

JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Arya Changiarath Sivadasan

written by

Multi-Scale Simulations of Protein Phase Separation in Gene Regulation

This dissertation is submitted to the Faculty of Physics, Mathematics and Informatics of the Johannes Gutenberg University of Mainz for the degree of “Doctor of Natural Sciences”

Multi-Scale Simulations of Protein Phase Separation in Gene Regulation

Arya Changiarath Sivadasan

Doctoral Dissertation: June 17, 2025

Supervisors: Prof. Dr. Friederike Schmid and Prof. Dr. Lukas S Stelzl

Abstract

The formation of distinct phase-separated condensates of bio-macromolecules underpins specific regulation in cells. In this thesis, I elucidate how phase-separated condensates regulate biochemical processes by providing distinct chemical environments through particle-based multi-scale simulations. The disordered C-terminal domain of RNA polymerase II (CTD) forms condensates in both unphosphorylated and phosphorylated (pCTD) states. Understanding the molecular driving forces behind CTD phase separation offers insights into how these condensates regulate transcription.

To uncover the molecular determinants of CTD phase separation, I integrate nuclear magnetic resonance (NMR) data with atomistic molecular dynamics simulations using the refined hierarchical chain growth (RHCG) method. This reveals local structural preferences and highlights the key role of tyrosine-proline contacts in driving condensation. These insights aid in validating the coarse-grained simulations, which capture how temperature and post-translational modifications (PTMs) influence condensate dynamics. I further explore whether unphosphorylated and phosphorylated CTD form mixed or distinct condensates by calculating interfacial properties from coarse-grained models.

To understand how phosphorylation modulates CTD condensate behaviour, an algorithm for simulating stochastic phosphorylation dynamics is developed. This approach provides a mechanistic understanding of how phosphorylation alters the phase behaviour of CTD condensates. By dynamically tuning protein interactions, phosphorylation modulates condensate stability, facilitating transitions between different functional states. This coupling between phosphorylation dynamics and phase separation highlights the regulatory potential of post-translational modifications in transcriptional condensates.

Finally, I introduce an active learning framework combining Bayesian optimization and coarse-grained simulations to study sequence determinants of phase behavior. This approach identifies sequence-property relationships for disordered proteins, including their self-interactions and interactions in phase-separated condensates. Using second virial coefficients, I predict peptide self-interactions and identify sequences that bind to CTD condensates. This approach allows the rational design of multiphasic condensates with tunable morphologies.

These findings highlight the importance of phase separation and post-translational modifications in regulating transcription. The active learning methods offer key insights for designing functional condensates and controlling cellular processes.

Zusammenfassung

Die Bildung unterschiedlicher phasengetrennter Kondensate von Biomakromolekülen bildet die Grundlage für eine spezifische Regulierung in Zellen. In der vorliegenden Arbeit wird dargelegt, wie phasengetrennte Kondensate die Regulation biochemischer Prozesse bewirken, indem sie unterschiedliche chemische Umgebungen mittels partikelbasierter Multiskalensimulationen bereitstellen. Die vorliegende Arbeit untersucht die Bildung von phasengetrennten Kondensaten der ungeordneten C-terminalen Domäne der RNA-Polymerase II (CTD) in verschiedenen Phosphorylierungszuständen (unphosphoryliert und phosphoryliert, pCTD). Das Verständnis der molekularen Triebkräfte hinter der CTD-Phasentrennung gibt Aufschluss darüber, wie diese Kondensate die Transkription regulieren.

Zur Aufdeckung der molekularen Determinanten der CTD-Phasentrennung werden kernmagnetische Resonanzdaten (NMR) mit atomistischen Molekulardynamiksimulationen integriert, wobei die Methode des verfeinerten hierarchischen Kettenwachstums (RHCG) Anwendung findet. Hierdurch werden lokale strukturelle Präferenzen aufgedeckt und die Schlüsselrolle von Tyrosin-Prolin-Kontakten bei der Kondensation hervorgehoben. Diese Erkenntnisse tragen zur Validierung der grobkörnigen Simulationen bei, welche die Beeinflussung der Kondensatdynamik durch Temperatur und posttranslationale Modifikationen (PTMs) erfassen. Darüber hinaus wird untersucht, ob unphosphorylierte und phosphorylierte CTD-Kondensate gemischt oder unterschiedlich sind, indem die Grenzflächeneigenschaften aus grobkörnigen Modellen berechnet werden.

Um ein tieferes Verständnis dafür zu erlangen, wie die Phosphorylierung das CTD-Kondensatverhalten moduliert, wird ein Algorithmus zur Simulation der stochastischen Phosphorylierungsdynamik entwickelt. Dieser Ansatz ermöglicht ein mechanistisches Verständnis dafür, wie die Phosphorylierung das Phasenverhalten von CTD-Kondensaten verändert. Die Phosphorylierung moduliert die Kondensatstabilität und erleichtert Übergänge zwischen verschiedenen Funktionszuständen durch die dynamische Abstimmung von Proteininteraktionen. Die Kopplung zwischen Phosphorylierungsdynamik und Phasentrennung verdeutlicht das regulatorische Potenzial posttranslationaler Modifikationen in Transkriptionskondensaten.

Schließlich wird ein aktiver Lernrahmen vorgestellt, der Bayes'sche Optimierung und grobkörnige Simulationen kombiniert, um Sequenzdeterminanten des Phasenverhaltens zu untersuchen. Der dargestellte Ansatz identifiziert Sequenz-Eigenschaftsbeziehungen für ungeordnete Proteine, einschließlich ihrer Selbstwechselwirkungen und Wechselwirkungen in phasengetrennten Kondensaten. Mittels zweiter Virialkoeffizienten werden Peptid-

Selbstwechselwirkungen vorhergesagt und Sequenzen identifiziert, die an CTD-Kondensate binden. Dieser Ansatz ermöglicht das rationale Design von mehrphasigen Kondensaten mit abstimmbarer Morphologie.

