff, Ashish Lal, Toren Finkel, Nicholas P. Restifo, Luca Gattinoni, Vijay G Subramanian, Ken R. Duffy, Marian L Turner, Philip D. Hodgkin, Rita L Strack, Daniel E Strongin, Dibyendu Bhattacharyya, Wen Tao, Allison Berman, Hal E Broxmeyer, Robert J Keenan, Benjamin S Glick, Daniel B Stetson, Markus Mohrs, R Lee Reinhardt, Jody L Baron, Zhi-En Wang, Laurent Gapin, Mitchell Kronenberg, Richard M. Locksley, Christian Stemberger, Katharina M. Huster, Martina Koffler, Florian Anderl, Matthias Schiemann, Hermann Wagner, Dirk H. Busch, Patricia Graef, Marcus Odendahl, Julia Albrecht, Georg Dössinger, Florian Anderl, Veit R. Buchholz, Georg Gasteiger, Matthias Schiemann, Götz U. Grigoleit, Friedhelm R. Schuster, Arndt Borkhardt, Birgitta Versluys, Torsten Tonn, Erhard Seifried, Hermann Einsele, Lothar Germeroth, Dirk H. Busch, Michael Neuenhahn, Timothy a Springer, Michael L. Dustin, G J Spangrude, F Sacchi, H R Hill, D E Van Epps, R a Daynes, Qing Song, Qing Han, Elizabeth M. Bradshaw, Sally C Kent, Khadir Raddassi, Gerald T Nepom, David a. Hafler, J. Christopher Love, Berend Snijder, Raphael Sacher, Pauli Rämö, Eva-Maria Damm, Prisca Liberali, Lucas Pelkmans, Kathleen Karen M Kendall A Smith, Joanne M Davidson, Paul Garside, Alison M Skelley, Oktay Kirak, Heikyung Suh, Rudolf Jaenisch, Joel Voldman, Michael Sixt, Federica Sallusto, Suzanne Ostrand-rosenberg, Andrea M. Siegel, Jennifer Heimall, Alexandra F. Freeman, Amy P. Hsu, Erica Brittain, Jason M. Brenchley, Daniel C. Douek, Gary H. Fahle, Jeffrey I. Cohen, Steven M. Holland, Joshua D. Milner, Raz Shimoni, Kim Pham, Mohammed Yassin, Min Gu, Sarah M. Russell, Eric Shifrut, Keyue Shen, V Kaye Thomas, Michael L. Dustin, Lance C

Kam, Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Y Lu, John J Trombetta, Dave David Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z Levin, Hongkun Park, Aviv Regev, Sheel Shah, Eric Lubeck, Wen Zhou, Long Cai, Maayan Schwarzkopf, Ting-Fang He, Alon Greenbaum, Chang Ho Sohn, Antti Lignell, Harry M. T. Choi, Viviana Gradinaru, Niles a. Pierce, Long Cai, Shou Serizawa, Kazunari Miyamichi, Hitoshi Sakano, Seila Selimović, Francesco Piraino, Hojae Bae, Marco Rasponi, Alberto Redaelli, Ali Khademhosseini, Kimberly S. Schluns, Leo Lefrançois, Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, Albert Cardona, Yael S Schiffenbauer, Yael Kalma, Ella Trubniykov, Orit Gal-Garber, Lilach Weisz, Asaf Halamish, Michael Sister, Gideon Berke, Jürgen Scheller, Athena Chalaris, Dirk Schmidt-Arras, Stefan Rose-John, Alexander Scheffold, Kenneth M Murphy, Thomas Höfer, Nadine S. Schaadt, Juan C.L. Carlos López Alfonso, Ralf Schönmeyer, Anne Grote, Germain Forestier, Cédric Wemmert, Nicole Krönke, Mechthild Stoeckelhuber, Hans H. Kreipe, Haralampos Hatzikirou, Friedrich Feuerhake, Yonatan Savir, Nir Waysbort, Yaron E Antebi, Tsvi Tlusty, Nir Friedman, A Saparov, F H Wagner, R Zheng, J R Oliver, H Maeda, R D Hockett, Casey T Weaver, Federica Sallusto, Antonio Lanzavecchia, Koichi Araki, Rafi Ahmed, D Lenig, R Förster, M Lipp, Antonio Lanzavecchia, Asako Sakaue-Sawano, Hiroshi Kurokawa, Toshifumi Morimura, Aki Hanyu, Hiroshi Hama, Hatsuki Osawa, Saori Kashiwagi, Kiyoko Fukami, Takaki Miyata, Hiroyuki Miyoshi, Takeshi Imamura, Masaharu Ogawa, Hisao Masai, Atsushi Miyawaki, Catherine a. Sabatos, Junsang Doh, Sumone Chakravarti, Rachel S. Friedman, Priya G. Pandurangi, Aaron J. Tooley, Matthew F. Krummel, ,

Jeffrey C. Rathmell Ryan D. Michalek, Rachel L. Rutishauser, Gislâine a. Martins, Sergey Kalachikov, Anmol Chandele, Ian a. Parish, Eric Meffre, Joshy Jacob, Kathryn Calame, Susan M. Kaech, Markus Rottmar, Maria Håkanson, Michael Smith, Katharina Maniura-Weber, Robert Rothlein, D Marlin, Timothy a Springer, Michael L. Dustin, S D Marlin, Timothy a Springer, Grazisa Rossetti, Mark Collinge, Jeffrey R Bender, Raffaela Molteni, Ruggero Pardi, I L Ross, C M Browne, David A Hume, Adam Rosenthal, Andrew S Alice Macdonald, Joel Voldman, M. Rosas-Ballina, P. S. Olofsson, M. Ochani, S. I. Valdes-Ferrer, Y. a. Levine, C. Reardon, M. W. Tusche, V. a. Pavlov, U. Andersson, S. Chavan, T. W. Mak, K. J. Tracey, Michael S. Rooney, Sachet a. Shukla, Catherine J. Chuan Wu, Gad Getz, Nir Hacohen, Noga Ron-Harel, Daniel Santos, Jonathan M. Ghergurovich, Peter T. Sage, Anita Reddy, Scott B. Lovitch, Noah Dephoure, F. Kyle Satterstrom, Michal Sheffer, Jessica B. Spinelli, Steven Gygi, Joshua D. Rabinowitz, Arlene H. Sharpe, Marcia C. Haigis, S Romagnani, Jan C Rohr, Carmen Gerlach, Lianne Kok, Ton N M Schumacher, K a Roebuck, A Finnegan, I Rivière, M J Sunshine, Dan R Littman, Hannah Richards, M Paula Longhi, Kate Wright, Hannah Richards, M Paula Longhi, Kate Wright, Awen Gallimore, Ann Ager, Jacqueline R Rettig, Albert Folch, Nicholas P. Restifo, Luca Gattinoni, R Lee Reinhardt, A Khoruts, R Merica, T Zell, Marc K. K Jenkins, Steven L. Reiner, William C Adams, Keil J Regehr, Maribella Domenech, Justin T Koepsel, Kristopher C Carver, Stephanie J Ellison-Zelski, William L Murphy, Linda a Schuler, Elaine T Alarid, David J. Beebe, Brandon Razooky, Edgar Gutierrez, Valery H Terry, Celsa a Spina, Alex Groisman, Leor S Weinberger, Jonathan M Raser, Erin K O'Shea, Rajesh R. Rao, Q.-J. Qingsheng Li, Melanie R Gubbels Bupp, Protul a. Shrikant, Gwendalyn J. Randolph, Andreas Ramming, Katja Thümmler, Hendrik Schulze-Koops, Alla Skapenko, Jim Jianhua Qin, Nannan Ye, Xin Liu, Bingcheng Lin, Dong Qin, Younan Xia, George M Whitesides, Vesna Pulko, John S Davies, Carmine Mar-

tinez, Marion C Lanteri, Michael P Busch, Michael S Diamond, Kenneth Knox, Erin C Bush, Peter a Sims, Shripad Sinari, Dean Billheimer, Elias K Haddad, Kristy O Murray, Anne M Wertheimer, Janko Nikolich-Žugich, Gavin C. Preston, Carmen Feijoo-Carnero, Nick Schurch, Victoria H. Cowling, Doreen a. Cantrell, Michal Polonsky, Irina Zaretsky, Nir Friedman, Benjamin M Chain, Nir Friedman, Courtney R. R. Plumlee, Brian S. S Sheridan, Basak B. B Cicek, Leo Lefrançois, Joshua J. Obar, Sara L. Colpitts, Evan R. Jellison, W. Nicholas Haining, Leo Lefrancois, Kamal M. Khanna, Kim Pham, Raz Shimoni, Mandy J Ludford-Menting, Cameron J Nowell, Pavel Lobachevsky, Ze'ev Bomzon, Min Gu, Terence P Speed, C Jane McGlade, Sarah M. Russell, Faruk Sacirbegovic, Sarah M. Russell, Stephen P. Persaud, Chelsea R Parker, Wan-Lin Lo, K. Scott Weber, Paul M. Allen, Georgia Perona-Wright, Katja Mohrs, Katrin D Mayer, Markus Mohrs, Marion Pepper, Marc K. K Jenkins, Karin Pelka, Eicke Latz, Erika L. Edward J. Pearce, Matthew C. Walsh, Pedro J. Cejas, Gretchen M. Harms, Hao Shen, Li San Lu Wang, Russell G. Jones, Yongwon Choi, Erika L. Edward J. Pearce, William E Paul, Jinfang Zhu, Kristen E. Pauken, E. John Wherry, Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, David N Louis, Orit Rozenblatt-Rosen, Mario L Suvà, Aviv Regev, Bradley E Bernstein, Elodie Parzy, Sylvain Miraux, Jean michel Franconi, Eric Thiaudière, Drew M. M. Pardoll, N Parameswaran, R Suresh, V Bal, S Rath, A George, Wyming Lee Pang, Pushpa Pandiyan, Lixin Zheng, Satoru Sayaka Ishihara, Jennifer Reed, Michael J Lenardo, Tiago Paixão, Tiago P Carvalho, Dinis Pedro Calado, Jorge Carneiro, Ertugrul M Ozbudak, Mukund Thattai, Iren Kurtser, Alan D Grossman, Alexander van Oudenaarden, Patrick a. Ott, F. Stephen Hodi, Caroline Robert, John J O'Shea, William E Paul, Yasushi Onishi, Zoltan Fehervari, Tomoyuki Yamaguchi, Shimon Sakaguchi, Adebola O Ogunniyi, Craig M Story, Eliseo Papa, Eduardo Guillen,

J. Christopher Love, Joshua J. Obar, Leo Lefrançois, Kamal M. Khanna, Leo Lefrançois, Leo Lefrancois, John J O'Shea, Riitta Lahesmaa, Golnaz Vahedi, Arian Laurence, Yuka Kanno, Katherine Nutsch, Chyi Hsieh, Mizuki Nishino, Nikhil H. Ramaiya, Hiroto Hatabu, F. Stephen Hodi, Simone a. Nish, Dominik Schenten, Thomas F Wunderlich, Scott D. Pope, Yan Gao, Namiko Hoshi, Shuang Yu, Xiting Yan, Han Byoel Hae Ock Heung Kyu Lee, Lesley Pasman, Igor Brodsky, Brian Yordy, Hui Hongyu Zhao, Jens Bruning, Ruslan Medzhitov, Evan W Newell, Mark M. Davis, Taku Naito, Hirokazu Tanaka, Yoshinori Naoe, Ichiro Taniuchi, Ognjenka Nadazdin, Svjetlan Boskovic, Toru Murakami, Georges Tocco, Rex-Neal Richard D Smith, Robert B Colvin, David H Sachs, James Allan, Joren C Madsen, Tatsuo Kawai, a Benedict Cosimi, Gilles Benichou, By Erin Murphy, Kazuko Shibuya, Nancy Hosken, Peter Openshaw, Anne O Garra, Mainofl Kenneth, Davisfl Kenneth, Brian Munsky, Gregor Neuert, Alexander van Oudenaarden, Adriana M Mujal, Julia K Gilden, Audrey Gérard, Makoto Kinoshita, Matthew F. Krummel, Scott N. Mueller, Laura K. Mackay, R Mostoslavsky, N Singh, T Tenzen, M Goldmit, C Gabay, S Elizur, P Qi, B E Reubinoff, A. Chess, H Cedar, Y Bergman, Mathieu Morel, Denis Bartolo, Jean-Christophe Galas, Maxime Dahan, Vincent Studer, Amy E Moran, Keli L Holzapfel, Yan Xing, Nicole R Cunningham, Jonathan S Maltzman, Jennifer Punt, Kristin a Hogquist, Markus Mohrs, Kanade Shinkai, Katja Mohrs, Richard M. Locksley, Masayuki Mizui, Hitoshi Kikutani, Keiichi Mitsuyama, Michio Sata, Stefan Rose-John, D. M. Mitchell, E. V. Ravkov, M. a. Williams, K T Miner, M Croft, Michelle Mark J Miller, Sindy H Wei, Ian Parker, Michael D Cahalan, Ian Parker, Olga Safrina, Ian Parker, Michael D Cahalan, Ivar Meyvantsson, David J. Beebe, Jerome T Mettetal, Dale Muzzey, Juan M Pedraza, Ertugrul M Ozbudak, Alexander van Oudenaarden, R Mercado, S Vijh, S E Allen, K Kerksiek, I M Pilip, E G Pamer, Luis Mendoza, Fátima Pardo, Erik Meijering, Oleh Dzyubachyk, Ihor Smal, J P Meador, J J

Lech, S D Rice, J E Hose, J W Short, S D Rice, T K Collier, N L Scholz, T K Collier, N L Scholz, S E Campana, L Carlsson, G Duker, R S Kass, D M Roden, D Darbar, H Cheng, W J Lederer, E V Freund, a P Farrell, B a Block, a Di Maio, S Thompson, B a Block, H a Shiels, a P Farrell, M S Lipnick, B a Block, J Leroy, J Y Le Guennec, C H Orchard, E White, Martin Md, Kim Mt, Shan Q, Sompallae R, Xue Hh, Harty Jt, Badovinac Vp, K Kai Mckinstry, Tara M Strutt, Bianca Bautista, Wenliang Zhang, Yi Kuang, Andrea M Cooper, Susan L. Swain, Nicholas McGranahan, Charles Swanton, J Cooper McDonald, George M Whitesides, Maxine A Mcclain, Christopher T Culbertson, Stephen C Jacobson, Nancy L Allbritton, Christopher E Sims, J Michael Ramsey, Jewish Natl, H H McAdams, A Arkin, Polly Matzinger, Tirumalai Kamala, Frederick Masson, Martina Minnich, Moshe Olshansky, Ivan Bilic, Adele M Mount, Axel Kallies, Terence P Speed, Meinrad Busslinger, Stephen L Nutt, Gabrielle T Belz, David Masopust, Susan M. Kaech, E. John Wherry, Rafi Ahmed, Vaiva D. Vezys, Amanda L Marzo, Leo Lefrançois, Kimberly D Klonowski, Agnes Le Bon, Persephone Borrow, David F Tough, Leo Lefrançois, Agnes Le Bon, Persephone Borrow, David F Tough, Leo Lefrançois, H. Marvi, Chaohui Gong, Nick Gravish, Henry Astley, David L Hu, Daniel I Goldman, Fernando O. Martinez, Scott M Siamon Gordon, Janet G M Markle, Daniel N Frank, Steven Mortin-Toth, Charles E Robertson, Leah M Feazel, Ulrike Rolle-Kampczyk, Martin von Bergen, Kathy D McCoy, Andrew J Macpherson, Jayne S Danska, Joshua S Marcus, W French Anderson, Stephen R Quake, Judith N. Mandl, João P. Monteiro, Nienke Vrisekoop, Ronald N. N. Germain, Nicole Malandro, Sadna Budhu, Nicholas F. Kuhn, Chengyu Cailian Liu, Judith T. Murphy, Czrina Cortez, Hong Zhong, Xia Yang, Gabrielle Rizzuto, Grégoire Gregoire Altan-Bonnet, Taha Merghoub, Jedd D. Wolchok, Cindy S Ma, Philip D. Hodgkin, Stuart G Tangye, Yasmin a. Lyons, Sherry Y. Wu, Willem W. Overwijk, Keith a. Baggerly, Anil K. Sood, Chong T Luo, Wei Will Liao, Saida Dadi, Ahmed

Toure, Ming O Li, Hervé Luche, Odile Weber, Tata Nageswara Rao, Carmen Blum, Hans Jörg Fehling, Eric Lubeck, Ahmet F. Coskun, Timur Zhiyentayev, Mubhij Ahmad, Long Cai, Michael I Love, Wolfgang Huber, Simon Anders, J. Christopher Love, Jehnna L Ronan, Gijsbert M Grotenbreg, Annemarthe G van der Veen, Hidde L Ploegh, M. Long, a. J. Adler, C a London, Abul K. Abbas, A Kelso, Inés Llaudó, Linda Cassis, Joan Torras, Oriol Bestard, Marcel La Franquesa, Josep M Cruzado, Gema Cerezo, Esther Castaño, Jordi Petriz, Immaculada Herrero-Fresneda, Josep M Grinyó, Núria Lloberas, Jean Livet, Tamily a Weissman, Hyuno Kang, Ryan W Draft, Ju Lu, Robyn a Bennis, Joshua R Sanes, Jeff W Lichtman, Zhiduo Liu, Michael Y. Gerner, Nicholas Van Panhuys, Andrew G. Arnold J Levine, Alexander Y. Rudensky, Ronald N. N. Germain, Pentao Liu, Nancy a. Jenkins, Neal G. Copeland, Yihan Yin C. Lin, Chang Ho Sohn, Chiraj K. Dalal, Long Cai, Michael B. Elowitz, Jin-Ming Jian-Xin Xin Lin, Peng Li, Delong Liu, Hyun Tak Hulin Jin, Jianping He, Mohammed Ata Ur Rasheed, Yrina Rochman, Li San Lu Wang, Kairong Cui, Chengyu Cailian Liu, Brian L. Kelsall, Rafi Ahmed, Warren J. Leonard, Wei Will Liao, Dustin E Schones, Jangsuk Oh, Yongzhi Cui, Kairong Cui, Tae-Young Roh, Keji Zhao, Warren J. Leonard, Jin-Ming Jian-Xin Xin Lin, Warren J. Leonard, Hong erh Liang, R Lee Reinhardt, Jennifer K Bando, Brandon M Sullivan, I-Cheng Ho, Richard M. Locksley, JiChu Li, Gail Huston, Susan L. Swain, Klaus Ley, Carlo Laudanna, Myron I. Cybulsky, Sussan Nourshargh, Sven Létourneau, Carsten Krieg, Giuseppe Pantaleo, Onur Boyman, Fabrice Lemaître, Hélène D Moreau, Laura Vedele, Philippe Bousso, Véronique Lecault, Michael Vaninsberghe, Sanja Sekulovic, David J H F Knapp, Stefan Wohrer, Francis Viel, Thomas Mclaughlin, Asefeh Jarandehei, Michelle Mark J Miller, Didier Falconnet, K Adam, David G Kent, Michael R Copley, Fariborz Taghipour, Connie J Eaves, R Keith Humphries, M James, Carl L Hansen, Antonio Lanzavecchia, Federica Sallusto, Bart N Lambrecht, Hamida Hammad, Galit Lahav, Nitzan