Die Ergebnisse betonen die Relevanz der Phasentrennung und posttranslationaler Modifikationen bei der Regulierung der Transkription. Die Methoden des aktiven Lernens bieten wichtige Erkenntnisse für die Gestaltung funktioneller Kondensate und die Steuerung zellulärer Prozesse.gma

Contents

Abstract **i**

Zusammenfassung **iv**

1 Background **1**
1.1 How Cells Are Organized? 1
1.1.1 The Structure and Function of Proteins 3
1.1.2 Biomolecular Condensates and Phase Separation 6
1.1.3 Transcriptional condensates 9
1.1.4 Experimental Techniques for Investigating Condensates and Intrinsically Disordered Proteins 11
1.2 Theory 14
1.2.1 Thermodynamics 14
1.2.2 Understanding Polymer behaviour 19
1.2.3 Surface tension 23
1.2.4 Classical Molecular Dynamics - Theory 29
1.2.5 Sampling in molecular dynamics 32
1.2.6 Atomistic and coarse-grained simulations 33
1.2.7 Challenges in Studying Biomolecular Condensates 35
1.3 Motivation 37

2 Phase-Separated Promoter and Gene-Body Condensates: A Phosphorylation-Driven Mechanism for Transcription Control **39**
2.1 Introduction 40
2.2 Methods 41
2.3 Results and Discussion 46
2.4 Conclusion 67

3 Modelling Intrinsically Disordered Proteins: The Hierarchical Chain Growth Approach **69**
3.1 Introduction 69
3.2 Methods 73
3.3 Results and Discussions 78

3.3.1 Tyr-Pro interactions are important for the contacts of CTD with the Mediator complex . . . 87
3.4 Conclusions . . . 87

4 Stochastic Models for phosphorylation and their application in molecular dynamics simulations of protein condensates 89
4.1 Introduction . . . 89
4.2 Methods . . . 91
4.3 Results and Discussions . . . 97
4.4 Conclusion . . . 106

5 Sequence determinants of protein phase separation through molecular dynamics and active learning 107

6 Summary and outlook 129

Bibliography 133

1 Background

1.1 How Cells Are Organized?

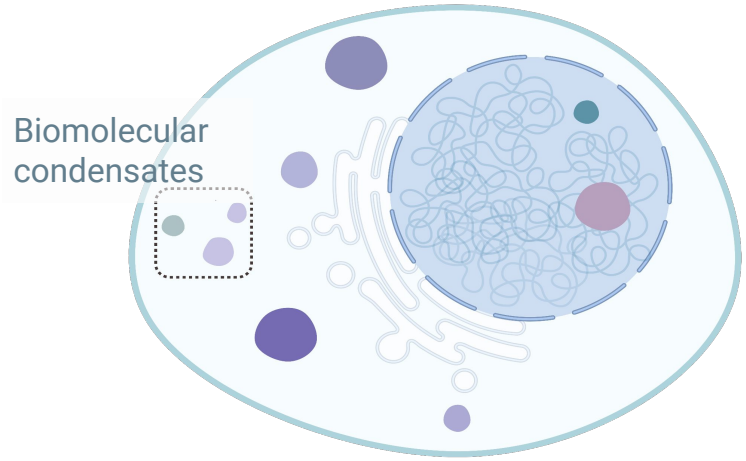

Fig. 1.1: Condensates in living cells are structures formed by liquid-liquid phase separation.

Cells are the fundamental units of life, capable of self-replication and carrying out complex biochemical processes. To efficiently organise these processes, eukaryotic cells have evolved intricate systems of compartmentalization. These compartments can be broadly categorized into membrane-bound organelles and membraneless organelles [1, 2]. Membrane-bound organelles, such as the nucleus, mitochondria, endoplasmic reticulum, Golgi apparatus, lysosomes, and vacuoles, are enclosed by lipid membranes, creating specialized environments for various biochemical processes. These membranes serve as critical diffusion barriers, establishing distinct environments that allow for specialized biochemical reactions. For instance, the nucleus houses the cell's genetic material and is the site of transcription, while mitochondria are responsible for producing chemical energy in the form of ATP. Such compartmentalization is essential for maintaining cellular homeostasis and ensuring that metabolic processes are efficiently regulated.

In contrast, membraneless organelles, like nucleoli, stress granules, and P bodies, do not have a surrounding membrane. Membraneless organelles have recently gained attention as an important area of study. They form through liquid-liquid phase separation (LLPS), where proteins and nucleic acids assemble into dynamic, droplet-like structures (Fig 1.1) [3]. This process is driven by multivalent interactions between molecules, often involving proteins

with intrinsically disordered regions (IDRs) and RNA. These interactions concentrate specific biomolecules, triggering a phase separation into liquid-like droplets. These compartments dynamically form and dissolve in response to cellular signals, effectively organizing various cellular processes. Membraneless organelles are crucial for functions such as regulating transcription, processing RNA, and responding to stress.

Initially identified in the P granules of C. elegans embryos, biomolecular condensates have since been found in many cellular contexts across a wide range of organisms, from bacteria to humans [4]. Biomolecular condensates play diverse roles in cellular function; they act as dynamic reaction hubs by concentrating enzymes and substrates to enhance biochemical reactions[5, 6]. For example, transcriptional condensates formed by RNA polymerase II and associated factors promote efficient gene expression by creating local environments rich in transcription machinery[7]. Condensates also play roles in signal transduction, stress response, and RNA processing, among other cellular processes.

The composition and properties of these condensates vary widely, from simple, single-component assemblies to complex, multicomponent structures like the nucleolus and stress granules. Some condensates exhibit multiphasic organization, where distinct liquid phases coexist within a single condensate, enabling separate biochemical environments [8, 9]. A prime example of a multiphasic condensate is the nucleolus[10], which is crucial for ribosome biogenesis. It consists of three distinct phases, each with a specific composition and function. They enable the coexistence of distinct biochemical environments within a single structure and support sequential biochemical reactions within one structure.

Dysfunction of biomolecular condensates can lead to various diseases, such as neurodegenerative disorders[11]. When these condensates transition from a dynamic, liquid-like state to a more solid-like state, they can form harmful protein aggregates that disrupt normal cellular functions. Understanding what drives these transitions is key for developing potential therapeutic strategies.

The study of biomolecular condensates is increasingly important in cell biology as it sheds light on how cells are organized and function. Research in this area helps us understand key biological processes and their roles in health and disease, potentially leading to novel therapies for neurodegenerative disorders and cancers. Additionally, the unique properties of these condensates offer potential uses in synthetic biology and drug delivery.

Current methods for studying biomolecular condensates include advanced microscopy techniques, in vitro reconstitution assays, and computational approaches[12, 13]. Super-resolution microscopy offers high-resolution imaging of condensates in living cells. Techniques like fluorescence recovery after photobleaching (FRAP) and single-particle tracking help study the dynamics and material properties of condensates. In vitro reconstitution assays enable the systematic investigation of condensate formation under controlled conditions. Computational methods, including molecular dynamics simulations and machine learning,

help model and predict condensate behavior. These approaches provide valuable insights into condensate formation and dynamics. However, these methods have limitations. Microscopy techniques may introduce artefacts due to fluorescent tagging, and in vitro studies often fail to capture the full complexity of the cellular environment. Computational models, while useful, may oversimplify biological processes.

Addressing these challenges is crucial for advancing our understanding of biomolecular condensates. Developing new approaches to studying condensates in living cells remains an active area of research. Overcoming current limitations will pave the way for developing novel therapeutic approaches and biotechnological applications based on the principles of phase separation in living systems.

1.1.1 The Structure and Function of Proteins

Proteins are essential for nearly all biological processes [14]. They consist of long chains of amino acids linked by peptide bonds, forming polypeptides. The sequence of these 20 naturally occurring amino acids defines the protein's primary structure, which determines how the protein folds and functions. Protein folding is a hierarchical process that results in secondary, tertiary, and sometimes quaternary structures. The secondary structure includes localized folding patterns like alpha helices and beta sheets, stabilized by hydrogen bonds. The tertiary structure represents the overall three-dimensional structures of a single polypeptide chain influenced by interactions among side chains.

The driving forces behind protein folding include hydrogen bonds, ionic interactions, and hydrophobic effects. These interactions depend on the chemical properties of amino acid side chains, which can be polar, nonpolar, charged, or neutral [15, 16]. This structural diversity allows proteins to perform a wide range of functions, such as catalyzing biochemical reactions, facilitating molecular transport, and providing structural support.

Proteins often contain distinct functional units called protein domains, which can evolve independently and perform specific roles within multi-domain proteins [17, 18]. These domains contribute to protein stability and functions by enabling modular interactions. However, protein stability is influenced by external factors such as temperature, pH, and chemical agents. High temperatures can disrupt hydrogen bonds, leading to denaturation, while changes in pH can alter the charge distribution of amino acids, affecting ionic and hydrogen bonds.

Proteins exist in dynamic states that are crucial for their functionality. The folded state, or native state, represents the well-defined three-dimensional conformation required for specific biological functions. However, some proteins remain unfolded or partially folded under physiological conditions. This flexibility is particularly evident in intrinsically disordered proteins (IDPs), which lack a three-dimensional structure [19]. IDPs can adopt various

conformations depending on their environment or binding partners, enabling them to participate in diverse cellular processes.

Over evolutionary time scales, proteins undergo mutations—changes in their amino acid sequences—that can enhance stability or create beneficial interactions with other molecules[20]. This process has resulted in homologous proteins across organisms that share similar structures but have adapted to perform distinct functions in specific biological contexts[21]. Such adaptability allows organisms to thrive in diverse environments and respond to changing conditions.

In addition to folding and sequence variation, proteins undergo post-translational modifications (PTMs), chemical alterations that occur after synthesis, which play a vital role in regulating protein function [22, 23]. PTMs can significantly alter a protein's activity, stability, and interactions with other molecules. One of the most common PTMs is phosphorylation, where enzymes called kinases add a phosphate group to specific amino acids. This modification introduces a 2 charge at physiological pH, altering protein structure and interactions, which can activate or deactivate proteins depending on the context. Conversely, phosphorylation can be reversed by enzymes called phosphatases through dephosphorylation, restoring the unmodified state of the protein. The balance between phosphorylation and dephosphorylation acts as a molecular switch that regulates cellular signalling pathways. Other PTMs, such as acetylation, ubiquitination, and glycosylation, further expand the functional range of proteins. PTMs are highly dynamic and reversible, allowing cells to adapt rapidly to environmental cues. For example, during stress responses, PTMs regulate metabolic enzymes or signaling molecules to support cellular adaptation. These modifications are vital for fine-tuning protein functions, enabling cells to respond effectively to physiological challenges. PTMs play crucial roles in cell growth and immune responses, maintaining cellular homeostasis.

Understanding protein dynamics, including folding, modifications, and adaptability, is essential for explaining how proteins function in biological systems and how their dysregulation contributes to diseases like cancer and neurodegenerative disorders.

Intrinsically Disordered Proteins

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are unique subsets of proteins or protein segments that lack a stable three-dimensional structure under physiological conditions. Unlike folded proteins, IDPs exist as dynamic conformational ensembles, allowing them to remain flexible and adaptable. This intrinsic disorder enables IDPs to engage in diverse interactions with various molecular partners, making them critical players in cellular processes such as molecular recognition, post-translational modifications (PTMs), scaffolding, and liquid-liquid phase separation [24, 25].

IDPs typically exhibit low sequence complexity and are enriched in polar and charged amino acids, which contribute to their inability to form a stable hydrophobic core. The flexibility of IDPs facilitates various interactions that promote phase separation. These include electrostatic interactions among charged residues, π-π interactions between aromatic amino acids, and cation-π interactions. For instance, positively charged amino acids like lysine and arginine interact with negatively charged ones such as aspartate and glutamate, stabilizing transient complexes and facilitating the formation of condensates. Aromatic amino acids like phenylalanine, tyrosine, and tryptophan participate in π-π interactions, where their electron-rich rings stack against each other, contributing to the stability of protein complexes and enhancing phase separation. Cation-π interactions further promote phase separation by allowing positively charged side chains to interact with the electron cloud of aromatic rings.

The concept of stickers and spacers further illustrates multivalency—the ability of a molecule to engage in multiple simultaneous interactions[26]. Stickers are specific amino acid residues or motifs within IDPs that mediate interactions, such as charged residues or aromatic amino acids involved in electrostatic or π-π interactions. Spacers, on the other hand, are the disordered regions between stickers that regulate their spacing and accessibility.

The sequence composition of IDRs is crucial for determining their phase separation propensity. Factors such as charge distribution, amino acid composition, and the arrangement of hydrophobic residues dictate their behaviour [27, 25]. IDRs rich in polar or charged residues tend to be more soluble, while those with more hydrophobic residues are more likely to promote phase separation. Specific combinations of amino acids can enhance this propensity, as sequences that alternate between hydrophobic and polar residues create favourable conditions for multivalent interactions. The arrangement of hydrophobic residues is also critical; when these residues cluster together, they form hydrophobic cores that drive condensation. Longer and more flexible IDRs allow for greater conformational variability, increasing the likelihood of interactions that lead to phase separation.