Rosenfeld, Alex Sigal, Naama Geva-Zatorsky, Andrew G. Arnold J Levine, Michael B. Elowitz, Uri Alon, Brian H. Ladle, Kelvin W. Kun-Po Li, Maggie J. Phillips, Alexandra B. Pucsek, Azeb Haile, Jonathan D. Powell, Elizabeth M. Jaffee, David a. Hildeman, Christopher J. Gamper, Bruno Kyewski, Ludger Klein, V a Kuznetsov, G D Knott, R F Bonner, Johana Kuncová-Kallio, Pasi J Kallio, Shantha Kumar, Marianne J. Skeen, Yaffa Adiri, Hyseuk Yoon, Vaiva D. Vezys, Aron E. Lukacher, Brian D. Evavold, H. Kirk Ziegler, Jeremy M. Boss, Brahma V. Kumar, Wenji Ma, Michelle Miron, Tomer Granot, Rebecca S. Guyer, Dustin J. Carpenter, Takashi Senda, Xiaoyun Sun, Siu Hong Ho, Harvey Lerner, Amy L. Friedman, Yufeng Shen, Donna L. Farber, C C Ku, M Murakami, Akemi Sakamoto, John W Kappler, Philippa Marrack, John H Koschwanez, Robert H Carlson, Deirdre R Meldrum, Mirjam Kool, Monique a M Willart, Menno van Nimwegen, Ingrid Bergen, Philippe Pouliot, J Christian Virchow, Neil Rogers, Fabiola Osorio, Caetano Reis E Sousa, Caetano Reis E Sousa, Hamida Hammad, Bart N Lambrecht, Robyn M. Kondrack, Judith Harbertson, Joyce T. Tan, Meghan E. McBreen, Charles D. Surh, Linda M. Bradley, Stefan Kobel, Ana Valero, Jonas Latt, Philippe Renaud, Matthias P Lutolf, Ichiko Kinjyo, Wolfgang Weninger, Philip D. Hodgkin, Jim Jianhua Qin, Sioh-Yang Tan, Cameron J. Wellard, Paulus Mrass, William Ritchie, Atsushi Doi, Lois L. Cavanagh, Michio Tomura, Asako Sakaue-Sawano, Osami Kanagawa, Atsushi Miyawaki, Philip D. Hodgkin, Wolfgang Weninger, Kevin R King, Sihong Wang, Daniel Irimia, Arul Jayaraman, Mehmet Toner, Martin L Yarmush, Stephanie H Seol-Hee Sangmin Kim, Choong-Eun Lee, W Kern, D a Puotinen, Kyeorda L Kemp, Steven D Levin, Paul J Bryce, Paul L Stein, A Kelso, P Groves, a B Troutt, K Francis, T a Kelly, D D Jeanfavre, D W McNeil, J R Woska, P L Reilly, E a Mainolfi, K. M Kishimoto, G H Nabozny, R Zinter, B J Bormann, Robert Rothlein, Katherine Kedzierska, Vanessa Venturi, Kenneth Field, Miles P. Davenport, Stephen J Turner, Peter C. Doherty, Sophie

a. Valkenburg, Peter C. Doherty, Miles P. Davenport, Vanessa Venturi, Ross M Kedl, John W Kappler, Philippa Marrack, Stefan H E Kaufmann, Suk jo Kang, Hong erh Liang, Boris Reizis, Richard M. Locksley, D Kamimura, K Ishihara, T Hirano, Gerard E Kaiko, Jay C Horvat, Kenneth W Beagley, Philip M Hansbro, Robin Kageyama, Jennifer L. Cannons, Fang Zhao, Isharat Yusuf, Christopher Lao, Michela Locci, Pamela L. Schwartzberg, Shane Crotty, Susan M. Kaech, Joyce T. Tan, E. John Wherry, Bogumila T. Konieczny, Charles D. Surh, Rafi Ahmed, Scott Hemby, Ellen Kersh, Rafi Ahmed, Weiguo Cui, S Z Josefowicz, L F Lu, Alexander Y. Rudensky, Margaret A Martha S. Jordan, Alan G Baxter, Simon a Jones, Jürgen Scheller, Stefan Rose-John, Lucy H L. Jones, R. Alli, B. Li, T. L. Geiger, Stephen J Jenkins, Dominik Ruckerl, Peter C Cook, Lucy H L. Jones, Fred D Finkelman, Nico van Rooijen, Andrew S Alice Macdonald, Judith E Allen, Edith M Janssen, Nathalie M Droin, Edward E Lemmens, Michael J Pinkoski, Steven J Bensinger, Benjamin D Ehst, Thomas S Griffith, Douglas R. Green, Stephen P Schoenberger, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadia Nadav Cohen, Steffen Jung, Amos Tanay, Ido Amit, Rob J. De Boer, Alan S. Perelson, Y Itoh, Ronald N. N. Germain, Irina Issaeva, Ariel Aharon Cohen, Eran Eden, Cellina Cohen-Saidon, Tamar Danon, Lydia Cohen, Uri Alon, Satoru Sayaka Ishihara, Akihiko Nishikimi, Eiji Umemoto, Masayuki Miyasaka, Makoto Saegusa, Koko Katagiri, Membrane Immunochemistry, Developmental Biology, Se Jin Im, Masao Hashimoto, Michael Y. Gerner, Judong Jeong Eon Junghwa Lee, Haydn T. Kissick, Matheus C. Burger, Qiang Shan, J. Scott Hale, Judong Jeong Eon Junghwa Lee, Tahseen H. Nasti, Arlene H. Sharpe, Gordon J. Freeman, Ronald N. N. Germain, Helder I. Nakaya, Hai-Hui Xue, Rafi Ahmed, Hirohito Ichii, Akemi Sakamoto, Masafumi Arima, Masahiko Hatano, Yoshikazu Kuroda, Takeshi Tokuhisa, Morgan Huse, Björn F Lillemeier, Michael S Kuhns, Daniel A N S. Chen, Mark M. Davis, Christopher

a Hunter, Simon a Jones, Jens Y Humrich, Henner Morbach, Reinmar Undeutsch, Philipp Enghard, Stefan Rosenberger, Olivia Weigert, Lutz Kloke, Juliane Heimann, Timo Gaber, Susan Brandenburg, Alexander Scheffold, Jochen Huehn, Andreas Radbruch, Gerd-Rüdiger Burmester, Gabriela Riemekasten, David A Hume, Washington Dc, Jane Hu-Li, C Pannetier, Liying Guo, Max Löhning, Hua Gu, C Watson, M Assenmacher, Andreas Radbruch, William E Paul, Tian Hong, Jianhua Xing, Liwu Li, John J Tyson, Soongweon Hong, Qiong Pan, Luke P. Lee, Brian D. Hondowicz, Dowon An, Jason M. Schenkel, Ki-Wook Karen S. Kyu Tae Kim, Holly R. Steach, Akshay T. Krishnamurty, Gladys J. Keitany, Esteban N. Garza, Kathryn A. Fraser, James J. Moon, William A. Altemeier, David Masopust, Marion Pepper, Mirja Hommel, Bruno Kyewski, Kristyna Holzerova, Jitka Zurmanova, Jan Neckar, Frantisek Kolar, Olga Novakova, G a Holländer, S Zuklys, C Morel, E Mizoguchi, K Mobisson, S Simpson, Cox Terhorst, W Wishart, D E Golan, a K Bhan, S J Burakoff, Thea Hogan, Andrey Shuvaev, Daniel Commenges, Andrew Yates, Robin E Callard, Rodolphe Thiebaut, Benedict Seddon, Robert M. Hoffman, Meng Yang, B Hoffman, D a Liebermann, Thomas Höfer, Oleg Krichevsky, Grégoire Gregoire Altan-Bonnet, Mirjam E Hoekstra, Feline E Dijkgraaf, Ton N M Schumacher, Jan C Rohr, Philip D. Hodgkin, William R Heath, Alan G Baxter, Mark R. Dowling, Ken R. Duffy, David R. Hodge, Elaine M. Hurt, William L. Farrar, Keiji Hirota, João H Duarte, Marc Veldhoen, Eve Hornsby, Ying Li, Daniel J Cua, Helena Ahlfors, Christoph Wilhelm, Mauro Tolaini, Ursula Menzel, Anna Garefalaki, Alexandre J Potocnik, Brigitta Stockinger, T Hirano, K Ishihara, M Hibi, Rodrigo Hess Michelini, Andrew L Doedens, Ananda W. Goldrath, Stephen M Hedrick, Tracy S P Heng, Michio W Painter, R C Henderson, Julie Helft, Alexandra Jacquet, Nathalie T Joncker, Isabelle Grandjean, Guillaume Dorothée, Adrien Kissenpfennig, Bernard Malissen, Polly Matzinger, Olivier Lantz, Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yi-

han Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, Christopher K. Glass, Sylvia Heink, Nir Yogev, Christoph Garbers, Marina Herwerth, Lilian Aly, Christiane Gasperi, Veronika Husterer, Andrew L. Croxford, Katja Möller-Hackbarth, Harald S Bartsch, Karl Sotlar, Stefan Krebs, Tommy Regen, Helmut Blum, Bernhard Hemmer, Thomas Misgeld, Thomas F Wunderlich, Juan Hidalgo, Mohamed Oukka, Stefan Rose-John, Marc Schmidt-Supprian, Ari Waisman, Thomas Korn, Ahmed N Hegazy, Michael Peine, Caroline Helmstetter, Isabel Panse, Anja Fröhlich, Andreas Bergthaler, Lukas Flatz, Daniel D Pinschewer, Andreas Radbruch, Max Löhning, Daniel Hebenstreit, Andrew Deonarine, M Madan Babu, Sarah a Teichmann, E. D. Hawkins, Marian L Turner, Mark R. Dowling, C van Gend, Philip D. Hodgkin, J. F. Markham, L P McGuinness, Philip D. Hodgkin, Yuval Hart, Shlomit Reich-Zeliger, Yaron E Antebi, Irina Zaretsky, Avraham E Avi Mayo, Uri Alon, Nir Friedman, Robert Haralick, K. Shanmugan, I. Dinstein, Qing Han, Elizabeth M. Bradshaw, Björn Nilsson, David a. Hafler, J. Christopher Love, N. Bagheri, Elizabeth M. Bradshaw, David a. Hafler, D. a. Lauffenburger, J. Christopher Love, Arnold Han, Evan W Newell, Jacob Glanville, Nielsen Fernandez-Becker, Chaitan Khosla, Yueh-Hsiu Chien, Mark M. Davis, Anna-Katerina Hadjantonakis, Virginia E Papaioannou, Liying Guo, Jane Hu-Li, William E Paul, Karolin Guldevall, Bruno Vanherberghen, Thomas Frisk, Johan Hurtig, Athanasia E Christakou, Otto Manneberg, Sara Lindström, Helene Andersson-Svahn, Martin Wiklund, Björn Önfelt, Hua Gu, Yong rui Zou, Klaus Rajewsky, Jane L. Grogan, Richard M. Locksley, Markus Mohrs, Brian Harmon, Dee A. Lacy, John W. Sedat, Richard M. Locksley, J P Griffin, I M Orme, Yuan Gong, Adebola O Ogunniyi, J. Christopher Love, B Y D W Goldman, F Chang, L A Gifford, E J Goetzl, H R Bourne, Samy Gobaa, Sylke Hoehnel, Marta Roccio, Andrea Negro, Stefan Kobel, Matthias P Lutolf, Hila Gingold, Disa Tehler, Nanna R Christoffersen, Morten M Nielsen, Fazila Asmar, Susanne M Kooi-

tra, Nicolaj S Christophersen, Lise Lotte Christensen, Michael Borre, Karina D Sørensen, Lars D Andersen, Claus L Andersen, Esther Hulleman, Tom Wurdinger, Elisabeth Ralfkiær, Kristian Helin, Kirsten Grønbæk, Torben Orntoft, Sebastian M Waszak, Orna Dahan, Jakob Skou Pedersen, Anders H Lund, Yitzhak Pilpel, Jane Gilmour, Paul Lavender, K Gijbels, S Brocke, J S Abrams, L Steinman, Amanda V. Gett, Federica Sallusto, Antonio Lanzavecchia, Jens Geginat, Philip D. Hodgkin, Bram Gerritsen, Aridaman Pandit, Ronald N. N. Germain, Martin Meier-Schellersheim, Aleksandra Nita-Lazar, Iain D C Fraser, M Gerloni, S Xiong, S Mukerjee, Stephen P Schoenberger, M Croft, M Zanetti, Carmen Gerlach, Jan C Rohr, Leïla Perié, Nienke van Rooij, Jeroen W J van Heijst, Arno Velds, Jos Urbanus, Shalin H Naik, Heinz Jacobs, Joost B Beltman, Rob J de Boer, Ton N M Schumacher, Audrey Gérard, Omar Khan, Peter Beemiller, Erin Oswald, Joyce Hu, Mehrdad Matloubian, Matthew F. Krummel, Rachel S. Friedman, Jordan Jacobelli, Matthew F. Krummel, Rebekka Geiger, Thomas Duhen, Antonio Lanzavecchia, Federica Sallusto, Marc a Gavin, Jeffrey P Rasmussen, Jason D Fontenot, Valeria Vasta, Vincent C Manganiello, Joseph a Beavo, Alexander Y. Rudensky, Luca Gattinoni, Daniel E Speiser, Mathias Lichterfeld, Chiara Bonini, Enrico Lugli, Yun Ji, Zoltan Pos, Chrystal M. Paulos, Máire F. Quigley, Jorge R. Almeida, Emma Gostick, Zhiya Yu, Carmine Carpenito, Ena Wang, Daniel C. Douek, David a. Price, Carl H. June, Francesco M. Marincola, Mario Roederer, Nicholas P. Restifo, Paul A Garrity, Daniel A N S. Chen, Ellen V Rothenberg, Barbara J Wold, Karina García-Martínez, Kalet León, Vitaly V Ganusov, Dejan Milutinovic, Rob J De Boer, S Gallucci, M Lolkema, Polly Matzinger, Thomas F Gajewski, Hans Schreiber, Yang-Xin Fu, Julien Gagnon, Sheela Ramanathan, Chantal Leblanc, Alexandre Cloutier, Patrick P McDonald, Subburaj Ilangumaran, Kirsten L. Frieda, James M. A Y Linton, Sahand Hormoz, Jong-Hoon Joonhyuk Choi, Ke Huan K. Chow, Zakary S. Singer, Mark W. Budde, Michael B.