Post-translational modifications (PTMs) further regulate the phase behaviour by inducing structural changes that stabilize or destabilize their conformations [23]. Phosphorylation is a particularly important PTM for IDPs, as it alters their electrostatic properties by introducing negative charges. This modification acts as a dynamic switch, enabling cells to regulate condensate formation in response to environmental cues or signalling events.

The ability of IDRs to interact with RNA is another key feature of IDPs. Many RNA-binding proteins (RBPs) contain IDRs that enhance RNA recognition and binding [28]. This flexibility facilitates the assembly of biomolecular condensates essential for processes like transcription and translation.

Within IDRs, short linear motifs (SLiMs) play an important role in mediating protein-protein interactions and serving as sites for PTMs [29]. SLiMs are compact sequences, typically 3–15 residues long, that do not require a stable tertiary structure to function. These motifs

often fold upon binding to their partners, enhancing interaction specificity. SLiMs are highly conserved despite being located in rapidly evolving disordered regions, highlighting their functional importance.

Another notable feature found in many IDPs and RBPs is prion-like domains (PrLDs) [30]. PrLDs are IDRs characterized by low sequence complexity and enrichment in polar amino acids. These domains have a strong propensity for self-assembly and aggregation. PrLDs drive liquid-liquid phase separation but can also form pathological aggregates linked to neurodegenerative diseases. Proteins like TDP-43 and FUS contain PrLDs critical for cellular processes but are prone to toxic aggregation when dysregulated.

The interplay among IDRs, SLiMs, PrLDs, and PTMs is vital for regulating protein function. PTMs can modulate the binding affinity of SLiMs or PrLDs for their interaction partners, thereby influencing signalling pathways or cellular responses. Multiple SLiMs or PrLDs within an IDR enable multivalent interactions with various partners, creating complex regulatory networks essential for cellular function.

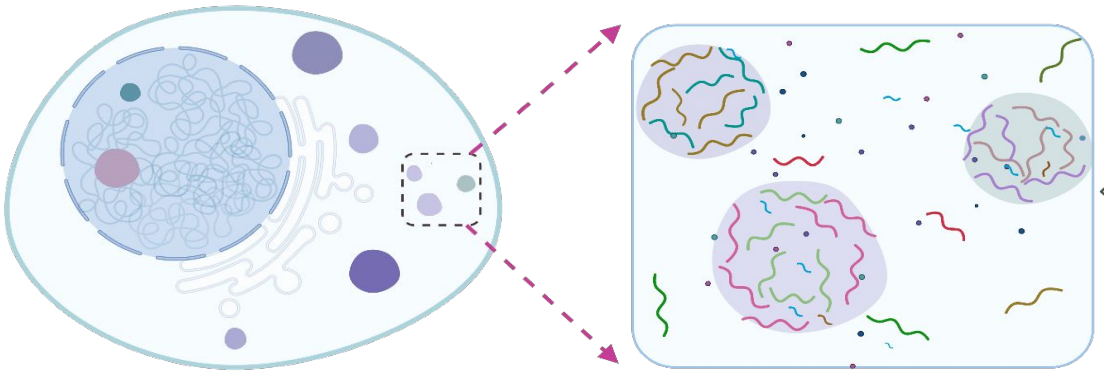

Fig. 1.2: Biomolecular condensates are membraneless structures within cells that form through a process called liquid-liquid phase separation. This phenomenon occurs when certain proteins and nucleic acids, often intrinsically disordered proteins (IDPs), aggregate to create dense, dynamic environments that concentrate biomolecules.

1.1.2 Biomolecular Condensates and Phase Separation

Biomolecular condensates are dynamic, membraneless structures that form through liquid-liquid phase separation (LLPS) within cells[1]. This process occurs when the concentration of specific proteins and nucleic acids exceeds a critical threshold, leading to their aggregation. The formation of these condensates is driven by weak, reversible multivalent interactions, including electrostatic attractions, hydrophobic interactions, and π-π stacking between

aromatic residues. Intrinsically disordered regions (IDRs) within proteins play a crucial role in facilitating these flexible and dynamic interactions. Their internal structure is often heterogeneous, containing different types of biomolecules in varying concentrations, which can enhance functionality through cooperative interactions.

The formation and behaviour of biomolecular condensates are influenced by various factors, including thermodynamics, concentration changes, and environmental conditions. Nucleation initiates the process, followed by growth through coalescence or incorporation of additional biomolecules. Changes in ionic strength, molecular crowding, and temperature can affect the stability and size of condensates. Larger condensates are thermodynamically favoured due to their lower surface tension, often growing at the expense of smaller ones.

Biomolecular condensates serve several key functions in cellular processes. They enhance biochemical reactions by increasing the local concentration of reactants[31]. For example, in the nucleolus, ribosomal RNA synthesis occurs within condensates that help concentrate the necessary components for ribosome assembly[32]. Condensates can also sequester specific proteins or nucleic acids to prevent unwanted interactions, allowing cells to control signalling pathways and metabolic processes more effectively. For instance, P granules in germ cells help regulate the availability of specific mRNAs and proteins essential for development[33]. By creating localized environments for specific reactions, condensates help maintain cellular homeostasis by isolating reactions from the bulk cytoplasm and ensuring optimal conditions for biochemical activities.

The formation and behaviour of biomolecular condensates are influenced by factors such as post-translational modifications, temperature changes, and variations in ion or metabolite concentrations. Dysfunctions in these processes are linked to diseases, including neurodegenerative disorders like Alzheimer's disease, where abnormal protein aggregates can form. Understanding how these structures function can reveal potential therapeutic targets; for example, modifying condensate dynamics may help address issues related to abnormal protein aggregation or disrupted signalling.

Material and Biophysical Properties of Biomolecular Condensates

Biomolecular condensates possess unique material and biophysical properties [1, 34, 35, 36]. These properties include higher viscosity, which allows condensates to resist flow and maintain their shape, and the ability to exhibit gel-like behavior [37, 38]. This viscosity is crucial for creating a localized environment where biochemical reactions can occur more efficiently. Viscosity varies widely among condensates, affecting the diffusion rates of molecules within them. These condensates are not static structures; instead, they can rapidly assemble and disassemble in response to changes in cellular conditions such as temperature, pH, or ionic strength. This flexibility enables cells to adapt to environmental changes or stress. Molecular concentration, composition, and post-translational modifications can significantly

alter condensate properties and dynamics. These condensates can be classified into various material states, including liquid, hydrogel, liquid-crystal, glassy, and amyloid-like fibril states. Each state has distinct material properties that influence biological functions[39]. For example, liquid-like condensates are highly fluid and dynamic, while hydrogels are more stable with limited mobility.

The concentration of components within a condensate is crucial; the saturation concentration (C_{sat}) indicates when phase separation occurs, and the partition coefficient (K) describes the concentration ratio between dense and dilute phases. These parameters help maintain cellular homeostasis by buffering fluctuations in biomolecule concentrations.

The timescale of interactions between proteins affects the material state of condensates; longer-lived interactions promote solid-like behaviour[40]. Surface tension plays a key role in determining the shape and stability of condensates by minimizing surface area, which drives them to adopt spherical shapes. Some biomolecular condensates consist of multiple coexisting liquid phases. Interfacial tension controls the arrangement and stability of these phases, often causing them to form layered structures to minimize energy. Distinct phases within a condensate enable selective biomolecule partitioning, enhancing biochemical reaction efficiency.

Understanding these properties elucidates the roles of biomolecular condensates in cellular processes and their dysregulation in diseases. Their ability to concentrate specific molecules while excluding others, combined with their dynamic nature, makes them effective reaction hubs for regulating biochemical processes [41].

Biomedical Significance

Biomolecular condensates are critical in disease pathogenesis due to dysregulation in their formation, composition, and behaviour [42, 43, 11]. A key process in many condensate-related diseases is the transition from a functional liquid-like state to a pathological solid-like state, often leading to the formation of protein aggregates characteristic of numerous neurodegenerative disorders. This phase transition can be triggered by factors such as the ageing of condensates, post-translational modifications, mutations in condensate-forming proteins, concentration changes of components, and environmental factors like pH or temperature shifts[44]. In conditions like amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), proteins such as TDP-43 and FUS can irreversibly transition from their normal droplet state into an amyloid state, resulting in loss of function and toxic intermediates[45, 46, 47]. Dysregulated condensates disrupt essential cellular processes[48]. In cancer, abnormal condensate formation by transcription factors can interfere with gene expression. These disruptions result in the loss of normal cellular functions, toxic effects due to sequestration of essential components or harmful aggregates, altered cellular organization, and the nucleation of pathological aggregates.

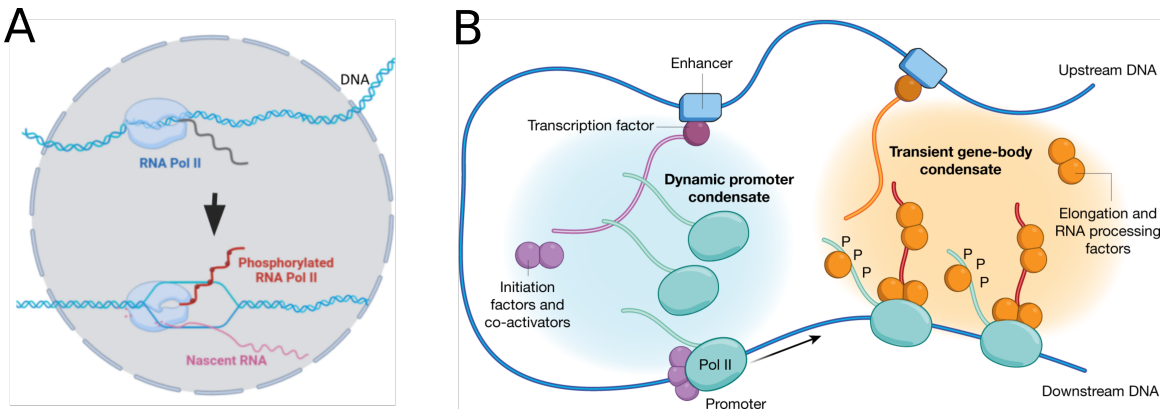

Fig. 1.3: A) The transcription process by Pol II involves initiation, where Pol II binds to DNA, and elongation is driven by CTD phosphorylation for RNA synthesis. B) Hypothetical model of an active Pol II gene (blue) with promoter (turquoise) and gene-body (orange) condensates. The promoter condensate, formed by transcription factors and enhancers (blue boxes), recruits Pol II with an unphosphorylated CTD (turquoise ovals), co-activators, and initiation factors (violet). The gene-body condensate includes nascent RNA (red), elongation and RNA processing factors (orange), and phosphorylated Pol II. Pol II shuttles between condensates depending on CTD phosphorylation (adapted from [49]).

1.1.3 Transcriptional condensates

Gene expression is the fundamental process by which the information encoded in DNA is converted into functional products, primarily proteins and RNAs, that perform cellular functions[50, 51, 52]. In eukaryotic cells, transcription initiates gene expression by copying DNA into RNA. RNA polymerase II (Pol II) is the enzyme responsible for transcribing protein-coding genes and some non-coding RNAs. Pol II is a complex of 12 subunits (RPB1-12), with its largest subunit containing a conserved C-terminal domain (CTD) made up of heptad repeats (YSPTSPS). The number of repeats varies by species, with 26 in yeast and up to 52 in vertebrates. The CTD serves as a platform for recruiting transcription and RNA processing factors.

Pol II requires additional proteins for transcription steps and is regulated by factors such as chromatin structure (the complex of DNA and proteins that packages genetic material), CTD phosphorylation (the addition of phosphate groups to the CTD), and elongation factors like TFIIS (a protein that helps Pol II overcome pausing and backtracking). Transcription is broadly divided into initiation and elongation phases (Fig 1.3A). During initiation, Pol II and general transcription factors assemble at the promotor region of the gene to form the pre-initiation complex (PIC), leading to DNA unwinding and the commencement of RNA synthesis(Fig 1.3B). In elongation, Pol II synthesizes RNA along the DNA template, a process regulated by P-TEFb kinase (positive transcription elongation factor b), which phosphorylates the CTD.