Elowitz, Long Cai, Thais a Fornari, Paula B Donate, Claudia Macedo, Elza T Sakamoto-Hojo, Eduardo a Donadi, Geraldo a Passos, Michael Flossdorf, Jens Rössler, Veit R. Buchholz, Dirk H. Busch, Thomas Höfer, Gary S Firestein, William D Roeder, Julie A Laxer, K A Y S Townsend, Casey T Weaver, Joanne T Hom, James M. A Y Linton, Bruce E Torbett, Andrew L Glasebrook, Pamela J Fink, Deborah W Hendricks, Ofer Feinerman, Garrit Jentsch, Karen E Tkach, Jesse W Coward, Matthew M Hathorn, Michael W Sneddon, Thierry Emonet, Kathleen Karen M Kendall A Smith, Grégoire Gregoire Altan-Bonnet, Donna L. Farber, Naomi a. Yudanin, Nicholas P. Restifo, Shannon Faley, Kevin Seale, Jacob Hughey, David K Schaffer, Scott VanCompernolle, Brett McKinney, Franz Baudenbacher, Derya Unutmaz, John P Wikswo, Michael B. Elowitz, Andrew G. Arnold J Levine, Eric D Siggia, Peter S Swain, Jamil El-Ali, Peter K Sorger, Klavs F Jensen, Jackson G Egen, Antonio Gigliotti Rothfuchs, Carl G Feng, Marcus a Horwitz, Alan Sher, Ronald N. N. Germain, Michael L. Dustin, Trever G Bivona, Mark R Philips, Omer Dushek, Milos Aleksic, Richard J Wheeler, Hao Zhang, Shaun-Paul Cordoba, Yan-Chun Peng, Ji-Li Chen, Vincenzo Cerundolo, Tao Dong, Daniel Coombs, Philip Anton van der Merwe, Thomas Duhen, Rebekka Duhen, Antonio Lanzavecchia, Federica Sallusto, Daniel J. Campbell, Ken R. Duffy, Philip D. Hodgkin, Cameron J. Wellard, J. F. Markham, J. H. S. Zhou, R. Holmberg, E. D. Hawkins, J. Hasbold, Mark R. Dowling, Philip D. Hodgkin, A. Kan, S. Heinzel, J. H. S. Zhou, J. M. Marchingo, Cameron J. Wellard, J. F. Markham, Philip D. Hodgkin, Hans Dooms, Kristen Wolslegel, Patricia Lin, Abul K. Abbas, Junsang Doh, Matthew F. Krummel, Miju Kim, Matthew F. Krummel, Dienz, Rincon, Gianfranco Di Genova, Natalia Savelyeva, Amy Suchacki, Stephen M Thirdborough, Freda K Stevenson, Dino Di Carlo, Liz Y Wu, Luke P. Lee, Assaf Deutsch, Naomi Zurgil, Ihar Hurevich, Yana Shafran, Elena Afrimzon, Pnina Lebovich, Mordechai Deutsch, Elissa K Deenick, Amanda V. Gett, Philip D. Hodgkin,

Rita De Gasperi, Anne B Rocher, Miguel a Gama Sosa, Susan L Wearne, Gissel M Perez, Victor L Friedrich, Patrick R Hof, Gregory a Elder, Jaime De Calisto, Ninghai Wang, Guoxing Wang, Burcu Yigit, Pablo Engel, Cox Terhorst, Daniel Day, Kim Pham, Mandy J Ludford-Menting, Jane Oliaro, David Izon, Sarah M. Russell, Min Gu, Mark M. Davis, John D Altman, Evan W Newell, Dang, Dang, Xiaojiong Lu, Yuefen Lou, Eric V. Dang, Joseph Barbi, Huang-Yu Yang, Dilini Jinasena, Hong Yu, Ying Zheng, Zachary Bordman, Juan Fu, Young Kim, Hung-Rong Yen, Weibo Luo, Karen Zeller, Larissa Shimoda, Suzanne L. Topalian, Gregg L. Semenza, Chi V. Dang, Drew M. M. Pardoll, Fan Pan, Andrew L. Croxford, Thorsten Buch, Rémi J Creusot, Lindy L Thomsen, John P Tite, Benjamin M Chain, M. Cote, C. Fos, a. J. Canonigo-Balancio, Klaus Ley, S. Becart, A. Altman, Ahmet F. Coskun, Long Cai, Neal G. Copeland, Nancy a. Jenkins, D L Court, Denis Comte, Maria P. Karampetsou, Katalin Kis-Toth, Nobuya Yoshida, Sean J. Bradley, Masayuki Mizui, Michihito Kono, Julie R. Solomon, Vasileios C. Kyttaris, George C. Tsokos, Ariel Aharon Cohen, Tomer Kalisky, Avraham E Avi Mayo, Naama Geva-Zatorsky, Tamar Danon, Irina Issaeva, Ronen Benjamine Kopito, Natalie Perzov, Ron Milo, Alex Sigal, Uri Alon, R L Coffman, Steven L. Reiner, Hubert Cochet, Michel Montaudon, Guillaume Calmettes, Jean michel Franconi, Sylvain Miraux, Elodie Parzy, Maria L. Ciocca, Burton E Barnett, Janis K Burkhardt, John T. Chang, Steven L. Reiner, Woosung Chung, Hye Hyeon Eum, Han Byoel Hae Ock Heung Kyu Lee, Kyung Min Lee, Han Byoel Hae Ock Heung Kyu Lee, Ki-Wook Karen S. Kyu Tae Kim, Han Suk Ryu, Stephanie H Seol-Hee Sangmin Kim, Judong Jeong Eon Junghwa Lee, Yeon Hee Park, Zhengyan Kan, Wonshik Han, Woong Yang Park, Seeyoung Choi, Ronald H Schwartz, Paul J Choi, Long Cai, Kirsten L. Frieda, X Sunney Xie, Jong-Hoon Joonhyuk Choi, B K Cho, V P Rao, Q Ge, H N Eisen, Ji-Li Chen, Pei Yu Chiou, Aaron T. Ohta, Ming C. Wu, A. Chess, I Simon, H Cedar, R Axel, Ye Chen, Elizabeth Adams, Fred-

erico S Regateiro, David J Vaux, Alexander G Betz, Kristian G Andersen, Herman Waldmann, Duncan Howie, Daniel A N S. Chen, Ira Mellman, Pratip K Chattopadhyay, Todd M Gierahn, Mario Roederer, J. Christopher Love, Madhumouli Chatterjee, Thomas Rauen, Katalin Kis-Toth, Vasileios C. Kyttaris, Christian M. Hedrich, Cox Terhorst, George C. Tsokos, To-Ha Thai, Cox Terhorst, George C. Tsokos, Christian M. Hedrich, Thomas Rauen, Christina Ioannidis, Cox Terhorst, George C. Tsokos, C C Chao, R Jensen, M O Dailey, Qing Chang, Eirini Bournazou, Pasquale Sansone, Marjan Berishaj, Sizhi Paul Gao, Laura Daly, Jared Wels, Till Theilen, Selena Granitto, Xinmin Zhang, Jesse W Cotari, Mary L Alpaugh, Elisa de Stanchina, Katia Manova, Ming O Li, Massimiliano Bonafe, Claudio Ceccarelli, Mario Taffurelli, Donatella Santini, Grégoire Gregoire Altan-Bonnet, Rosandra Kaplan, Larry Norton, Norihiro Nishimoto, Dennis Huszar, David Lyden, Jacqueline Bromberg, John T. Chang, Vikram R. Palanivel, Ichiko Kinjyo, Felix Schambach, Andrew M Intlekofer, Arnob Banerjee, Sarah a Longworth, Kristine E Vinup, Paulus Mrass, Jane Oliaro, Nigel Killeen, Jordan S Orange, Sarah M. Russell, Wolfgang Weninger, Steven L. Reiner, Maria L. Ciocca, Ichiko Kinjyo, Vikram R. Palanivel, Courtney E. McClurkin, Caitlin S. DeJong, Erin C. Mooney, Jiyeon S. Kim, Natalie C. Steinel, Jane Oliaro, Catherine C. Yin, Bogdan I. Florea, Herman S. Overkleeft, Leslie J. Berg, Sarah M. Russell, Gary a. Koretzky, Margaret A Martha S. Jordan, Steven L. Reiner, Chih Hao Chang, Jonathan D. Curtis, Leonard B. Maggi, Brandon Faubert, Alejandro V. Villarino, David O'Sullivan, Stanley Ching Cheng Huang, Gerritje J.W. Van Der Windt, Julianna Blagih, Jing Qiu, Jason D. Weber, Erika L. Edward J. Pearce, Russell G. Jones, Erika L. Edward J. Pearce, Ornella Cazzalini, a Ivana Scovassi, Monica Savio, Lucia a Stivala, Ennio Prosperi, R B Cattell, Julienne L. Carstens, Pedro Correa de Sampaio, Dalu Yang, Souptik Barua, Huamin Wang, Arvind Anjana Rao, James P. Allison, Valerie S. LeBleu, Raghu Kalluri, Silvia Carroll, Mohamed

Al-Rubeai, Corey M Carlson, Bart T Endrizzi, Jinghai Wu, Xiaojie Ding, Michael a Weinreich, Elizabeth R Walsh, Maqsood a Wani, Jerry B Lingrel, Kristin a Hogquist, Stephen C Jameson, Leo M Carlin, Efstathios G Stamatiades, Cedric Auffray, Richard N Hanna, Leanne Glover, Gema Vizcay-Barrena, Catherine C Hedrick, H Terence Cook, Sandra Diebold, Frederic Geissmann, Doreen a. Cantrell, Kathleen Karen M Kendall A Smith, Michael a Cannarile, Nicholas a Lind, Richard Rivera, Alison D Sheridan, Kristin a Camfield, Bei Bei Wu, Kitty P Cheung, Zhaoqing Ding, Ananda W. Goldrath, Robin E Callard, Dinis Pedro Calado, Tiago Paixão, Dan Holmberg, Matthias Haury, Long Cai, Nir Friedman, X Sunney Xie, Nigel J Burroughs, Philip Anton van der Merwe, F M Burnet, R P Bucy, A Panoskaltsis-Mortari, G Q Huang, JiChu Li, L Karr, M Ross, J H Russell, Kenneth M Murphy, Casey T Weaver, Veit R. Buchholz, Michael Flossdorf, Inge Hensel, Lorenz Kretschmer, Bianca Weissbrich, Patricia Gräf, Admar Verschoor, Matthias Schiemann, Thomas Höfer, Dirk H. Busch, J Brockdorff, Morten M Nielsen, A Svejgaard, P Dobson, C Röpke, C Geisler, N Odum, S B Kanner, Morten M Nielsen, N Borregaard, C Geisler, A Svejgaard, N Odum, M F Brizzi, P Defilippi, A Rosso, M Venturino, G Garbarino, A Miyajima, L Silengo, G Tarone, L Pegoraro, Onur Boyman, Jonathan Sprent, Philippe Bousso, Rémy T. Boscacci, Friederike Pfeiffer, Kathrin Gollmer, Ana Isabel Checa Sevilla, Ana Maria Martin, Silvia Fernandez Soriano, Daniela Natale, Sarah Henrickson, Ulrich H. Von Andrian, Yoshinori Fukui, Mario Mellado, Urban Deutsch, Britta Engelhardt, Jens V. Stein, Jeffrey a Bluestone, Charles R Mackay, John J O'Shea, Brigitta Stockinger, Mark Bix, Hélène Beuneu, Fabrice Lemaître, Jacques Deguine, Hélène D Moreau, Isabelle Bouvier, Zacarias Garcia, Matthew L Albert, Philippe Bousso, J. Adam Best, David a. Blair, Jamie Knell, Edward Yang, Viveka Mayya, Andrew L Doedens, Michael L. Dustin, Ananda W. Goldrath, Paul Monach, Susan a. Shinton, Richard R. Hardy, Radu Jianu, Daphne David Koller, Jim Collins, Roi Gazit, Brian S. Garrison, Der-

rick J. Rossi, Kavitha Narayan, Katelyn Sylvia, Joonsoo Kang, Anne Fletcher, Kutlu Elpek, Angelique Bellemare-Pelletier, Deepali Malhotra, Shannon Turley, J. Adam Best, Vladimir Jojic, Daphne David Koller, Tal Shay, Aviv Regev, Nadia Nadav Cohen, Patrick Brennan, Michael Brenner, Taras Kreslavsky, Natalie a. Bezman, Joseph C. Sun, Charlie C. Kim, Lewis L. Lanier, Jim John H Jennifer Miller, Brian Brown, Miriam Merad, Emmanuel L. Gautier, Claudia Jakubzick, Gwendalyn J. Randolph, Francis Kim, Tata Nageswara Rao, Amy Wagers, Tracy S P Heng, Michio W Painter, Jeffrey Ericson, Scott Davis, Ayla Ergun, Michael Mingueneau, Diane Mathis, Christophe Benoist, Sten-Erik Bergström, Eva Bergdahl, Karl-Gösta Sundqvist, D C Bennett, Sean C. Bendall, Erin F. Simonds, Peng Qiu, El ad Ad David Amir, Peter O Krutzik, Rachel Finck, Robert V Bruggner, Rachel Melamed, Angelica Trejo, Olga I Ornatsky, Robert S Balderas, Sylvia K Plevritis, Karen Sachs, Dana Pe'er, Scott D Tanner, Garry P. Nolan, Dana Pe, Scott D Tanner, Garry P. Nolan, Simone Becattini, Daniela Latorre, Federico Mele, Mathilde Foglierini, Corinne De Gregorio, Antonino Cassotta, Blanca Fernandez, Sander Kelderman, Ton N M Schumacher, Davide Corti, Antonio Lanzavecchia, Federica Sallusto, Susan Beatty, Charles Brooks, Shawna Dean, Mark Hanna, Dan Janiak, Chen Kung, Jia Ni, Anne Samuel, Kunal Thaker, Thomas Barthlott, Halima Moncrieffe, Marc Veldhoen, Christopher J Atkins, Jillian Christensen, Anne O'Garra, Brigitta Stockinger, Arnob Banerjee, Scott M Siamon Gordon, Andrew M Intlekofer, Michael a Paley, Erin C. Mooney, Tulia Lindsten, E. John Wherry, Steven L. Reiner, Vladimir P. Badovinac, Jodie S. Haring, John T. Harty, H S Azzam, A Grinberg, K Lui, Hao Shen, E W Shores, P E Love, Janilyn Arsenio, Patrick J Metz, John T. Chang, Boyko Kakaradov, Patrick J Metz, Gene W Yeo, John T. Chang, Stephanie H Seol-Hee Sangmin Kim, Gene W Yeo, John T. Chang, Silvia Ariotti, Joost B Beltman, Grzegorz Chodaczek, Mirjam E Hoekstra, Anna E van Beek, Raquel Gomez-Eerland, Laila Ritsma, Jacco van Rheenen, Athana-

sius F M Marée, Tomasz Zal, Rob J de Boer, John B a G Haanen, Ton N M Schumacher, Tamar Huberman Arieli, Clemente F Arias, Miguel a Herrero, Jose a Cuesta, Francisco J Acosta, Cristina Fernandez-Arias, Victor Appay, Rene a W van Lier, Federica Sallusto, Mario Roederer, Rustom Antia, Vitaly V Ganusov, Rafi Ahmed, K Mark Ansel, Ivana Djuretic, Bogdan Tanasa, Arvind Anjana Rao, El ad Ad David Amir, Kara L. Davis, Michelle D. Tadmor, Erin F. Simonds, Jacob H. Levine, Sean C. Bendall, Daniel K. Shenfeld, Smita Krishnaswamy, Garry P. Nolan, Dana Pe'Er, Oral Alpan, Eric Bachelder, Eda Isil, Heinz Arnheiter, Polly Matzinger, Uri Alon, Afonso R M Almeida, Inês F Amado, Joseph Reynolds, Julien Berges, Grant Lythe, Carmen Molina-París, Antonio a Freitas, Juan C.L. Carlos López Alfonso, Nadine S. Schaadt, Ralf Schönmeyer, N. Brieu, Germain Forestier, Cédric Wemmert, Friedrich Feuerhake, Haralampos Hatzikirou, Rama S. Akondy, Mark Fitch, Srilatha Edupuganti, Shu Yang, Haydn T. Kissick, Kelvin W. Kun-Po Li, Ben a. Youngblood, Hossam a. Abdelsamed, Donald J. McGuire, Kristen W. Cohen, Gabriela Alexe, Shashi Nagar, Megan M. McCausland, Satish Gupta, Pramila Tata, W. Nicholas Haining, M. Juliana McElrath, David Zhang, Bin Hu, William J. Greenleaf, Jorg J. Goronzy, Mark J. Mulligan, Marc Hellerstein, Rafi Ahmed, David Gray, Michael J. Bevan, Steven L. Reiner, Douglas T. Fearon, Shimrit Adutler-Lieber, Irina Zaretsky, Helena Sabany, Elena Kartvelishvily, Ofra Golani, Benjamin Geiger, Nir Friedman, Ilia Platzman, Janosch Deeg, Nir Friedman, Joachim P. Spatz, and Benjamin Geiger. Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science*, 8, 2014.

[98] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23, 1995.

[99] Oliviero Carugo and Kristina Djinovic Carugo. Half a century of ramachandran plots. *Acta Crystallographica Section D: Biological Crystallography*, 69, 2013.

[100] Scott A. Hollingsworth and P. Andrew Karplus. A fresh look at the ramachan-

dran plot and the occurrence of standard structures in proteins. *Biomolecular Concepts*, 1, 2010.

[101] Alice Qinhua Zhou, Corey S. O'Hern, and Lynne Regan. Revisiting the ramachandran plot from a new angle. *Protein Science*, 20, 2011.

[102] Shuguang Yuan, HC Stephen Chan, and Zhenquan Hu. Using pymol as a platform for computational drug design. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(2):e1298, 2017.

[103] Zheng Yang, Keren Lasker, Dina Schneidman-Duhovny, Ben Webb, Conrad C Huang, Eric F Pettersen, Thomas D Goddard, Elaine C Meng, Andrej Sali, and Thomas E Ferrin. Ucsf chimera, modeller, and imp: an integrated modeling system. *Journal of structural biology*, 179(3):269–278, 2012.

[104] Jane S. Richardson. The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, 34, 1981.

[105] Jack D. Dunitz. Pauling's left-handed -helix. *Angewandte Chemie - International Edition*, 40, 2001.

[106] H. A. Scheraga. Effect of side chain-backbone electrostatic interactions on the stability of alpha-helices. *Proceedings of the National Academy of Sciences of the United States of America*, 82, 1985.

[107] Thomas E. Creighton. Stability of alpha-helices. *Nature*, 326, 1987.

[108] XY Zhang, Jian Shao, SX Jiang, Biao Wang, and Yue Zheng. Structure-dependent electrical conductivity of protein: Its differences between alpha-domain and beta-domain structures. *Nanotechnology*, 26(12):125702, 2015.

[109] Jinlong He, Lin Zhang, and Ling Liu. The hydrogen-bond configuration modulates the energy transfer efficiency in helical protein nanotubes. *Nanoscale*, 13, 2021.

[110] P. R. Gascoyne, R. Pethig, and A. Szent-Györgyi. Water structure-dependent charge transport in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 78, 1981.

[111] R. P. Sheridan, R. M. Levy, and F. R. Salemme. alpha-helix dipole model and electrostatic stabilization of 4-alpha-helical proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 79, 1982.

[112] A Wada. The -helix as an electric macro-dipole. *Advances in Biophysics*, 9, 1976.

[113] Daniel Van Belle, Ignace Couplet, Martine Prevost, and Shoshana J. Wodak. Calculations of electrostatic properties in proteins. *Journal of Molecular Biology*, 198, 1987.

[114] T Herz, P Otto, and T Clark. On the band gap in peptide alpha-helices. *International Journal of Quantum Chemistry*, 79, 2000.

[115] S. Marqusee, V. H. Robbins, and R. L. Baldwin. Unusually stable helix formation in short alanine-based peptides. *Proceedings of the National Academy of Sciences of the United States of America*, 86, 1989.

[116] B A French, K L Chow, E N Olson, and R J Schwartz. Heterodimers of myogenic helix-loop-helix regulatory factors and e12 bind a complex element governing myogenic induction of the avian cardiac alpha-actin promoter. *Molecular and Cellular Biology*, 11, 1991.

[117] W. Kabsch, H. G. Mannherz, and D. Suck. Three-dimensional structure of the complex of actin and dnase i at 4.5 a resolution. *The EMBO journal*, 4, 1985.