The phosphorylation status is crucial for its function throughout the transcription process. At the beginning of transcription, Ser5 residues are phosphorylated by CDK7 (cyclin-dependent

kinase 7), while Ser2 residues are increasingly phosphorylated by CDK9 (P-TEFb) during elongation. This results in a phosphorylation pattern where Ser5 phosphorylation peaks near the transcription start site, and Ser2 phosphorylation accumulates toward the end of the gene. Finally, termination occurs when Pol II recognises specific signals in the DNA that prompt it to stop RNA synthesis and release the completed transcript.

Recent structural studies have revealed that transcription involves the formation of distinct transient condensates at gene promoters and within gene bodies (regions of a gene that are transcribed into RNA) (Fig 2.1B, C)[53, 54, 32]. These condensates concentrate the necessary factors for transcription initiation and elongation, allowing RNA polymerase II (Pol II) to shuttle between them in a phosphorylation-dependent manner. This organization is crucial for regulating gene activity and ensuring proper cellular responses. Different types of transcriptional condensates control various steps of gene activation, including forming pre-initiation complexes and facilitating elongation. The phosphorylation of the CTD of Pol II is key to transitioning from initiation to elongation, enabling Pol II to move from initiation to elongation condensates. Newly synthesized RNA influences condensate formation and dissolution, creating feedback mechanisms that regulate the transcription process.

Our understanding of transcriptional condensates is still limited, with several open questions[55]. While many studies support phase separation, some researchers believe that weak, transient interactions might be enough to explain the clustering of transcription factors and other components without needing a separate phase. How different condensates coexist and selectively sort biomolecules during transcription remains unclear. These condensates are difficult to study because of their small size and complex interactions, which makes traditional methods less effective in living cells [55, 56].

This thesis addresses these challenges by exploring how transcriptional condensates regulate gene transcription using multiscale molecular dynamics simulations. It comprises four key projects: First, we explore the intrinsically disordered nature of the CTD using NMR spectroscopy and molecular dynamics simulations. Second, we examine how CTD phosphorylation drives Pol II's phase behaviour, forming multiphasic condensates to separate initiation and elongation phases. In the third project, we analyze the stochastic effects of CTD phosphorylation in nonequilibrium states to understand complex cellular behaviours in dynamic environments. Finally, we apply active learning techniques to efficiently map protein sequence parameters, enhancing predictions about biomolecular condensate formation.

The objective is to uncover the molecular mechanisms that allow cells to maintain functional organization under dynamic conditions, bridging knowledge gaps in transcriptional condensate understanding and offering a perspective on cellular gene regulation.

1.1.4 Experimental Techniques for Investigating Condensates and Intrinsically Disordered Proteins

Experimental techniques for studying condensates and intrinsically disordered proteins (IDPs) include a variety of methods[57]. Microscopy techniques are crucial for visualizing condensates. Fluorescence microscopy enables imaging with nanometer resolution. Time-lapse microscopy tracks phase separation in real time under various conditions. Fluorescence recovery after photobleaching (FRAP) assesses molecular diffusion within condensates, helping distinguish between liquid-like and gel-like states. Single-molecule fluorescence techniques, such as Förster resonance energy transfer (FRET), provide insights into protein-RNA interactions driving condensate formation. Mass spectrometry identifies protein components within condensates, though it lacks spatial information.

For IDPs, Nuclear Magnetic Resonance (NMR) spectroscopy is particularly useful, offering atomic-level insights into their dynamic conformations and structural preferences. NMR reveals transient structural features. Small-angle X-ray scattering (SAXS) further characterizes the overall shape and size distributions of IDPs in solution.

Nuclear Magnetic Resonance (NMR) Spectroscopy

For the study of IDPs, Nuclear Magnetic Resonance (NMR) spectroscopy emerges as a particularly powerful tool, offering atomic-level insights into their dynamic conformations and structural preferences[58]. NMR exploits the magnetic properties of certain atomic nuclei, such as hydrogen and carbon, to study molecular structure and dynamics. When placed in a strong magnetic field and exposed to radio waves, these nuclei emit signals that provide detailed information about their chemical environment, interactions with neighboring atoms, and the overall molecular structure. The resulting NMR spectrum serves as a unique fingerprint of the molecule, with peaks at different frequencies corresponding to different atoms.

NMR provides extensive information on molecular structure and dynamics, making it invaluable for studying IDPs. Chemical shifts indicate the electronic environment of nuclei, while spin-spin coupling, observed as peak splitting, reveals magnetic interactions between neighbouring nuclei. Relaxation time measurements provide data on molecular motion and dynamics, making NMR well-suited for investigating the highly flexible nature of IDPs. Unlike structured proteins, IDPs exhibit distinctive spectral characteristics in NMR, with signals clustered in a narrow region of the spectrum, indicating high flexibility and lack of a defined structure.

NMR is particularly powerful for studying intrinsically disordered proteins (IDPs), which lack a stable structure. Unlike structured proteins, IDPs are highly flexible, making them challenging to investigate with conventional methods. However, NMR captures

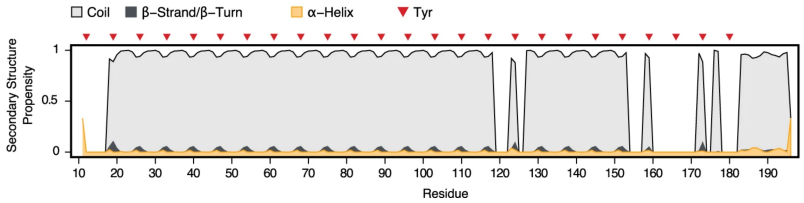

Fig. 1.4: Secondary structure propensity of yCTD based on NMR chemical shifts, with Tyr residues indicated by red triangles. The analysis confirms a dynamic, predominantly random coil behaviour, with a slight propensity for turn formation at the Y1 heptad position.

their dynamic nature at the atomic level. For IDPs, NMR shows distinctive spectral characteristics; the protein's signals cluster in a narrow region of the spectrum, indicating high flexibility and lack of a defined structure, contrasting with structured proteins that exhibit more spread-out and distinct signals. NMR can also reveal local structural preferences.

NMR Spectroscopy Techniques for Studying Intrinsically Disordered Proteins

Intrinsically disordered proteins (IDPs) exist as dynamic ensembles of conformations, making them difficult to study using traditional structural biology techniques. Several NMR-based methods have proven effective for investigating their structural and dynamic properties[59, 60, 61]. Secondary chemical shifts, calculated by comparing observed shifts to random coil values, reveal transient structural preferences such as α-helices and β-sheets (Fig 1.4). The secondary chemical shifts of Cα, Cβ, N, and HN nuclei provide insights into these preferences. Residual Dipolar Coupling (RDC) provides information on the alignment of molecules in a magnetic field by analyzing the angles between internuclear vectors and magnetic field vectors. It captures the dynamic nature of IDPs and detects transient structural preferences, aiding in understanding their flexibility. The Nuclear Overhauser Effect (NOE) provides distance information between protons in NMR spectroscopy, effective up to about 5 Å. In IDPs, sequential and medium-range NOEs are common, while long-range NOEs are rare in NOESY spectra. The absence of long-range NOEs in IDPs does not rule out short-lived contacts, as their rapid dynamics can weaken NOE signals and complicate peak assignments. However, NOE data is still useful for identifying short-range interactions within IDPs. The hydrodynamic radius (Rh) measures the effective size of IDPs in solution, reflecting their flexibility. Rh is typically larger for IDPs than for structured proteins, indicating their dynamic nature. Techniques like pulsed-field gradient NMR (PFG-NMR) measure diffusion coefficients to calculate Rh.

The dynamic nature of IDPs results in averaged chemical shifts that do not capture the full range of conformational states, leading to underdetermined datasets that are difficult to interpret. The study of IDPs relies on conformational ensembles derived from techniques such as NMR, SAXS, single-molecule FRET, and molecular dynamics (MD) simulations. Recent advancements in MD simulations provide insights into IDP dynamics and interactions. A combination of these methods is essential for understanding the complex behaviour and

functions of IDPs.

Caenorhabditis elegans(*C. elegans*)

The nematode Caenorhabditis elegans (C. elegans) is a widely used model organism in biological research due to its simplicity, genetic tractability, and well-characterized biology[62, 63]. Its short life cycle (3 days), transparent body, and ease of maintenance enable high-throughput experiments and direct observation of cellular processes in living tissues[63]. These features make it particularly valuable for studying liquid-liquid phase separation (LLPS), where its genetic tools allow precise manipulation of proteins involved in biomolecular condensates like P granules [64].

Beyond LLPS, C. elegans serves as a key system for toxicology, drug discovery, and aging research, leveraging its conserved molecular pathways (60–80% genetic similarity to humans)[65]. Its ability to model complex biological phenomena in an ethical, cost-effective, and physiologically relevant context provides advantages over in vitro systems and mammalian models[66].

1.2 Theory

1.2.1 Thermodynamics

Here, we examined the crucial roles of intrinsically disordered proteins (IDPs) and regions (IDRs) in cellular function, particularly in the context of biomolecular condensates and phase separation. To fully understand the behaviour of IDPs and IDRs in biomolecular condensates, we must now turn our attention to the underlying physical principles that govern these phenomena. The following sections will introduce the key concepts from polymer physics and thermodynamics that are fundamental to describing the behaviour of disordered proteins in solution and their propensity for phase separation. We will begin by examining the Flory-Huggins theory, a fundamental model in polymer physics that provides a framework for understanding the mixing behaviour of polymers in solution. This theory provides a basis for exploring how IDPs drive phase separation and contribute to condensate formation.

Intrinsically disordered proteins (IDPs) and polymers share fundamental characteristics as macromolecules composed of hundreds of thousands of atoms. Both consist of smaller repeating units — monomers in polymers and amino acids in IDPs — arranged in different sequences. This repetitive structure, while more complex in IDPs, allows for simplified theoretical modelling despite their large size. The conformational diversity of IDPs and polymers stems from their chain-like structure, which permits significant flexibility. This flexibility enables a diverse range of conformations and is essential for their function. Importantly, local interactions among molecules tend to average out due to multiple interactions with surrounding molecules, allowing for a simplified description of their behaviour. As a result, both polymers and IDPs can often be characterized using a few effective quantities rather than detailed atomic-level descriptions. Key parameters include topology, local stiffness along the backbone, bulkiness, and the compatibility or incompatibility of the building blocks. For IDPs, these parameters are influenced by their specific amino acid sequences and can include partially structured regions, adding complexity to their behaviour[67].

The behaviour of both IDPs and polymers is often governed by bulk properties rather than by specific local structures, as they are influenced by their environment and interactions with other molecules. Because of these similarities, concepts from polymer physics, such as the Flory-Huggins theory, can be applied to IDPs[68, 69]. This is particularly useful for understanding processes like phase separation and the formation of biomolecular condensates in cells. By combining ideas from polymer physics with the unique aspects of proteins, we can better understand the complex behaviours of IDPs in cellular organization and function, helping us explain important cellular processes.

Liquid-liquid phase separation (LLPS) involves a homogeneous mixture of bio-macromolecules separating into two distinct liquid phases. This can be illustrated by everyday examples such as oil and water separating into layers due to their immiscibility. In biological contexts,

consider a water-soluble polymer in an aqueous solution: under specific conditions, it can separate into a concentrated polymer-rich phase and a dilute polymer-poor phase. This behaviour is crucial for understanding cellular function as biomolecular condensates emerge from similar processes, enabling cells to compartmentalize biochemical reactions and adapt to environmental changes without relying on membranes.

In two-component systems, phase separation can be mathematically described using volume fractions. For a polymer-solvent mixture, ϕ_p represents the polymer fraction and ϕ_s the solvent fraction, with the relationship $\phi_p + \phi_s = 1$ in a closed system. In a binary mixture of a polymer and a poor solvent, there exists a specific concentration threshold known as the saturation concentration (ϕ_{sat}). Beyond this threshold, the system separates into a dense polymer-rich phase and a dilute phase. The volume fractions of the polymer in these coexisting phases are referred to as ϕ_{dilute} and ϕ_{dense}, where ϕ_{dilute} is equal to ϕ_{sat}[70, 71].

The Gibbs phase rule, expressed as $F = C - P + 2$, is a fundamental principle in thermodynamics that relates the number of degrees of freedom (F) to the number of components (C) and phases (P) in a system at equilibrium[72]. The degree of freedom refers to the number of intensive variables (such as temperature, pressure, or composition) that can be independently varied without disturbing the equilibrium state of the system. This rule helps in understanding the conditions under which phase transitions occur and for predicting the number of coexisting phases under given conditions.