[118] R Li, N Mitra, H Gratkowski, G Vilaire, R Litvinov, C Nagasami, J W Weisel, J D Lear, W F DeGrado, and J S Bennett. Activation of integrin alphaiibbeta3 by modulation of transmembrane helix associations. *Science*, 300, 2003.

[119] Wei Yang, Motomu Shimaoka, Jian Feng Chen, and Timothy A. Springer. Activation of integrin -subunit i-like domains by one-turn c-terminal -helix deletions. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 2004.

[120] Johannes Eble. Collagen-binding integrins as pharmaceutical targets. *Current Pharmaceutical Design*, 11, 2005.

[121] Paula Tapial Martínez, Pilar López Navajas, and Daniel Lietha. Fak structure and regulation by membrane interactions and force in focal adhesions. *Biomolecules*, 10, 2020.

[122] J. T. Parsons, K. H. Martin, J. K. Slack, J. M. Taylor, and S. A. Weed. Focal adhesion kinase: A regulator of focal adhesion dynamics and cell movement. *Oncogene*, 19, 2000.

[123] Saumitra Singh. Transmembrane alpha helix prediction using neural networks. *International Journal of Scientific Research in Science, Engineering and Technology*, 2020.

[124] Alexander Bevacqua, Sachit Bakshi, and Yu Xia. Principal component analysis of alpha-helix deformations in transmembrane proteins. *PLoS ONE*, 16, 2021.

[125] Zhao Qin, Laurent Kreplak, and Markus J. Buehler. Nanomechanical properties of vimentin intermediate filament dimers. *Nanotechnology*, 20, 2009.

[126] Jérémie Bertaud, Joshua Hester, Daniel D. Jimenez, and Markus J. Buehler. Energy landscape, structure and rate effects on strength properties of alpha-helical proteins. *Journal of Physics Condensed Matter*, 22, 2010.

[127] Zhao Qin, Laurent Kreplak, and Markus J. Buehler. Hierarchical structure controls nanomechanical properties of vimentin intermediate filaments. *PLoS ONE*, 4, 2009.

[128] Jamie J. Kwan, Neil Warner, Joban Maini, Kelvin W. Chan Tung, Hoshang Zakaria, Tony Pawson, and Logan W. Donaldson. Saccharomyces cerevisiae ste50 binds the mapkkk ste11 through a head-to-tail sam domain interaction. *Journal of Molecular Biology*, 356, 2006.

[129] Roger Armen, Darwin O V Alonso, and Valerie Daggett. The role of {alpha}-, 310-, and {pi}-helix in helix-¿coil transitions. *Protein Science*, 12, 2003.

[130] Sanguk Kim and T A Cross. 2d solid state nmr spectral simulation of 310, [alpha], and [pi]-helices. *Journal of Magnetic Resonance*, 168, 2004.

[131] Mary E. Karpen, Pieter L. De Haseth, and Kenneth E. Neet. Differences in the amino acid distributions of 310-helices and -helices. *Protein Science*, 1, 1992.

[132] Jia K.E. Sun and Andrew J. Doig. Addition of side-chain interactions to 310-helix/coil and - helix/310-helix/coil theory. *Protein Science*, 7, 1998.

[133] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J.C. Berendsen. Gromacs: Fast, flexible, and free. *Journal of Computational Chemistry*, 26, 2005.

[134] Christine S Schwaiger, Pär Bjelkmar, Berk Hess, and Erik Lindahl. 310-helix conformation facilitates the transition of a voltage sensor s4 segment toward the down state. *Biophysical journal*, 100(6):1446–1454, 2011.

[135] Todd M. Weaver. The -helix translates structure into function. *Protein Science*, 9, 2008.

[136] Yiteng Zhang, Gennady P. Berman, and Sabre Kais. The radical pair mechanism and the avian chemical compass: Quantum coherence and entanglement. *International Journal of Quantum Chemistry*, 115, 2015.

[137] Daniela Passarella and Michel Goedert. Beta-sheet assembly of tau and neurodegeneration in drosophila melanogaster. *Neurobiology of Aging*, 72, 2018.

[138] Giorgia Zandomeneghi, Mark R.H. Krebs, Margaret G. McCammon, and Marcus Fändrich. Ftir reveals structural differences between native -sheet proteins and amyloid fibrils. *Protein Science*, 13, 2009.

[139] Francesca Lugli, Francesca Toschi, Fabio Biscarini, and Francesco Zerbetto. Electric field effects on short fibrils of a amyloid peptides. *Journal of Chemical Theory and Computation*, 6, 2010.

[140] Wendy A. Loughlin, Joel D. A. Tyndall, Matthew P. Glenn, Timothy A. Hill, and David P. Fairlie. Update 1 of: Beta-strand mimetics. *Chemical Reviews*, 110, 2010.

[141] D. A. Kirschner, H. Inouye, L. K. Duffy, A. Sinclair, M. Lind, and D. J. Selkoe. Synthetic peptide homologous to beta protein from alzheimer disease forms

amyloid-like fibrils in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 84, 1987.

[142] Ralph Michael, Cees Otto, Aufried Lenferink, Ellen Gelpi, Gustavo A. Montenegro, Jurja Rosandić, Francisco Tresserra, Rafael I. Barraquer, and Gijs F.J.M. Vrensen. Absence of amyloid-beta in lenses of alzheimer patients: A confocal raman microspectroscopic study. *Experimental Eye Research*, 119, 2014.

[143] P. E. Fraser, J. T. Nguyen, W. K. Surewicz, and D. A. Kirschner. ph-dependent structural transitions of alzheimer amyloid peptides. *Biophysical Journal*, 60, 1991.

[144] Peggy Cebe, Xiao Hu, David L. Kaplan, Evgeny Zhuravlev, Andreas Wurm, Daniela Arbeiter, and Christoph Schick. Beating the heat-fast scanning melts silk beta sheet crystals. *Scientific Reports*, 3, 2013.

[145] Keiji Numata, Peggy Cebe, and David L. Kaplan. Mechanism of enzymatic degradation of beta-sheet crystals. *Biomaterials*, 31, 2010.

[146] U Carlsson and B H Jonsson. Folding of beta-sheet proteins. *Curr Opin Struct Biol*, 5, 1995.

[147] Sinan Keten and Markus J. Buehler. Strength limit of entropic elasticity in beta-sheet protein domains. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 78, 2008.

[148] Xiao Hu, David Kaplan, and Peggy Cebe. Determining beta-sheet crystallinity in fibrous proteins by thermal analysis and infrared spectroscopy. *Macromolecules*, 39, 2006.

[149] P. C. Painter, L. E. Mosher, and C. Rhoads. Low-frequency modes in the raman spectra of proteins. *Biopolymers*, 21, 1982.

[150] I. A. Balakhnina, N. N. Brandt, A. A. Mankova, and A. Yu Chikishev. The problem of manifestation of tertiary structure in the vibrational spectra of proteins. *Vibrational Spectroscopy*, 114, 2021.

[151] David A. Turton, Hans Martin Senn, Thomas Harwood, Adrian J. Lapthorn,

Elizabeth M. Ellis, and Klaas Wynne. Terahertz underdamped vibrational motion governs protein-ligand binding in solution. *Nature Communications*, 5, 2014.

[152] Toru Fujimori, Takafumi Yamashino, Takahiko Kato, and Takeshi Mizuno. Circadian-controlled basic/helix-loop-helix factor, pil6, implicated in light-signal transduction in arabidopsis thaliana. *Plant and Cell Physiology*, 45, 2004.

[153] Min Ni, James M. Tepperman, and Peter H. Quail. Pif3, a phytochrome-interacting factor necessary for normal photoinduced signal transduction, is a novel basic helix-loop-helix protein. *Cell*, 95, 1998.

[154] José A. Caro, Kyle W. Harpole, Vignesh Kasinath, Jackwee Lim, Jeffrey Granja, Kathleen G. Valentine, Kim A. Sharp, and A. Joshua Wand. Entropy in molecular recognition by proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 114, 2017.

[155] Kristin R. Gleitsman, Raghuvir N. Sengupta, and Daniel Herschlag. Slow molecular recognition by rna. *RNA*, 23, 2017.

[156] Chung Kim Dong and Joon Kang Dae. Molecular recognition and specific interactions for biosensing applications. *Sensors*, 8, 2008.

[157] Jinqiao Dong and Anthony P. Davis. Molecular recognition mediated by hydrogen bonding in aqueous media. *Angewandte Chemie - International Edition*, 60, 2021.

[158] Denise M. Ferkey, Piali Sengupta, and Noelle D. L'Etoile. Chemosensory signal transduction in caenorhabditis elegans. *Genetics*, 217, 2021.

[159] Robert Bekus and Thomas Schrader. Artificial signal transduction. *ChemistryOpen*, 9, 2020.

[160] Davide Provasi, Andrea Bortolato, and Marta Filizola. Exploring molecular mechanisms of ligand recognition by opioid receptors with metadynamics. *Biochemistry*, 48, 2009.

[161] Eric J. Sundberg and Roy A. Mariuzza. Molecular recognition in antibody-antigen complexes. *Advances in Protein Chemistry*, 61, 2002.

[162] Löic Etheve, Juliette Martin, and Richard Lavery. Dynamics and recognition within a protein-dna complex: A molecular dynamics study of the skn-1/dna interaction. *Nucleic Acids Research*, 44, 2016.

[163] Xiu Luo, Xinlu Wang, Yina Gao, Jingpeng Zhu, Songqing Liu, Guangxia Gao, and Pu Gao. Molecular mechanism of rna recognition by zinc-finger antiviral protein. *Cell Reports*, 30, 2020.

[164] Yu Hsien Hwang Fu, Sowmya Chandrasekar, Jae Ho Lee, and Shu Ou Shan. A molecular recognition feature mediates ribosome-induced srp-receptor assembly during protein targeting. *Journal of Cell Biology*, 218, 2019.

[165] Hans Joachim Gabius, Maré Cudic, Tammo Diercks, Herbert Kaltner, Jürgen Kopitz, Kevin H. Mayo, Paul V. Murphy, Stefan Oscarson, René Roy, Andreas Schedlbauer, Stefan Toegel, and Antonio Romero. What is the sugar code? *ChemBioChem*, 2021.

[166] Yukiko Kamiya, Daiki Kamiya, Kazuo Yamamoto, Beat Nyfeler, Hans Peter Hauri, and Koichi Kato. Molecular basis of sugar recognition by the human l-type lectins ergic-53, vipl, and vip36. *Journal of Biological Chemistry*, 283, 2008.

[167] Seiji Shinkai. "dynamic" molecular recognition and chirality segregation utilizing concepts of molecular machines and molecular assemblies. *Proceedings of the Japan Academy Series B: Physical and Biological Sciences*, 95, 2019.

[168] Hua Jiang and Bradley D. Smith. Dynamic molecular recognition on the surface of vesicle membranes. *Chemical Communications*, 2006.

[169] Joshua M. Brockman and Khalid Salaita. Mechanical proofreading: A general mechanism to enhance the fidelity of information transfer between cells. *Frontiers in Physics*, 7, 2019.

[170] Yonatan Savir and Tsvi Tiusty. Conformational proofreading: The impact of

conformational changes on the specificity of molecular recognition. *PLoS ONE*, 2, 2007.

[171] Yonatan Savir and Tsvi Tlusty. Molecular recognition as an information channel: The role of conformational changes. 2009.

[172] Samuel H. Sternberg, Benjamin Lafrance, Matias Kaplan, and Jennifer A. Doudna. Conformational control of dna target cleavage by crispr-cas9. *Nature*, 527, 2015.

[173] Harshad Ghodke, Hong Wang, Ching L. Hsieh, Selamawit Woldemeskel, Simon C. Watkins, Vesna Rapić-Otrin, and Bennett Van Houten. Single-molecule analysis reveals human uv-damaged dna-binding protein (uv-ddb) dimerizes on dna via multiple kinetic intermediates. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 2014.

[174] Aiman Alam-Nazki and J. Krishnan. Spatial control of biochemical modification cascades and pathways. *Biophysical Journal*, 108, 2015.

[175] Joseph W Kim and John Z Luo. The biochemical cascades of the human pancreatic ÃÂ²-cells: The role of micrornas. *Journal of Bioanalysis Biomedicine*, 07, 2015.

[176] Mengmeng Zhang and Shuqun Zhang. Mitogen-activated protein kinase cascades in plant signaling. *Journal of Integrative Plant Biology*, 64, 2022.

[177] Boris N. Kholodenko. Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. *European Journal of Biochemistry*, 267, 2000.

[178] Karanvir Saini and Dennis E. Discher. Forced unfolding of proteins directs biochemical cascades. *Biochemistry*, 58, 2019.

[179] Hiroshi Yasui, Teru Hideshima, Paul G. Richardson, and Kenneth C. Anderson. Novel therapeutic strategies targeting growth factor signalling cascades in multiple myeloma. *British Journal of Haematology*, 132, 2006.

[180] Yangyang Guo, Yanyi Xiao, Hengyue Zhu, Hangcheng Guo, Ying Zhou, Yang-

ping Shentu, Chenfei Zheng, Chaosheng Chen, and Yongheng Bai. Inhibition of proliferation-linked signaling cascades with atractylenolide i reduces myofibroblastic phenotype and renal fibrosis. *Biochemical Pharmacology*, 183, 2021.

[181] Erica Mari Nell, Lize Van Der Merwe, Jill Cook, Christopher J. Handley, Malcolm Collins, and Alison V. September. The apoptosis pathway and the genetic predisposition to achilles tendinopathy. *Journal of Orthopaedic Research*, 30, 2012.

[182] Robin Schwarzer, Lucie Laurien, and Manolis Pasparakis. New insights into the regulation of apoptosis, necroptosis, and pyroptosis by receptor interacting protein kinase 1 and caspase-8. *Current Opinion in Cell Biology*, 63, 2020.

[183] Yunli Zhao, Qiuli Wu, and Dayong Wang. An epigenetic signal encoded protection mechanism is activated by graphene oxide to inhibit its induced reproductive toxicity in caenorhabditis elegans. *Biomaterials*, 79, 2016.

[184] Helen M. Beere. 'the stress of dying': The role of heat shock proteins in the regulation of apoptosis. *Journal of Cell Science*, 117, 2004.

[185] Aviad Ben-Shmuel, Noah Joseph, Batel Sabag, and Mira Barda-Saad. Lymphocyte mechanotransduction: The regulatory role of cytoskeletal dynamics in signaling cascades and effector functions. *Journal of Leukocyte Biology*, 105, 2019.

[186] L. Ramage, G. Nuki, and D. M. Salter. Signalling cascades in mechanotransduction: Cell-matrix interactions and mechanical loading. *Scandinavian Journal of Medicine and Science in Sports*, 19, 2009.

[187] Jiacheng Huang, Lele Zhang, Dalong Wan, Lin Zhou, Shusen Zheng, Shengzhang Lin, and Yiting Qiao. Extracellular matrix and its therapeutic potential for cancer treatment. *Signal Transduction and Targeted Therapy*, 6, 2021.

[188] Stephen J.P. Pratt, Rachel M. Lee, and Stuart S. Martin. The mechanical microenvironment in breast cancer. *Cancers*, 12, 2020.

[189] Song Li, Ngan F. Huang, and Steven Hsu. Mechanotransduction in endothelial cell migration. *Journal of Cellular Biochemistry*, 96, 2005.

[190] Jimena Canales, Diego Morales, Constanza Blanco, Jose Rivas, Nicolas Diaz, Ioannis Angelopoulos, and Oscar Cerda. A tr(i)p to cell migration: New roles of trp channels in mechanotransduction and cancer. *Frontiers in Physiology*, 10, 2019.

[191] Zuping Wu, Chenchen Zhou, Quan Yuan, Demao Zhang, Jing Xie, and Shujuan Zou. Ctgf facilitates cell-cell communication in chondrocytes via pi3k/akt signalling pathway. *Cell Proliferation*, 54, 2021.

[192] Éric Chevalier, Audrey Loubert-Hudon, Erin L. Zimmerman, and Daniel P. Matton. Cell-cell communication and signalling pathways within the ovule: From its inception to fertilization. *New Phytologist*, 192, 2011.

[193] Anna Bigas and Lluis Espinosa. Notch signaling in cell–cell communication pathways. *Current Stem Cell Reports*, 2, 2016.

[194] Kalpana Manthiram, Qing Zhou, Ivona Aksentijevich, and Daniel L. Kastner. The monogenic autoinflammatory diseases define new pathways in human innate immunity and inflammation. *Nature Immunology*, 18, 2017.

[195] Abu Sadat Md Sayem, Aditya Arya, Hamed Karimian, Narendiran Krishnasamy, Ameya Ashok Hasamnis, and Chowdhury Faiz Hossain. Action of phytochemicals on insulin signaling pathways accelerating glucose transporter (glut4) protein translocation. *Molecules*, 23, 2018.

[196] Yi Zeng, Le Zhang, and Zhiping Hu. Cerebral insulin, insulin signaling pathway, and brain angiogenesis. *Neurological Sciences*, 37, 2016.

[197] Huimin Shao, Zhongyu Han, Natalia Krasteva, and Dayong Wang. Identification of signaling cascade in the insulin signaling pathway in response to nanopolystyrene particles. *Nanotoxicology*, 13, 2019.

[198] Boris N. Kholodenko, Nora Rauch, Walter Kolch, and Oleksii S. Rukhlenko. A

systematic analysis of signaling reactivation and drug resistance. *Cell Reports*, 35, 2021.

[199] Maryam Rahmati, Eduardo A. Silva, Janne E. Reseland, Catherine A. Heyward, and Håvard J. Haugen. Biological responses to physicochemical properties of biomaterial surface. *Chemical Society Reviews*, 49, 2020.

[200] Shensi Shen, Guillermo Rodrigo, Satya Prakash, Eszter Majer, Thomas E. Landrain, Boris Kirov, José Antonio Daròs, and Alfonso Jaramillo. Dynamic signal processing by ribozyme-mediated rna circuits to control gene expression. *Nucleic Acids Research*, 43, 2015.

[201] Maximilian Hörner and Wilfried Weber. Molecular switches in animal cells. *FEBS Letters*, 586, 2012.

[202] Silvia Turroni, Manlio Tolomeo, Gianfranco Mamone, Gianluca Picariello, Elisa Giacomini, Patrizia Brigidi, Marinella Roberti, Stefania Grimaudo, Rosaria Maria Pipitone, Antonietta Di Cristina, and Maurizio Recanatini. A natural-like synthetic small molecule impairs bcr-abl signaling cascades and induces megakaryocyte differentiation in erythroleukemia cells. *PLoS ONE*, 8, 2013.

[203] Charles W.E. Tomlinson, Katy A.S. Cornish, Andrew Whiting, and Ehmke Pohl. Structure-functional relationship of cellular retinoic acid-binding proteins i and ii interacting with natural and synthetic ligands. *Acta Crystallographica Section D: Structural Biology*, 77, 2021.

[204] Jacob D. Durrant and J. Andrew McCammon. Hbonanza: A computer algorithm for molecular-dynamics-trajectory hydrogen-bond analysis. *Journal of Molecular Graphics and Modelling*, 31, 2011.

[205] Adam Hospital, Josep Ramon Goñi, Modesto Orozco, and Josep L. Gelpí. Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, 8, 2015.

[206] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and

David A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25, 2004.

[207] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of Computational Chemistry*, 31, 2010.

[208] Song Hi Lee and Jahun Kim. Transport properties of bulk water at 243–550 k: a comparative molecular dynamics simulation study using spc/e, tip4p, and tip4p/2005 water models. *Molecular Physics*, 117, 2019.

[209] Maria Tsanai, Pim W J.M. Frederix, Carsten F.E. Schroer, Paulo C.T. Souza, and Siewert J. Marrink. Coacervate formation studied by explicit solvent coarse-grain molecular dynamics with the martini model. *Chemical Science*, 12, 2021.

[210] David B. Kony, Wolfgang Damm, Serge Stoll, Wilfred F. Van Gunsteren, and Philippe H. Hünenberger. Explicit-solvent molecular dynamics simulations of the polysaccharide schizophyllan in water. *Biophysical Journal*, 93, 2007.

[211] Liu Ming Yan, Chao Sun, and Hui Ting Liu. Opposite phenomenon to the flying ice cube in molecular dynamics simulations of flexible tip3p water. *Advances in Manufacturing*, 1, 2013.

[212] Song Hi Lee. Temperature dependence on structure and self-diffusion of water: A molecular dynamics simulation study using spc/e model. *Bulletin of the Korean Chemical Society*, 34, 2013.

[213] Song Hi Lee. Temperature dependence of the thermal conductivity of water: A molecular dynamics simulation study using the spc/e model. *Molecular Physics*, 112, 2014.

[214] Matthew Carter Childers and Valerie Daggett. Validating molecular dynamics simulations against experimental observables in light of underlying conformational ensembles. *Journal of Physical Chemistry B*, 122, 2018.

[215] J. Lobaugh and Gregory A. Voth. A quantum model for water: Equilibrium and dynamical properties. *Journal of Chemical Physics*, 106, 1997.

[216] Ramu Anandakrishnan, Aleksander Drozdetski, Ross C. Walker, and Alexey V. Onufriev. Speed of conformational change: Comparing explicit and implicit solvent molecular dynamics simulations. *Biophysical Journal*, 108, 2015.

[217] Nazish Hoda and Ronald G. Larson. Explicit- and implicit-solvent molecular dynamics simulations of complex formation between polycations and polyanions. *Macromolecules*, 42, 2009.

[218] Paul L. Barclay and Duan Z. Zhang. Periodic boundary conditions for arbitrary deformations in molecular dynamics simulations. *Journal of Computational Physics*, 435, 2021.

[219] Carsten Kutzner, Szilárd Páll, Martin Fechner, Ansgar Esztermann, Bert L. de Groot, and Helmut Grubmüller. More bang for your buck: Improved use of gpu nodes for gromacs 2018. *Journal of Computational Chemistry*, 40, 2019.

[220] Gonzalo Nicolas Barreales, Marcos Novalbos, Miguel A. Otaduy, and Alberto Sanchez. Mdscale: Scalable multi-gpu bonded and short-range molecular dynamics. *Journal of Parallel and Distributed Computing*, 157, 2021.

[221] Szilárd Páll, Artem Zhmurov, Paul Bauer, Mark Abraham, Magnus Lundborg, Alan Gray, Berk Hess, and Erik Lindahl. Heterogeneous parallelization and acceleration of molecular dynamics simulations in gromacs. *Journal of Chemical Physics*, 153, 2020.

[222] Hun Joo Myung, Ryuji Sakamaki, Kwang Jin Oh, Tetsu Narumi, Kenji Yasuoka, and Sik Lee. Accelerating molecular dynamics simulation using graphics processing unit. *Bulletin of the Korean Chemical Society*, 31, 2010.

[223] Khanittha Kerdpol, Jintawee Kicuntod, Peter Wolschann, Seiji Mori, Chompoonut Rungnim, Manaschai Kunaseth, Hisashi Okumura, Nawee Kungwan, and Thanyada Rungrotmongkol. Cavity closure of 2-hydroxypropyl--

cyclodextrin: Replica exchange molecular dynamics simulations. *Polymers*, 11, 2019.

[224] Jim Pfaendtner and Massimiliano Bonomi. Efficient sampling of high-dimensional free-energy landscapes with parallel bias metadynamics. *Journal of Chemical Theory and Computation*, 11, 2015.

[225] Christopher D. Fu and Jim Pfaendtner. Lifting the curse of dimensionality on enhanced sampling of reaction networks with parallel bias metadynamics. *Journal of Chemical Theory and Computation*, 14, 2018.

[226] Federico Fogolari, Alessandra Corazza, Sara Fortuna, Miguel Angel Soler, Bryan Van Schouwen, Giorgia Brancolini, Stefano Corni, Giuseppe Melacini, and Gennaro Esposito. Distance-based configurational entropy of proteins from molecular dynamics simulations. *PLoS ONE*, 10, 2015.

[227] Alexander Kantardjiev and Petko M. Ivanov. Entropy rules: Molecular dynamics simulations of model oligomers for thermoresponsive polymers. *Entropy*, 22, 2020.

[228] Heiko Schäfer, Alan E. Mark, and Wilfred F. Van Gunsteren. Absolute entropies from molecular dynamics simulation trajectories. *Journal of Chemical Physics*, 113, 2000.

[229] D. B. Amirkulova and A. D. White. Recent advances in maximum entropy biasing techniques for molecular dynamics. *Molecular Simulation*, 45, 2019.

[230] Jia Bo Le and Jun Cheng. Modeling electrochemical interfaces from ab initio molecular dynamics: water adsorption on metal surfaces at potential of zero charge. *Current Opinion in Electrochemistry*, 19, 2020.

[231] Usman L. Abbas, Qi Qiao, Manh Tien Nguyen, Jian Shi, and Qing Shao. Molecular dynamics simulations of heterogeneous hydrogen bond environment in hydrophobic deep eutectic solvents. *AIChE Journal*, 68, 2022.

[232] Federico Coppola, Fulvio Perrella, Alessio Petrone, Greta Donati, and Nadia Rega. A not obvious correlation between the structure of green fluorescent

protein chromophore pocket and hydrogen bond dynamics: A choreography from ab initio molecular dynamics. *Frontiers in Molecular Biosciences*, 7, 2020.

[233] Eyber Domingos Alves, Douglas X. de Andrade, Agnaldo R. de Almeida, and Guilherme Colherinhas. Molecular dynamics study of hydrogen bond in peptide membrane at 150–300 k. *Journal of Molecular Liquids*, 349, 2022.

[234] Ce Zhou, Xingxing Li, Zhongliang Gong, Chuancheng Jia, Yuanwei Lin, Chunhui Gu, Gen He, Yuwu Zhong, Jinlong Yang, and Xuefeng Guo. Direct observation of single-molecule hydrogen-bond dynamics with single-bond resolution. *Nature Communications*, 9, 2018.

[235] Peter Lykos. Modeling the hydrogen bond within molecular dynamics. *Journal of Chemical Education*, 81, 2004.

[236] Xin Zheng Li, Brent Walker, and Angelos Michaelides. Quantum nature of the hydrogen bond. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 2011.

[237] A. Michaelides. Quantum nature of the hydrogen bond. *Acta Crystallographica Section A Foundations of Crystallography*, 67, 2011.

[238] Kuan Pern Tan, Khushboo Singh, Anirban Hazra, and M. S. Madhusudhan. Peptide bond planarity constrains hydrogen bond geometry and influences secondary structure conformations. *Current Research in Structural Biology*, 3, 2021.

[239] P. Siani, H. Khandelia, M. Orsi, and L. G. Dias. Parameterization of a coarse-grained model of cholesterol with point-dipole electrostatics. *Journal of Computer-Aided Molecular Design*, 32, 2018.

[240] Jean François Olivieri, James T. Hynes, and Damien Laage. Confined water's dielectric constant reduction is due to the surrounding low dielectric media and not to interfacial molecular ordering. *Journal of Physical Chemistry Letters*, 12, 2021.

[241] M. A. Belyanchikov, M. Savinov, Z. V. Bedran, P. Bednyakov, P. Proschek,

J. Prokleska, V. A. Abalmasov, J. Petzelt, E. S. Zhukova, V. G. Thomas, A. Dudka, A. Zhugayevych, A. S. Prokhorov, V. B. Anzin, R. K. Kremer, J. K.H. Fischer, P. Lunkenheimer, A. Loidl, E. Uykur, M. Dressel, and B. Gorshunov. Dielectric ordering of water molecules arranged in a dipolar lattice. *Nature Communications*, 11, 2020.

[242] Yosuke KATAOKA and Yuri YAMADA. Phase diagram of a lennard-jones system by molecular dynamics simulations. *Journal of Computer Chemistry, Japan*, 13, 2014.

[243] Joseph E. Davis, Obaidur Rahaman, and Sandeep Patel. Molecular dynamics simulations of a dmpc bilayer using nonadditive interaction models. *Biophysical Journal*, 96, 2009.

[244] Ratan K. Mishra, Krishan Kanhaiya, Jordan J. Winetrout, Robert J. Flatt, and Hendrik Heinz. Force field for calcium sulfate minerals to predict structural, hydration, and interfacial properties. *Cement and Concrete Research*, 139, 2021.

[245] Chamila C. Dharmawardhana, Krishan Kanhaiya, Tzu Jen Lin, Amanda Garley, Marc R. Knecht, Jihan Zhou, Jianwei Miao, and Hendrik Heinz. Reliable computational design of biological-inorganic materials to the large nanometer scale using interface-ff. *Molecular Simulation*, 43, 2017.

[246] Wei Liu, Xiangpeng Xie, Wei Qian, Xiaozhuo Xu, and Yan Shi. Optimal linear filtering for networked control systems with random matrices, correlated noises, and packet dropouts. *IEEE Access*, 8, 2020.

[247] James A. Snyder, Tigran Abramyan, Jeremy A. Yancey, Aby A. Thyparambil, Yang Wei, Steven J. Stuart, and Robert A. Latour. Development of a tuned interfacial force field parameter set for the simulation of protein adsorption to silica glass. *Biointerphases*, 7, 2012.

[248] Hanne S. Antila and Emppu Salonen. Polarizable force fields. *Methods in Molecular Biology*, 924, 2013.

[249] Zhifeng Jing, Chengwen Liu, Sara Y. Cheng, Rui Qi, Brandon D. Walker,

Jean Philip Piquemal, and Pengyu Ren. Polarizable force fields for biomolecular simulations: Recent advances and applications. *Annual Review of Biophysics*, 48, 2019.

[250] Nicholas A. Besley. Modeling of the spectroscopy of core electrons with density functional theory. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 11, 2021.

[251] Eric A. Mills and Steven S. Plotkin. Density functional theory for protein transfer free energy. *Journal of Physical Chemistry B*, 117, 2013.

[252] Stefano Angioletti-Uberti, Matthias Ballauff, and Joachim Dzubiella. Dynamic density functional theory of protein adsorption on polymer-coated nanoparticles. *Soft Matter*, 10, 2014.

[253] Vyshnavi Vennelakanti, Azadeh Nazemi, Rimsha Mehmood, Adam H. Steeves, and Heather J. Kulik. Harder, better, faster, stronger: Large-scale qm and qm/mm for predictive modeling in enzymes and proteins. *Current Opinion in Structural Biology*, 72, 2022.

[254] Hong Jiang. Band gaps from the tran-blaha modified becke-johnson approach: A systematic investigation. *Journal of Chemical Physics*, 138, 2013.

[255] Michael G. Medvedev, Ivan S. Bushmarinov, Jianwei Sun, John P. Perdew, and Konstantin A. Lyssenko. Density functional theory is straying from the path toward the exact functional. *Science*, 355, 2017.

[256] Annemie Bogaerts, Narjes Khosravian, Jonas Van Der Paal, Christof C.W. Verlackt, Maksudbek Yusupov, Balu Kamaraj, and Erik C. Neyts. Multilevel molecular modelling for plasma medicine. *Journal of Physics D: Applied Physics*, 49, 2015.

[257] Maria Andrea Mroginski, Franz Mark, Walter Thiel, and Peter Hildebrandt. Quantum mechanics/molecular mechanics calculation of the raman spectra of the phycocyanobilin chromophore in -c-phycocyanin. *Biophysical Journal*, 93, 2007.

[258] Anna Helena Mazurek, Łukasz Szeleszczuk, and Dariusz Maciej Pisklak. A review on combination of ab initio molecular dynamics and nmr parameters calculations. *International Journal of Molecular Sciences*, 22, 2021.

[259] Grigory Kolesov, Efthimios Kaxiras, and Efstratios Manousakis. Density functional theory beyond the born-oppenheimer approximation: Accurate treatment of the ionic zero-point motion. *Physical Review B*, 98(19), Nov 2018.

[260] Claudia Filippi, Francesco Buda, Leonardo Guidoni, and Adalgisa Sinicropi. Bathochromic shift in green fluorescent protein: A puzzle for qm/mm approaches. *Journal of Chemical Theory and Computation*, 8, 2012.

[261] Marc Etienne Moret, Enrico Tapavicza, Leonardo Guidoni, Ute F. Röhrig, Marialore Sulpizi, Ivano Tavernelli, and Ursula Rothlisberger. Quantum mechanical/molecular mechanical (qm/mm) car-parrinello simulations in excited states. *Chimia*, 59, 2005.

[262] Shideh Ahmadi, Lizandra Barrios Herrera, Morteza Chehelamirani, Jiří Hostaš, Said Jalife, and Dennis R. Salahub. Multiscale modeling of enzymes: Qm-cluster, qm/mm, and qm/mm/md: A tutorial review. *International Journal of Quantum Chemistry*, 118, 2018.

[263] Manuel Guidon, Florian Schiffmann, Jürg Hutter, and Joost Vandevondele. Ab initio molecular dynamics using hybrid density functionals. *Journal of Chemical Physics*, 128, 2008.

[264] Lili Cao and Ulf Ryde. On the difference between additive and subtractive qm/mm calculations. *Frontiers in Chemistry*, 6, 2018.

[265] Hans Martin Senn and Walter Thiel. Qm/mm methods for biomolecular systems. *Angewandte Chemie - International Edition*, 48, 2009.

[266] Brigitta Elsässer and Peter Goettig. Mechanisms of proteolytic enzymes and their inhibition in qm/mm studies. *International Journal of Molecular Sciences*, 22, 2021.

[267] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee,

and Lee G. Pedersen. A smooth particle mesh ewald method. *The Journal of Chemical Physics*, 103, 1995.

[268] Peng Wang, Yecheng Shao, Hongtao Wang, and Wei Yang. Accurate interatomic force field for molecular dynamics simulation by hybridizing classical and machine learning potentials. *Extreme Mechanics Letters*, 24, 2018.

[269] S. Bougueroua, R. Spezia, S. Pezzotti, S. Vial, F. Quessette, D. Barth, and M. P. Gaigeot. Graph theory for automatic structural recognition in molecular dynamics simulations. *Journal of Chemical Physics*, 149, 2018.

[270] Stefano Caprasecca, Lorenzo Cupellini, Sandro Jurinovich, Daniele Loco, Filippo Lipparini, and Benedetta Mennucci. A polarizable qm/mm description of environment effects on nmr shieldings: from solvated molecules to pigment–protein complexes. *Theoretical Chemistry Accounts*, 137, 2018.

[271] Sharon Hammes-Schiffer. Hydrogen tunneling and protein motion in enzyme reactions. *Accounts of Chemical Research*, 39, 2006.

[272] Judith P. Klinman and Amnon Kohen. Hydrogen tunneling links protein dynamics to enzyme catalysis. *Annual Review of Biochemistry*, 82, 2013.

[273] Brian J. Bahnson, Thomas D. Colby, Jodie K. Chin, Barry M. Goldstein, and Judith P. Klinman. A link between protein structure and enzyme catalyzed hydrogen tunneling. *Proceedings of the National Academy of Sciences of the United States of America*, 94, 1997.

[274] S. C. Tsai and J. P. Klinman. Probes of hydrogen tunneling with horse liver alcohol dehydrogenase at subzero temperatures. *Biochemistry*, 40, 2001.

[275] Tareq Hameduh, Yazan Haddad, Vojtech Adam, and Zbynek Heger. Homology modeling in the time of collective and artificial intelligence. *Computational and Structural Biotechnology Journal*, 18, 2020.

[276] Muhammed Tilahun Muhammed and Esin Aki-Yalcin. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chemical Biology and Drug Design*, 93, 2019.

[277] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T. Heer, Tjaart A.P. De Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede. Swiss-model: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46, 2018.

[278] Hanka Venselaar, Robbie P. Joosten, Bas Vroling, Coos A.B. Baakman, Maarten L. Hekkelman, Elmar Krieger, and Gert Vriend. Homology modelling and spectroscopy, a never-ending love story. *European Biophysics Journal*, 39, 2010.

[279] Yazan Haddad, Vojtech Adam, and Zbynek Heger. Ten quick tips for homology modeling of high-resolution protein 3d structures. *PLoS Computational Biology*, 16, 2020.

[280] Jeff Gauthier, Antony T. Vincent, Steve J. Charette, and Nicolas Derome. A brief history of bioinformatics. *Briefings in Bioinformatics*, 20, 2019.

[281] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18, 2017.

[282] Dilara Ayyildiz and Silvano Piazza. Introduction to bioinformatics, 2019.

[283] Maciej Pawel Ciemny, Aleksandra Elzbieta Badaczewska-Dawid, Monika Pikuzinska, Andrzej Kolinski, and Sebastian Kmiecik. Modeling of disordered protein structures using monte carlo simulations and knowledge-based statistical force fields. *International Journal of Molecular Sciences*, 20, 2019.

[284] Wouter Boomsma, Jes Frellsen, Tim Harder, Sandro Bottaro, Kristoffer E. Johansson, Pengfei Tian, Kasper Stovgaard, Christian Andreetta, Simon Olsson, Jan B. Valentin, Lubomir D. Antonov, Anders S. Christensen, Mikael Borg, Jan H. Jensen, Kresten Lindorff-Larsen, Jesper Ferkinghoff-Borg, and Thomas Hamelryck. Phaistos: A framework for markov chain monte carlo simulation and inference of protein structure. *Journal of Computational Chemistry*, 34, 2013.

[285] K. Jayanthi and C. Mahesh. Need of machine learning in bioinformatics. *International Journal of Innovative Technology and Exploring Engineering*, 8, 2019.

[286] David Baker. Protein folding, structure prediction and design. *Biochemical Society Transactions*, 42, 2014.

[287] Derek N. Woolfson. A brief history of de novo protein design: Minimal, rational, and computational: De novo protein design. *Journal of Molecular Biology*, 433, 2021.

[288] David Baker. What has de novo protein design taught us about protein folding and biophysics? *Protein Science*, 28, 2019.

[289] Po Ssu Huang, Scott E. Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537, 2016.

[290] Sergey Ovchinnikov, Hahnbeom Park, David E. Kim, Frank DiMaio, and David Baker. Protein structure prediction using rosetta in casp12. *Proteins: Structure, Function and Bioinformatics*, 86, 2018.

[291] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28, 2000.

[292] Robert A. Langan, Scott E. Boyken, Andrew H. Ng, Jennifer A. Samson, Galen Dods, Alexandra M. Westbrook, Taylor H. Nguyen, Marc J. Lajoie, Zibo Chen, Stephanie Berger, Vikram Khipple Mulligan, John E. Dueber, Walter R.P. Novak, Hana El-Samad, and David Baker. De novo design of bioactive protein switches. *Nature*, 572, 2019.

[293] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A.A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Pe-

tersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, 596, 2021.

[294] Mathias Uhlén, Linn Fagerberg, Bjö M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, Ing Marie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle Von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar Von Heijne, Jens Nielsen, and Fredrik Pontén. Tissue-based map of the human proteome. *Science*, 347, 2015.

[295] Bo Wen, Wen Feng Zeng, Yuxing Liao, Zhiao Shi, Sara R. Savage, Wen Jiang, and Bing Zhang. Deep learning in proteomics. *Proteomics*, 20, 2020.

[296] Bilal Aslam, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool. Proteomics: Technologies and their applications. *Journal of Chromatographic Science*, 55, 2017.

[297] Yang Woo Kwon, Han Seul Jo, Sungwon Bae, Youngsuk Seo, Parkyong Song, Minseok Song, and Jong Hyuk Yoon. Application of proteomics in cancer: Recent trends and approaches for biomarkers discovery. *Frontiers in Medicine*, 8, 2021.

[298] Thomas D. Burton and Nicholas S. Eyre. Applications of deep mutational scanning in virology. *Viruses*, 13, 2021.

[299] Vanessa E. Gray, Katherine Sitko, Floriane Z. Ngako Kameni, Miriam Williamson, Jason J. Stephany, Nicholas Hasle, and Douglas M. Fowler. Elucidating the molecular determinants of ab aggregation with deep mutational scanning. *G3: Genes, Genomes, Genetics*, 9, 2019.

[300] Sarah K. Hilton, Michael B. Doud, and Jesse D. Bloom. Phydms: Software

for phylogenetic analyses informed by deep mutational scanning. *PeerJ*, 2017, 2017.

[301] Sakineh Abbasi and Saeedeh Masoumi. Next-generation sequencing (ngs). *International Journal of Advanced Science and Technology*, 29, 2020.

[302] Jason R. Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95, 2010.

[303] Shadi Shokralla, Jennifer L. Spall, Joel F. Gibson, and Mehrdad Hajibabaei. Next-generation sequencing technologies for environmental dna research. *Molecular Ecology*, 21, 2012.

[304] Deeptak Verma, Gevorg Grigoryan, and Chris Bailey-Kellogg. Pareto optimization of combinatorial mutagenesis libraries. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16, 2019.

[305] E. Merino, J. Osuna, F. Bolivar, and X. Soberon. A general, pcr-based method for single or combinatorial oligonucleotide- directed mutagenesis on puc/m13 vectors. *BioTechniques*, 12, 1992.

[306] Tomer A. Roth, Gregory A. Weiss, Charles Eigenbrot, and Sachdev S. Sidhu. A minimized m13 coat protein defines the requirements for assembly into the bacteriophage particle. *Journal of Molecular Biology*, 322, 2002.

[307] Gregory A. Weiss, Tomer A. Roth, Pierre F. Baldi, and Sachdev S. Sidhu. Comprehensive mutagenesis of the c-terminal domain of the m13 gene-3 minor coat protein: The requirements for assembly into the bacteriophage particle. *Journal of Molecular Biology*, 332, 2003.

[308] David Gonzalez-Perez, James Ratcliffe, Shu Khan Tan, Mary Chen May Wong, Yi Pei Yee, Natsai Nyabadza, Jian He Xu, Tuck Seng Wong, and Kang Lan Tee. Random and combinatorial mutagenesis for improved total production of secretory target protein in escherichia coli. *Scientific Reports*, 11, 2021.

[309] Laila Berg, Trine Aakvik Strand, Svein Valla, and Trygve Brautaset. Combi-

natorial mutagenesis and selection to understand and improve yeast promoters. *BioMed Research International*, 2013, 2013.

[310] Michele C. Kieke, Eric V. Shusta, Eric T. Boder, Luc Teyton, K. Dane Wittrup, and David M. Kranz. Selection of functional t cell receptor mutants from a yeast surface-display library. *Proceedings of the National Academy of Sciences of the United States of America*, 96, 1999.

[311] Nimish Gera, Mahmud Hussain, and Balaji M. Rao. Protein selection using yeast surface display. *Methods*, 60, 2013.

[312] Deniz T. Yucesoy, Siddharth S Rath, Jacob L Rodriguez, Jonathan Francis-Landau, Oliver Nakano-Baker, and Mehmet Sarikaya. Deep directed evolution of solid binding peptides for quantitative big-data generation. *bioRxiv*, 2021.

[313] Frances H. Arnold. Directed evolution of enzymes and binding proteins; scientific background on the nobel prize for chemistry 2018. *The Royal Swedish Academy of Sciences*, 50005, 2018.

[314] Álvaro Sánchez, Jean C.C. Vila, Chang Yu Chang, Juan Diaz-Colunga, Sylvie Estrela, and María Rebolleda-Gomez. Directed evolution of microbial communities. *Annual Review of Biophysics*, 50, 2021.

[315] Oliver F. Brandenberg, Kai Chen, and Frances H. Arnold. Directed evolution of a cytochrome p450 carbene transferase for selective functionalization of cyclic compounds. *Journal of the American Chemical Society*, 141, 2019.

[316] Satoshi Fujii, Tomoaki Matsuura, and Tetsuya Yomo. In vitro directed evolution of alpha-hemolysin by liposome display. *Biophysics (Japan)*, 11, 2015.

[317] Cory D. Sago, Melissa P. Lokugamage, Fatima Z. Islam, Brandon R. Krupczak, Manaka Sato, and James E. Dahlman. Nanoparticles that deliver rna to bone marrow identified by in vivo directed evolution. *Journal of the American Chemical Society*, 140, 2018.

[318] Bioinformatics Published and John Wiley. Proteins : Structure , function , and bioinformatics. *Production*, 19, 2006.

[319] Anna Tramontano and Veronica Morea. Exploiting evolutionary relationships for predicting protein structures. *Biotechnology and Bioengineering*, 84, 2003.

[320] Jhih Siang Lai, Burkhard Rost, Bostjan Kobe, and Mikael Bodén. Evolutionary model of protein secondary structure capable of revealing new biological relationships. *Proteins: Structure, Function and Bioinformatics*, 88, 2020.

[321] Kulkarni Keya and Sundarrajan Priya. A study of phylogenetic relationships and homology of cytochrome c using bioinformatics. *Int. Res. J. of Science Engineering*, 4, 2016.

[322] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215, 1990.

[323] Weicai Ye, Ying Chen, Yongdong Zhang, and Yuesheng Xu. H-blast: A fast protein sequence alignment toolkit on heterogeneous computers with gpus. *Bioinformatics*, 33, 2017.

[324] Seung Yon Rhee. Bioinformatics. current limitations and insights for the future. *Plant Physiology*, 138, 2005.

[325] Michael Hsing and Artem Cherkasov. Indel pdb: A database of structural insertions and deletions derived from sequence alignments of closely related proteins. *BMC Bioinformatics*, 9, 2008.

[326] Eric Paul Bennett, Bent Larsen Petersen, Ida Elisabeth Johansen, Yiyuan Niu, Zhang Yang, Christopher Aled Chamberlain, Özcan Met, Hans H. Wandall, and Morten Frödin. Indel detection, the 'achilles heel' of precise genome editing: A survey of methods for accurate profiling of gene editing induced indels. *Nucleic Acids Research*, 48, 2021.

[327] Ryang Guk Kim and Jun Tao Guo. Systematic analysis of short internal indels and their impact on protein folding. *BMC Structural Biology*, 10, 2010.

[328] I. Miklós, G. A. Lunter, and I. Holmes. A "long indel" model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21, 2004.

[329] Benjamin I. Baarda, Fabian G. Martinez, and Aleksandra E. Sikora. Proteomics, bioinformatics and structure-function antigen mining for gonorrhea vaccines. *Frontiers in Immunology*, 9, 2018.

[330] Daniel J. Rigden. *From protein structure to function with bioinformatics: Second Edition*. 2017.

[331] Bernd Reif, Sharon E. Ashbrook, Lyndon Emsley, and Mei Hong. Solid-state nmr spectroscopy. *Nature Reviews Methods Primers*, 1, 2021.

[332] Abdul Hamid Emwas, Raja Roy, Ryan T. McKay, Leonardo Tenori, Edoardo Saccenti, G. A. Nagana Gowda, Daniel Raftery, Fatimah Alahmari, Lukasz Jaremko, Mariusz Jaremko, and David S. Wishart. Nmr spectroscopy for metabolomics research. *Metabolites*, 9, 2019.

[333] Stanislav Mazurenko, Zbynek Prokop, and Jiri Damborsky. Machine learning in enzyme engineering. *ACS Catalysis*, 10, 2020.

[334] Ehsaneddin Asgari and Mohammad R.K. Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE*, 10, 2015.

[335] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 2021.

[336] Tomer Tsaban, Julia K. Varga, Orly Avraham, Ziv Ben-Aharon, Alisa Khramushin, and Ora Schueler-Furman. Harnessing protein folding neural networks for peptide–protein docking. *Nature Communications*, 13, 2022.

[337] Sergio Gomes Ramalli, Andrew John Miles, Robert W. Janes, and B. A. Wallace. The pcddb (protein circular dichroism data bank): A bioinformatics resource for protein characterisations and methods development. *Journal of Molecular Biology*, 434, 2022.

[338] Alex Bateman. Uniprot: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47, 2019.

[339] Alex Bateman, Maria Jesus Martin, Sandra Orchard, Michele Magrane, Rahat Agivetova, Shadab Ahmad, Emanuele Alpi, Emily H. Bowler-Barnett, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Ray Coetzee, Austra Cukura, Alan da Silva, Paul Denny, Tunca Dogan, Thank God Ebenezer, Jun Fan, Leyla Garcia Castro, Penelope Garmiri, George Georghiou, Leonardo Gonzales, Emma Hatton-Ellis, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Petteri Jokinen, Vishal Joshi, Dushyanth Jyothi, Antonia Lock, Rodrigo Lopez, Aurelien Luciani, Jie Luo, Yvonne Lussi, Alistair MacDougall, Fabio Madeira, Mahdi Mahmoudy, Manuela Menchi, Alok Mishra, Katie Moulang, Andrew Nightingale, Carla Susana Oliveira, Sangya Pundir, Guoying Qi, Shriya Raj, Daniel Rice, Milagros Rodriguez Lopez, Rabie Saidi, Joseph Sampson, Tony Sawford, Elena Speretta, Edward Turner, Nidhi Tyagi, Preethi Vasudev, Vladimir Volynkin, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie Claude Blatter, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casals-Casas, Edouard de Castro, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Guillaume Keller, Arnaud Kerhornou, Vicente Lara, Philippe Le

Mercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Batista Neto, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Christian Sigrist, Karin Sonesson, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S. Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai Su Yeh, Jian Zhang, Patrick Ruch, and Douglas Teodoro. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49, 2021.

[340] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. Uniref: Comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23, 2007.

[341] Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R.N. Tivey, Simon C. Potter, Robert D. Finn, and Rodrigo Lopez. The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic Acids Research*, 47, 2019.

[342] Gaia Cantelli, Alex Bateman, Cath Brooksbank, Anton I. Petrov, Rahuman S. Malik-Sheriff, Michele Ide-Smith, Henning Hermjakob, Paul Flicek, Rolf Apweiler, Ewan Birney, and Johanna McEntyre. The european bioinformatics institute (embl-ebi) in 2021. *Nucleic Acids Research*, 50, 2022.

[343] Jose Manuel Rodriguez, Fernando Pozo, Daniel Cerdán-Velez, Tomás Di Domenico, Jesús Vázquez, and Michael L. Tress. Appris: Selecting functionally important isoforms. *Nucleic Acids Research*, 50, 2022.

[344] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. Dbsnp: The ncbi database of genetic variation. *Nucleic Acids Research*, 29, 2001.

[345] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. Ncbi reference se-

quences (refseq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35, 2007.

[346] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. Ncbi reference sequence (refseq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33, 2005.

[347] Jacob Rodriguez, Siddharth Rath, Jonathan Francis-Landau, Yekta Demirci, Burak Berk Üstündağ, and Mehmet Sarikaya. A generalized similarity metric for predicting peptide binding affinity. *A Generalized Similarity Metric for Predicting Peptide Binding Affinity*, 2019.

[348] Randi Vita, Swapnil Mahajan, James A. Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R. Cantrell, Daniel K. Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (iedb): 2018 update. *Nucleic Acids Research*, 47, 2019.

[349] Daniele Raimondi, Gabriele Orlando, Wim F. Vranken, and Yves Moreau. Exploring the limitations of biophysical propensity scales coupled with machine learning for protein sequence analysis. *Scientific Reports*, 9, 2019.

[350] Dan Hu, Xueliang Li, Xiaogang Liu, and Shenggui Zhang. The von neumann entropy of random multipartite graphs. *Discrete Applied Mathematics*, 232, 2017.

[351] Zhonghao Liu, Yuxin Cui, Zheng Xiong, Alierza Nasiri, Ansi Zhang, and Jianjun Hu. Deepseqpan, a novel deep convolutional neural network model for pan-specific class i hla-peptide binding affinity prediction. *Scientific Reports*, 9, 2019.

[352] Usman Ghani, Israel Desta, Akhil Jindal, Omeir Khan, George Jones, Sergey Kotelnikov, Dzmitry Padhorny, Sandor Vajda, and Dima Kozakov. Improved docking of protein models by a combination of alphafold2 and cluspro. *bioRxiv*, 2021.

[353] Tamás Hegedűs, Markus Geisler, Gergely László Lukács, and Bianka Farkas. Ins

and outs of alphafold2 transmembrane protein structure predictions. *Cellular and Molecular Life Sciences*, 79, 2022.

[354] Nazim Bouatta, Peter Sorger, and Mohammed AlQuraishi. Protein structure prediction by alphafold2: Are attention and symmetries all you need? *Acta Crystallographica Section D: Structural Biology*, 77, 2021.

[355] Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using alphafold2. *Nature Communications*, 13, 2022.

[356] Claus Lundegaard, Kasper Lamberth, Mikkel Harndahl, Søren Buus, Ole Lund, and Morten Nielsen. Netmhc-3.0: accurate web accessible predictions of human, mouse and monkey mhc class i affinities for peptides of length 8-11. *Nucleic acids research*, 36, 2008.

[357] Ilka Hoof, Bjoern Peters, John Sidney, Lasse Eggers Pedersen, Alessandro Sette, Ole Lund, Søren Buus, and Morten Nielsen. Netmhcpan, a method for mhc class i binding prediction beyond humans. *Immunogenetics*, 61, 2009.

[358] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. Netmhcpan-4.1 and netmhciipan-4.0: Improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic Acids Research*, 48, 2021.

[359] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12, 2021.

[360] Mohammed Alquraishi. End-to-end differentiable learning of protein structure a i m t l article end-to-end differentiable learning of protein structure. *Cell Systems*, 8, 2019.

[361] Mohammed AlQuraishi and Peter K. Sorger. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nature Methods*, 18, 2021.

[362] Mohammed AlQuraishi, Shengdong Tang, and Xide Xia. An affinity-structure

database of helix-turn-helix: Dna complexes with a universal coordinate system. *BMC Bioinformatics*, 16, 2015.

[363] Lin Zhu, Mehdi D. Davari, and Wenjin Li. Recent advances in the prediction of protein structural classes: Feature descriptors and machine learning algorithms. *Crystals*, 11, 2021.

[364] Jorge A. Anaya-Contreras, Héctor M. Moya-Cessa, and Arturo Zúñiga-Segundo. The von neumann entropy for mixed states. *Entropy*, 21, 2019.

[365] Kevin K. Yang, Zachary Wu, Claire N. Bedbrook, and Frances H. Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34, 2018.

[366] Kaifu Gao, Duc Duy Nguyen, Vishnu Sresht, Alan M. Mathiowetz, Meihua Tu, and Guo Wei Wei. Are 2d fingerprints still valuable for drug discovery? *Physical Chemistry Chemical Physics*, 22, 2020.

[367] Romanos Fasoulis, Georgios Paliouras, and Lydia E. Kavraki. Graph representation learning for structural proteomics. *Emerging Topics in Life Sciences*, 5, 2021.

[368] Benoit Playe and Veronique Stoven. Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity. *Journal of Cheminformatics*, 12, 2020.

[369] Rıza Özçelik, Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Chemboost: A chemical language based approach for protein – ligand binding affinity prediction. *Molecular Informatics*, 40, 2021.

[370] Mohammed Alquraishi, Grigoriy Koytiger, Anne Jenney, Gavin Macbeath, and Peter K. Sorger. A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks. *Nature Genetics*, 46, 2014.

[371] Zhao Chen, Yang Liu, and Hao Sun. Physics-informed learning of governing equations from scarce data. *Nature Communications*, 12, 2021.

[372] Bi Zhao, Akila Katuwawala, Christopher J. Oldfield, Gang Hu, Zhonghua Wu,

Vladimir N. Uversky, and Lukasz Kurgan. Intrinsic disorder in human rna-binding proteins. *Journal of Molecular Biology*, 433, 2021.

[373] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3, 2021.

[374] Vikas Dwivedi and Balaji Srinivasan. Physics informed extreme learning machine (pielm)–a rapid method for the numerical solution of partial differential equations. *Neurocomputing*, 391, 2020.

[375] Guofei Pang and George Em Karniadakis. *Physics-Informed Learning Machines for Partial Differential Equations: Gaussian Processes Versus Neural Networks*. 2020.

[376] V. T. and John C. Strikwerda. Finite difference schemes and partial differential equations. *Mathematics of Computation*, 55, 1990.

[377] C. M. Elliott and T. Ranner. A unified theory for continuous-in-time evolving finite element space approximations to partial differential equations in evolving domains. *IMA Journal of Numerical Analysis*, 41, 2021.

[378] Silke Prohl. Finite element methods for partial differential equations for option pricing. *SSRN Electronic Journal*, 2019.

[379] Davies. The finite element method an introduction with partial differential equations. *Journal of Chemical Information and Modeling*, 53, 2019.

[380] R. Rannacher. Adaptive galerkin finite element methods for partial differential equations. *Journal of Computational and Applied Mathematics*, 128, 2001.

[381] Yanan Guo, Xiaoqun Cao, Bainian Liu, and Mei Gao. Solving partial differential equations using deep learning and physical constraints. *Applied Sciences (Switzerland)*, 10, 2020.

[382] Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM Review*, 63, 2021.

[383] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural net-

works: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 2019.

[384] Arnd Koeppe, Franz Bamer, Michael Selzer, Britta Nestler, and Bernd Markert. Explainable artificial intelligence for mechanics: Physics-explaining neural networks for constitutive models. *Frontiers in Materials*, 8, 2022.

[385] Amirhossein Arzani, Jian Xun Wang, and Roshan M. D'Souza. Uncovering near-wall blood flow from sparse data with physics-informed neural networks. *Physics of Fluids*, 33, 2021.

[386] Khemraj Shukla, Ameya D. Jagtap, and George Em Karniadakis. Parallel physics-informed neural networks via domain decomposition. *Journal of Computational Physics*, 447, 2021.

[387] Liu Yang, Xuhui Meng, and George Em Karniadakis. B-pinns: Bayesian physics-informed neural networks for forward and inverse pde problems with noisy data. *Journal of Computational Physics*, 425, 2021.

[388] Shengze Cai, Zhiping Mao, Zhicheng Wang, Minglang Yin, and George Em Karniadakis. Physics-informed neural networks (pinns) for fluid mechanics: a review. *Acta Mechanica Sinica/Lixue Xuebao*, 37, 2021.

[389] Ehsan Kharazmi, Zhongqiang Zhang, and George E.M. Karniadakis. hp-vpinns: Variational physics-informed neural networks with domain decomposition. *Computer Methods in Applied Mechanics and Engineering*, 374, 2021.

[390] Zhiwei Fang. A high-efficient hybrid physics-informed neural networks based on convolutional neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[391] Taufeq Mohammed Razakh, Beibei Wang, Shane Jackson, Rajiv K. Kalia, Aiichiro Nakano, Ken ichi Nomura, and Priya Vashishta. Pnd: Physics-informed neural-network software for molecular dynamics applications. *SoftwareX*, 15, 2021.

[392] Alexander Lindsay, Roy Stogner, Derek Gaston, Daniel Schwen, Christopher Matthews, Wen Jiang, Larry K. Aagesen, Robert Carlsen, Fande Kong, Andrew Slaughter, Cody Permann, and Richard Martineau. Automatic differentiation in metaphysicl and its applications in moose. *Nuclear Technology*, 207, 2021.

[393] Damiano Mazza and Michele Pagani. Automatic differentiation in pcf. *Proceedings of the ACM on Programming Languages*, 5, 2021.

[394] Conal Elliott. The simple essence of automatic differentiation. *Proceedings of the ACM on Programming Languages*, 2, 2018.

[395] Jin Guo Liu and Kai Lai Xu. Automatic differentiation and its applications in physics simulation. *Wuli Xuebao/Acta Physica Sinica*, 70, 2021.

[396] Alan Julian Izenman. Random matrix theory and its applications. *Statistical Science*, 36, 2021.

[397] Luis Aparicio, Mykola Bordyuh, Andrew J. Blumberg, and Raul Rabadan. A random matrix theory approach to denoise single-cell data. *Patterns*, 1, 2020.

[398] David S. Dean, Pierre Le Doussal, Satya N. Majumdar, and Grégory Schehr. Noninteracting fermions in a trap and random matrix theory. *Journal of Physics A: Mathematical and Theoretical*, 52, 2019.

[399] Sayantan Choudhury and Arkaprava Mukherjee. A bound on quantum chaos from random matrix theory with gaussian unitary ensemble. *Journal of High Energy Physics*, 2019, 2019.

[400] Nicholas P. Baskerville, Diego Granziol, and Jonathan P. Keating. Appearance of random matrix theory in deep learning. *Physica A: Statistical Mechanics and its Applications*, 590, 2022.

[401] Gang Guo and Gabriel Martínez-Pinedo. Chiral effective field theory description of neutrino nucleon–nucleon bremsstrahlung in supernova matter. *The Astrophysical Journal*, 887, 2019.

[402] M. V. Berry and Pragya Shukla. Quantum metric statistics for random-matrix families. *Journal of Physics A: Mathematical and Theoretical*, 53, 2020.

[403] Pavel Kos, Marko Ljubotina, and Tomaž Prosen. Many-body quantum chaos: Analytic connection to random matrix theory. *Physical Review X*, 8, 2018.

[404] A. V. Andreev, O. Agam, B. D. Simons, and B. L. Altshuler. Quantum chaos, irreversible classical dynamics, and random matrix theory. *Physical Review Letters*, 76, 1996.

[405] Denis Ullmo. Bohigas-giannoni-schmit conjecture. *Scholarpedia*, 11, 2016.

[406] Mario Kieburg, Jacobus J.M. Verbaarschot, and Savvas Zafeiropoulos. Spectral properties of the wilson-dirac operator and random matrix theory. *Physical Review D - Particles, Fields, Gravitation and Cosmology*, 88, 2013.

[407] Jacobus Verbaarschot. Spectrum of the qcd dirac operator and chiral random matrix theory. *Physical Review Letters*, 72, 1994.

[408] Edward Witten. Matrix models and deformations of jt gravity: Matrix models and deformations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476, 2020.

[409] Yang Li, Zheng Wang, Zhu Hong You, Li Ping Li, and Xuegang Hu. Predicting protein-protein interactions via random ferns with evolutionary matrix representation. *Computational and Mathematical Methods in Medicine*, 2022, 2022.

[410] I. L. Aleiner, P. W. Brouwer, and L. I. Glazman. Quantum effects in coulomb blockade. *Physics Reports*, 358, 2002.

[411] Andrea Cappelli, Carlo A. Trugenberger, and Guillermo R. Zemba. Large n limit in the quantum hall effect. *Physics Letters B*, 306, 1993.

[412] David J.E. Callaway. Random matrices, fractional statistics, and the quantum hall effect. *Physical Review B*, 43, 1991.

[413] C. W.J. Beenakker. Random-matrix theory of majorana fermions and topological superconductors. *Reviews of Modern Physics*, 87, 2015.

[414] Damir Herman, T. Tzen Ong, Gonzalo Usaj, Harsh Mathur, and Harold U. Baranger. Level spacings in random matrix theory and coulomb blockade peaks

in quantum dots. *Physical Review B - Condensed Matter and Materials Physics*, 76, 2007.

[415] Lucas Cuadra and José Carlos Nieto-Borge. Modeling quantum dot systems as random geometric graphs with probability amplitude-based weighted links. *Nanomaterials*, 11, 2021.

[416] Richard Peng and Santosh Vempala. Solving sparse linear systems faster than matrix multiplication. 2021.

[417] Wei Ting Chang and Ravi Tandon. Random sampling for distributed coded matrix multiplication. volume 2019-May, 2019.

[418] Joël Bun, Jean Philippe Bouchaud, and Marc Potters. Overlaps between eigenvectors of correlated random matrices. *Physical Review E*, 98, 2018.

[419] Joël Bun, Jean Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666, 2017.

[420] Malik Tiomoko, Florent Bouchard, Guillaume Ginolhac, and Romain Couillet. Random matrix improved covariance estimation for a large class of metrics. *Journal of Statistical Mechanics: Theory and Experiment*, 2020, 2020.

[421] Noureddine El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics*, 36, 2008.

[422] Kion Fallah, Adam A. Willats, Ninghao Liu, and Christopher J. Rozell. Learning sparse codes from compressed representations with biologically plausible local wiring constraints. volume 2020-December, 2020.

[423] E. Gudowska-Nowak, M. A. Nowak, D. R. Chialvo, J. K. Ochab, and W. Tarnowski. From synaptic interactions to collective dynamics in random neuronal networks models: Critical role of eigenvectors and transient behavior. *Neural Computation*, 32, 2020.

[424] Carina Curto, Anda Degeratu, and Vladimir Itskov. Flexible memory networks. *Bulletin of Mathematical Biology*, 74, 2012.

[425] Paul Bourgade, Horng-Tzer Yau, and Jun Yin. Random band matrices in

the delocalized phase, i: Quantum unique ergodicity and universality. *arXiv preprint arXiv:1807.01559*, 2018.

[426] Paul Bourgade, Horng Tzer Yau, and Jun Yin. Random band matrices in the delocalized phase i: Quantum unique ergodicity and universality. *Communications on Pure and Applied Mathematics*, 73, 2020.

[427] Paul Bourgade, Laszlo Erdoos, Horng Tzer Yau, and Jun Yin. Universality for a class of random band matrices. *Advances in Theoretical and Mathematical Physics*, 21, 2017.

[428] Seth Lloyd. Quantum coherence in biological systems. volume 302, 2011.

[429] Johnjoe McFadden and Jim Al-Khalili. A quantum mechanical model of adaptive mutation. *BioSystems*, 50, 1999.

[430] Johnjoe McFadden and Jim Al-Khalili. The origins of quantum biology. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474, 2018.

[431] Youngchan Kim, Federico Bertagna, Edeline M. D'souza, Derren J. Heyes, Linus O. Johannissen, Eveliny T. Nery, Antonio Pantelias, Alejandro Sanchez Pedreño Jimenez, Louie Slocombe, Michael G. Spencer, Jim Al-Khalili, Gregory S. Engel, Sam Hay, Suzanne M. Hingley-Wilson, Kamalan Jeevaratnam, Alex R. Jones, Daniel R. Kattnig, Rebecca Lewis, Marco Sacchi, Nigel S. Scrutton, S. Ravi P. Silva, and Johnjoe McFadden. Quantum biology: An update and perspective. *Quantum Reports*, 3, 2021.

[432] AW Chin, Susana F Huelga, and Martin B Plenio. Coherence and decoherence in biological systems: principles of noise-assisted transport and the origin of long-lived coherences. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1972):3638–3657, 2012.

[433] Hong Bin Chen, Pin Yi Chiu, and Yueh Nan Chen. Vibration-induced coherence enhancement of the performance of a biological quantum heat engine. *Physical Review E*, 94, 2016.

[434] Edward T. Drabold and David J. Bayless. Quantitative photoresponse of the first photosynthetic biomaterials: Physical measurements and analysis of microalgae systems. *Physica Status Solidi (B) Basic Research*, 258, 2021.

[435] S. V. Kozyrev and I. V. Volovich. Dark states in quantum photosynthesis, 2018.

[436] Cvetelin Vasilev, Guy E. Mayneord, Amanda A. Brindley, Matthew P. Johnson, and C. Neil Hunter. Dissecting the cytochrome c2–reaction centre interaction in bacterial photosynthesis using single molecule force spectroscopy. *Biochemical Journal*, 476, 2019.

[437] Adriana Marais, Ilya Sinayskiy, Francesco Petruccione, and Rienk Van Grondelle. A quantum protective mechanism in photosynthesis. *Scientific Reports*, 5, 2015.

[438] Pavel Malý, J. Michael Gruber, Richard J. Cogdell, Tomá Mančal, and Rienk Van Grondelle. Ultrafast energy relaxation in single light-harvesting complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 2016.

[439] Jim Al-Khalili Johnjoe McFadden.

[440] L. Slocombe, J. S. Al-Khalili, and M. Sacchi. Quantum and classical effects in dna point mutations: Watson-crick tautomerism in at and gc base pairs. *Physical Chemistry Chemical Physics*, 23, 2021.

[441] Gizem Çelebi, Elif Özçelik, Emre Vardar, and Durmuş Demir. Time delay during the proton tunneling in the base pairs of the dna double helix. *Progress in Biophysics and Molecular Biology*, 167, 2021.

[442] Jennifer C. Brookes. Quantum effects in biology: Golden rule in enzymes, olfaction, photosynthesis and magnetodetection. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473, 2017.

[443] Eric R. Bittner, Adrian Madalan, Arkadiusz Czader, and Gregg Roman. Quantum origins of molecular recognition and olfaction in drosophila. *Journal of Chemical Physics*, 137, 2012.

[444] Ann Sophie Barwich. How to be rational about empirical success in ongoing science: The case of the quantum nose and its critics. *Studies in History and Philosophy of Science Part A*, 69, 2018.

[445] Sönke Johnsen, Erin Mattern, and Thorsten Ritz. Light-dependent magnetoreception: Quantum catches and opponency mechanisms of possible photosensitive molecules. *Journal of Experimental Biology*, 210, 2007.

[446] Jianming Cai and Martin B. Plenio. Chemical compass model for avian magnetoreception as a quantum coherent device. *Physical Review Letters*, 111, 2013.

[447] S. Ghosh, M. Petrin, and A. Maki. Spin-lattice relaxation in the triplet state of the buried tryptophan residue of ribonuclease t1. *Biophysical Journal*, 49, 1986.

[448] Klaus Schulten, Charles E. Swenberg, and Albert Weiler. A biomagnetic sensory mechanism based on magnetic field modulated coherent electron spin motion. *Zeitschrift fur Physikalische Chemie*, 111, 1978.

[449] Ernst Walter Knapp and Klaus Schulten. Magnetic field effect on the hyperfine-induced electron spin motion in radicals undergoing diamagnetic-paramagnetic exchange. *The Journal of Chemical Physics*, 71, 1979.

[450] Klaus Schulten. Ensemble averaged spin pair dynamics of doublet and triplet molecules. *The Journal of Chemical Physics*, 80, 1984.

[451] Jim Al-Khalili and Samuele Lilliu. Quantum biology. *Scientific Video Protocols*, 1, 2020.

[452] Cyril W. Smith. Quanta and coherence effects in water and living systems. *Journal of Alternative and Complementary Medicine*, 10, 2004.

[453] Jianfeng Bao, Xiaohong Cui, Shuhui Cai, Jianhui Zhong, Congbo Cai, and Zhong Chen. Brown adipose tissue mapping in rats with combined intermolecular double-quantum coherence and dixon water-fat mri. *NMR in Biomedicine*, 26, 2013.

[454] Onur Pusuluk, Tristan Farrow, Cemsinan Deliduman, and Vlatko Vedral. Emer-

gence of correlated proton tunnelling in water ice. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475, 2019.

[455] V. Garbuio, C. Andreani, S. Imberti, A. Pietropaolo, G. F. Reiter, R. Senesi, and M. A. Ricci. Proton quantum coherence observed in water confined in silica nanopores. *Journal of Chemical Physics*, 127, 2007.

[456] Arieh Warshel, Arno Papazyan, Peter A. Kollman, W. W. Cleland, Maurice M. Kreevoy, and Perry A. Frey. On low-barrier hydrogen bonds and enzyme catalysis. *Science*, 269, 1995.

[457] Jacob D. Graham, Allyson M. Buytendyk, Di Wang, Kit H. Bowen, and Kim D. Collins. Strong, low-barrier hydrogen bonds may be available to enzymes. *Biochemistry*, 53, 2014.

[458] Yongho Kim and Kwang Hyun Ahn. Theoretical study of the role of low-barrier hydrogen bonds in enzyme catalysis: A model of proton transfer in serine protease. *Theoretical Chemistry Accounts*, 106, 2001.

[459] Prashasti Kumar, Pratul K. Agarwal, M. Brett Waddell, Tanja Mittag, Engin H. Serpersu, and Matthew J. Cuneo. Low-barrier and canonical hydrogen bonds modulate activity and specificity of a catalytic triad. *Angewandte Chemie - International Edition*, 58, 2019.

[460] Agback, P., and T. Agback. Direct evidence of a low barrier hydrogen bond in the catalytic triad of a serine protease. *Nature*, 2018.

[461] Shaobo Dai, Lisa Marie Funk, Fabian Rabe von Pappenheim, Viktor Sautner, Mirko Paulikat, Benjamin Schröder, Jon Uranga, Ricardo A. Mata, and Kai Tittmann. Low-barrier hydrogen bonds in enzyme cooperativity. *Nature*, 573, 2019.

[462] Shaobo Dai, Lisa-Marie Funk, Fabian Rabe von Pappenheim, Viktor Sautner, Mirko Paulikat, Benjamin Shroder, Jon Uranga, Ricardo Mata, and Kai Tittmann. Low-barrier hydrogen bonds in enzyme cooperativity. *Nature*, 2019.

[463] Irena Cosic, Anthony N Hodder, Marie-Isabel Aguilar, and Milton TW Hearn.

Resonant recognition model and protein topography: model studies with myoglobin, hemoglobin and lysozyme. *European journal of biochemistry*, 198(1):113–119, 1991.

[464] Irena COSIC, Anthony N. HODDER, Marie-Isabel -I AGUILAR, and Milton T.W. HEARN. Resonant recognition model and protein topography: Model studies with myoglobin, hemoglobin and lysozyme. *European Journal of Biochemistry*, 198, 1991.

[465] Irena Cosic, Katarina Lazar, and Drasko Cosic. Prediction of tubulin resonant frequencies using the resonant recognition model (rrm). *IEEE Transactions on Nanobioscience*, 14, 2015.

[466] Irena Cosic, Drasko Cosic, and Katarina Lazar. Tesla, bioresonances and resonant recognition model. *Second International Congress Nikola Tesla - Disruptive Innovation*, 2017.

[467] Irena Cosic, Vasilis Paspaliaris, and Drasko Cosic. Analysis of protein-receptor interactions on an example of leptin-leptin receptor interaction using the resonant recognition model. *Applied Sciences (Switzerland)*, 9, 2019.

[468] Satyajit Mahapatra and Sitanshu Sekhar Sahu. Integrating resonant recognition model and stockwell transform for localization of hotspots in tubulin. *IEEE Transactions on Nanobioscience*, 20, 2021.

[469] Susana Margarita Montesino Castillo, Cristóbal Yera Gálvez, and José Luis Hernández Cáceres. Resonant recognition model study for interactions between sars cov 2 and human proteins. *Rev. cuba. inform. méd*, 13, 2021.

[470] Elena Pirogova, Qiang Fang, Eliada Lazoura, and Irena Cosic. Analysis of amino acid parameters in the resonant recognition model. 1998.

[471] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub

Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596, 2021.

[472] Leila T. Alexander, Rosalba Lepore, Andriy Kryshtafovych, Athanassios Adamopoulos, Markus Alahuhta, Ann M. Arvin, Yannick J. Bomble, Bettina Böttcher, Cécile Breyton, Valerio Chiarini, Naga babu Chinnam, Wah Chiu, Krzysztof Fidelis, Rhys Grinter, Gagan D. Gupta, Marcus D. Hartmann, Christopher S. Hayes, Tatjana Heidebrecht, Andrea Ilari, Andrzej Joachimiak, Youngchang Kim, Romain Linares, Andrew L. Lovering, Vladimir V. Lunin, Andrei N. Lupas, Cihan Makbul, Karolina Michalska, John Moult, Prasun K. Mukherjee, William Nutt, Stefan L. Oliver, Anastassis Perrakis, Lucy Stols, John A. Tainer, Maya Topf, Susan E. Tsutakawa, Mauricio Valdivia-Delgado, and Torsten Schwede. Target highlights in casp14: Analysis of models by structure providers. *Proteins: Structure, Function and Bioinformatics*, 89, 2021.

[473] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Applying and improving alphafold at casp14. *Proteins: Structure, Function and Bioinformatics*, 89, 2021.

[474] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glass-

man, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. Van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373, 2021.

[475] Aisha Al-Janabi. Has deepmind's alphafold solved the protein folding problem? *BioTechniques*, 72, 2022.

[476] Matthew K. Higgins. Can we alphafold our way out of the next pandemic? *Journal of Molecular Biology*, 433, 2021.

[477] Xianwen Yu, Xiaoning Zhang, Yang Cao, and Min Xia. Vaegan: A collaborative filtering framework based on adversarial variational autoencoders. volume 2019-August, 2019.

[478] Mihaly Varadi, John Berrisford, Mandar Deshpande, Sreenath S. Nair, Aleksandras Gutmanas, David Armstrong, Lukas Pravda, Bissan Al-Lazikani, Stephen Anyango, Geoffrey J. Barton, Karel Berka, Tom Blundell, Neera Borkakoti, Jose Dana, Sayoni Das, Sucharita Dey, Patrizio Di Micco, Franca Fraternali, Toby Gibson, Manuela Helmer-Citterich, David Hoksza, Liang Chin Huang, Rishabh Jain, Harry Jubb, Christos Kannas, Natarajan Kannan, Jaroslav Koca, Radoslav Krivak, Manjeet Kumar, Emmanuel D. Levy, F. Madeira, M. S. Madhusudhan, Henry J. Martell, Stuart MacGowan, Jake E. McGreig, Saqib Mir, Abhik Mukhopadhyay, Luca Parca, Typhaine Paysan-Lafosse, Leandro Radusky, Antonio Ribeiro, Luis Serrano, Ian Sillitoe, Gulzar Singh, Petr Skoda, Radka Svobodova, Jonathan Tyzack, Alfonso Valencia, Eloy Villasclaras Fernandez, Wim Vranken, Mark Wass, Janet Thornton, Michael Sternberg, Christine Orengo, and Sameer Velankar. Pdbe-kb: A community-driven resource for structural and functional annotations. *Nucleic Acids Research*, 48, 2020.

[479] Akira R. Kinjo, Gert Jan Bekker, Hiroshi Wako, Shigeru Endo, Yuko Tsuchiya, Hiromu Sato, Hafumi Nishi, Kengo Kinoshita, Hirofumi Suzuki, Takeshi Kawabata, Masashi Yokochi, Takeshi Iwata, Naohiro Kobayashi, Toshimichi Fujiwara, Genji Kurisu, and Haruki Nakamura. New tools and functions in data-out activities at protein data bank japan (pdbj). *Protein Science*, 27, 2018.

[480] Masashi Yokochi, Naohiro Kobayashi, Eldon L. Ulrich, Akira R. Kinjo, Takeshi Iwata, Yannis E. Ioannidis, Miron Livny, John L. Markley, Haruki Nakamura, Chojiro Kojima, and Toshimichi Fujiwara. Publication of nuclear magnetic resonance experimental data with semantic web technology and the application thereof to biomedical research of proteins. *Journal of Biomedical Semantics*, 7, 2016.

[481] Catherine L. Lawson, Ardan Patwardhan, Matthew L. Baker, Corey Hryc, Eduardo Sanz Garcia, Brian P. Hudson, Ingvar Lagerstedt, Steven J. Ludtke, Grigore Pintilie, Raul Sala, John D. Westbrook, Helen M. Berman, Gerard J. Kleywegt, and Wah Chiu. Emdatabank unified data resource for 3dem. *Nucleic Acids Research*, 44, 2016.

[482] Paul D. Adams, Pavel V. Afonine, Kumaran Baskaran, Helen M. Berman, John Berrisford, Gerard Bricogne, David G. Brown, Stephen K. Burley, Minyu Chen, Zukang Feng, Claus Flensburg, Aleksandras Gutmanas, Jeffrey C. Hoch, Yasuyo Ikegawa, Yumiko Kengaku, Eugene Krissinel, Genji Kurisu, Yuhe Liang, Dorothee Liebschner, Lora Mak, John L. Markley, Nigel W. Moriarty, Garib N. Murshudov, Martin Noble, Ezra Peisach, Irina Persikova, Billy K. Poon, Oleg V. Sobolev, Eldon L. Ulrich, Sameer Velankar, Clemens Vonrhein, John Westbrook, Marcin Wojdyr, Masashi Yokochi, and Jasmine Y. Young. Announcing mandatory submission of pdbx/mmcif format files for crystallographic depositions to the protein data bank (pdb). *Acta Crystallographica Section D: Structural Biology*, 75, 2019.

[483] John Westbrook, Nobutoshi Ito, Haruki Nakamura, Kim Henrick, and Helen M.

Berman. Pdbml: The representation of archival macromolecular structure data in xml. *Bioinformatics*, 21, 2005.

[484] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. volume 2016-December, 2016.

[485] Daria Grechishnikova. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific Reports*, 11, 2021.

[486] Nicholas Fowler and Michael P. Williamson. The accuracy of protein structures in solution determined by alphafold and nmr. *SSRN Electronic Journal*, 2022.

[487] Yuri D. Ivanov, Amir Taldaev, Andrey V. Lisitsa, Elena A. Ponomarenko, and Alexander I. Archakov. Prediction of monomeric and dimeric structures of cyp102a1 using alphafold2 and alphafold multimer and assessment of point mutation effect on the efficiency of intra-and interprotein electron transfer. *Molecules*, 27, 2022.

[488] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2022.

[489] Markus Zweckstetter. Nmr hawk-eyed view of alphafold2 structures. *Protein Science*, 30, 2021.

[490] T. Reid Alderson, Iva Pritišanac, Alan M. Moses, and Julie D. Forman-Kay. Systematic identification of conditionally folded intrinsically disordered regions by alphafold2. *bioRxiv*, 2022.

[491] Kiersten M. Ruff and Rohit V. Pappu. Alphafold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433, 2021.

[492] Isak Johansson- Akhe and Björn Wallner. Benchmarking peptide-protein docking and interaction prediction with alphafold-multimer. *bioRxiv*, 2021.

[493] Marina A. Pak, Karina A. Markhieva, Mariia S. Novikova, Dmitry S. Petrov, Ilya S. Vorobyev, Ekaterina S. Maksimova, Fyodor A. Kondrashov, and Dmitry N. Ivankov. Using alphafold to predict the impact of single mutations on protein stability and function. *bioRxiv*, 1, 2021.

[494] Alessia David, Suhail Islam, Evgeny Tankhilevich, and Michael J.E. Sternberg. The alphafold database of protein structures: A biologist's guide. *Journal of Molecular Biology*, 434, 2022.

[495] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Zídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50, 2022.

[496] Guy Jacoby, Merav Segal Asher, Tamara Ehm, Inbal Abutbul Ionita, Hila Shinar, Salome Azoulay-Ginsburg, Ido Zemach, Gil Koren, Dganit Danino, Michael M. Kozlov, Roey J. Amir, and Roy Beck. Order from disorder with intrinsically disordered peptide amphiphiles. *Journal of the American Chemical Society*, 143, 2021.

[497] Steven J. Metallo. Intrinsically disordered proteins are potential drug targets. *Current Opinion in Chemical Biology*, 14, 2010.

[498] Sarah E. Bondos, A. Keith Dunker, and Vladimir N. Uversky. On the roles of intrinsically disordered proteins and regions in cell communication and signaling. *Cell Communication and Signaling*, 19, 2021.

[499] Peter Tompa, Norman E. Davey, Toby J. Gibson, and M. Madan Babu. A million peptide motifs for the molecular biologist. *Molecular Cell*, 55, 2014.

[500] Manjeet Kumar, Marc Gouw, Sushama Michael, Hugo Sámano-Sánchez, Rita

Pancsa, Juliana Glavina, Athina Diakogianni, Jesús Alvarado Valverde, Dayana Bukirova, Jelena Signalyševa, Nicolas Palopoli, Norman E. Davey, Lucía B. Chemes, and Toby J. Gibson. Elm-the eukaryotic linear motif resource in 2020. *Nucleic Acids Research*, 48, 2020.

[501] Eugene Lin, Sudipto Mukherjee, and Sreeram Kannan. A deep adversarial variational autoencoder model for dimensionality reduction in single-cell rna sequencing analysis. *BMC Bioinformatics*, 21, 2020.

[502] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63, 2020.

[503] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[504] Jia Luo and Jinying Huang. Generative adversarial network: An overview. *Yi Qi Yi Biao Xue Bao/Chinese Journal of Scientific Instrument*, 40, 2019.

[505] Chao Tao, Hao Wang, Ji Qi, and Haifeng Li. Semisupervised variational generative adversarial networks for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 2020.

[506] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404, 2020.

[507] Cem Eteke. Wasserstein gan : Technical report. *COMP.541 Sprint Final Report*, 2018.

[508] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. volume 2017-December, 2017.

[509] Yuhong Zhang, Yuling Li, Yi Zhu, and Xuegang Hu. Wasserstein gan based

on autoencoder with back-translation for cross-lingual embedding mappings. *Pattern Recognition Letters*, 129, 2020.

[510] Madan Lal Mehta. *Random matrices*. Elsevier, 2004.

[511] Wen Jia Rao. Higher-order level spacings in random matrix theory based on wigner's conjecture. *Physical Review B*, 102, 2020.

[512] Peter J. Forrester. Analogies between random matrix ensembles and the one-component plasma in two-dimensions. *Nuclear Physics B*, 904, 2016.

[513] Sumeet Kulkarni, Khun Sang Phukon, Amit Reza, Sukanta Bose, Anirban Dasgupta, Dilip Krishnaswamy, and Anand S. Sengupta. Random projections in gravitational wave searches of compact binaries. *Physical Review D*, 99, 2019.

[514] Vladimir Vasilchuk. On the gaussian random matrix ensembles with additional symmetry conditions. *Symmetry, Integrability and Geometry: Methods and Applications (SIGMA)*, 2, 2006.

[515] Margherita Disertori, Martin Lohmann, and Sasha Sodin. The density of states of 1d random band matrices via a supersymmetric transfer operator. *Journal of Spectral Theory*, 11, 2021.

[516] Ioana Dumitriu and Tobias Johnson. The marčenko-pastur law for sparse random bipartite biregular graphs. *Random Structures and Algorithms*, 48, 2016.

[517] L. A. Pastur. A simple approach to the global regime of gaussian ensembles of random matrices. *Ukrainian Mathematical Journal*, 57, 2005.

[518] Ji Oon Lee and Kevin Schnelli. Local law and tracy–widom limit for sparse random matrices. *Probability Theory and Related Fields*, 171, 2018.

[519] T. W. Anderson. The non-central wishart distribution and certain problems of multivariate statistics. *The Annals of Mathematical Statistics*, 17, 1946.

[520] Frédéric Ouimet. A symmetric matrix-variate normal local approximation for the wishart distribution and some applications. *Journal of Multivariate Analysis*, 189, 2022.

[521] Paul Bourgade. Random band matrices. volume 4, 2018.

[522] P. Bourgade, F. Yang, H. T. Yau, and J. Yin. Random band matrices in the delocalized phase, ii: Generalized resolvent estimates. *Journal of Statistical Physics*, 174, 2019.

[523] Danko D. Georgiev and James F. Glazebrook. On the quantum dynamics of davydov solitons in protein -helices. *Physica A: Statistical Mechanics and its Applications*, 517, 2019.

[524] C. P. Dettmann, O. Georgiou, and G. Knight. Spectral statistics of random geometric graphs. *EPL*, 118, 2017.

[525] Prasanna Sahoo, Carrye Wilkins, and Jerry Yeager. Threshold selection using renyi's entropy. *Pattern Recognition*, 30, 1997.

[526] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindah. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2, 2015.

[527] Frank Jensen. *Introduction to computational chemistry*. John wiley & sons, 2017.

[528] Jan Kubelka, James Hofrichter, and William A Eaton. The protein folding 'speed limit'. *Folding and Design*, 14(1):76–88, 2004.

[529] Utsab R Shrestha, Jeremy C Smith, and Loukas Petridis. Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. *Communications biology*, 4(1):1–8, 2021.

[530] Martin B Plenio and Susana F Huelga. Dephasing-assisted transport: quantum networks and biomolecules. *New Journal of Physics*, 10(11):113019, 2008.

[531] Thomas Steiner. The hydrogen bond in the solid state. *Angewandte Chemie International Edition*, 41(1):48–76, 2002.

[532] Yangqi Gu, Vishok Srikanth, Aldo I. Salazar-Morales, Ruchi Jain, J. Patrick O'Brien, Sophia M. Yi, Rajesh Kumar Soni, Fadel A. Samatey, Sibel Ebru

Yalcin, and Nikhil S. Malvankar. Structure of geobacter pili reveals secretory rather than nanowire behaviour. *Nature*, September 2021.

[533] Bruce Damer and David Deamer. The hot spring hypothesis for an origin of life. *Astrobiology*, 20(4):429–452, 2020.

[534] Daniel Milshteyn, Bruce Damer, Jeff Havig, and David Deamer. Amphiphilic compounds assemble into membranous vesicles in hydrothermal hot spring water but not in seawater. *Life*, 8(2):11, 2018.

[535] David Deamer, Bruce Damer, and Vladimir Kompanichenko. Hydrothermal chemistry and the origin of cellular life. *Astrobiology*, 19(12):1523–1537, 2019.

[536] C. Wochnowski and C Wochnowski. A very simple model concerning the unified field theory basing on the kronig-penney-model. *Journal of High Energy Physics, Gravitation and Cosmology*, 05, 01 2019.

[537] Ida V Lundholm, Helena Rodilla, Weixiao Y Wahlgren, Annette Duelli, Gleb Bourenkov, Josip Vukusic, Ran Friedman, Jan Stake, Thomas Schneider, and Gergely Katona. Terahertz radiation induces non-thermal structural changes associated with fröhlich condensation in a protein crystal. *Structural Dynamics*, 2(5):054702, 2015.

Appendix A

CODE REPOSITORY

All code can be found at `https://github.com/rathsidd/QuantropyPro`

VITA

Siddharth Rath was a Ph.D. student at the University of Washington in the departments of Materials Science and Engineering and Molecular Engineering and Sciences from 2016-2022. His interests lay in applying data science to probe the unknown fundamental phenomena governing biology, at the confluence of mathematics, information theory, molecular biology and quantum mechanics. Any questions may be directed to rathsiduw@gmail.com or rathdidd@uw.edu or rathsid1991@gmail.com.