Flory-Huggins theory

The Flory-Huggins theory is a fundamental model in polymer science that describes the thermodynamics of polymer solutions and blends [73, 74]. It provides a framework for understanding phase behaviour, including mixing and demixing phenomena in these systems. The theory is based on a lattice model, where polymer chains and solvent molecules are assumed to occupy discrete sites on a regular lattice. Each lattice site is occupied by either a solvent molecule or a polymer segment.

The theory is based on several key assumptions that simplify its calculations. It assumes incompressibility, where the total volume of the solution equals the sum of the volumes of the polymer and solvent components. Polymers chains are considered flexible, allowing them to adopt any conformation within the lattice. Additionally, the mean-field approximation is applied, where each molecule experiences an average force field, ignoring the local fluctuations in concentration and energy.

The Flory-Huggins theory focuses on the Gibbs free energy of mixing (ΔG_{mix}), which is determined by both the entropy of mixing (ΔS_{mix}) and the enthalpy of mixing (Δh_{mix}). The main source of entropy gain during mixing comes from translational entropy—the increased ability of polymer chains and solvent molecules to move freely and occupy more spatial

positions. While polymer chains have many possible conformations that contribute to their configurational entropy, this contribution changes little upon mixing. The translational entropy per molecule can be expressed as:

$$T\Delta S_{mix} = -k_B T \left(\frac{\phi_1}{N1} \ln \phi_1 + \frac{\phi_2}{N2} \ln \phi_2 \right)$$

Here, k_B is Boltzmann's constant, and T is the absolute temperature. ϕ_1 and ϕ_2 represent the volume fractions of solvent and polymer in the solution, respectively. Their sum is always equal to $\phi_1 + \phi_2 = 1$. N_1 and N_2 are the lattice sites occupied by a single polymer molecule and a single solvent molecule, respectively. For a polymer solution $N_1 = N$ and $N_2 = 1$.

The enthalpy of mixing represents the change in interaction energy during mixing and depends on the Flory-Huggins interaction parameter χ, which characterises the polymer-solvent interactions in a mean-field approximation. The interaction parameter is defined as:

$$\chi = \frac{z}{2k_B T}(2u_{ps} - u_{pp} - u_{ss}),$$

where z is the coordination number (number of nearest neighbors on the lattice), and u_{ps}, u_{pp}, and u_{ss} are pairwise interaction energies for polymer-solvent, polymer-polymer, and solvent-solvent pairs, respectively. Using this definition, the enthalpy per lattice site can be expressed as:

$$\Delta H_{\text{mix}} = \chi\phi(1 - \phi)$$

Combining entropy and enthalpy contributions, the Gibbs free energy of mixing becomes:

$$\Delta G_{\text{mix}} = \Delta H_{\text{mix}} - T\Delta S_{\text{mix}},$$

which can be written as [75]:

$$\frac{\Delta G}{k_B T} = \frac{\phi}{N} \ln \phi + (1 - \phi) \ln(1 - \phi) + \chi\phi(1 - \phi) \tag{1.1}$$

This equation is called the Flory-Huggins equation for polymer solutions, which includes contributions from both entropy and interaction energy. The entropy terms favor mixing by increasing system disorder, while the interaction energy term, determined by the Flory-Huggins interaction parameter χ, can oppose mixing if χ is positive. At low χ values, entropy dominates, resulting in miscibility. However, as χ increases, the interaction energy becomes significant and may lead to phase separation if it outweighs the entropic contributions.

A higher χ value reflects greater energy costs for polymer-solvent interactions. When $\chi > 0$, the polymer is in a "poor solvent," encouraging phase separation. If χ exceeds a critical threshold (which depends on system-specific factors), phase separation occurs, forming two distinct phases: one with low polymer concentration and another with high polymer concentration. This behavior is observed in solutions containing intrinsically disordered

proteins (IDPs), which undergo liquid-liquid phase separation due to weak multivalent interactions modeled by Flory-Huggins theory. Conversely, a negative χ value indicates favorable polymer-solvent interactions, promoting mixing.

The Flory-Huggins interaction parameter χ is influenced by temperature, which is important for understanding polymer solutions. It can be expressed as:

$$\chi(T) = A + \frac{B}{T}$$

where A represents a constant entropic contribution (temperature-independent), and $\frac{B}{T}$ represents the enthalpic contribution. At low temperatures, the enthalpic term ($\frac{B}{T}$) has a greater effect, while at high temperatures, the entropic term (A) becomes more important. This analysis is not limited to temperature changes; other conditions can also affect the interaction parameter. The temperature dependence of χ results in two behaviours: the Lower Critical Solution Temperature (LCST), where χ increases with temperature, and the Upper Critical Solution Temperature (UCST), where χ decreases. Some systems can exhibit both behaviours. The Flory-Huggins theory predicts phase behaviour in polymer-solvent or polymer-polymer mixtures based on the temperature dependence of the χ parameter (Fig. 1.5). Phase behaviour is determined by the Gibbs free energy (ΔG) of the mixture. If ΔG for the blend is lower than that of the separate components, the mixture is fully miscible. Conversely, if ΔG is higher than that of two coexisting phases, phase separation occurs. At high temperatures, the Gibbs free energy favours a stable homogeneous phase, while at lower temperatures, a maximum in the Gibbs energy versus composition curve can lead to phase separation.

At equilibrium, a key condition is that the chemical potential (μ) of each component must be equal across all phases. This is expressed mathematically as

$$\mu_{1,\alpha}(T,p) = \mu_{1,\beta}(T,p) \quad \text{and} \quad \mu_{2,\alpha}(T,p) = \mu_{2,\beta}(T,p),$$

where subscripts 1 and 2 refer to different components, and α and β represent different phases. The chemical potential is a measure of how the Gibbs free energy of a system changes with the addition or removal of a component. This equilibrium condition is fundamental to understanding phase behaviour and can be derived from the principle of minimizing the total Gibbs free energy of the system. The chemical potential is defined as the partial molar Gibbs free energy:

$$\mu_i = \left(\frac{\partial G}{\partial n_i}\right)_{T,p,n_{j\neq i}},$$

where G is the Gibbs free energy and n_i is the number of moles of component i.

The T1 curve in Fig. 1.5 illustrates two minima in Gibbs free energy (G), indicating unstable compositions between these points where phase separation occurs. The compositions of the two coexisting phases (ϕ' and ϕ'') are determined by constructing a double tangent

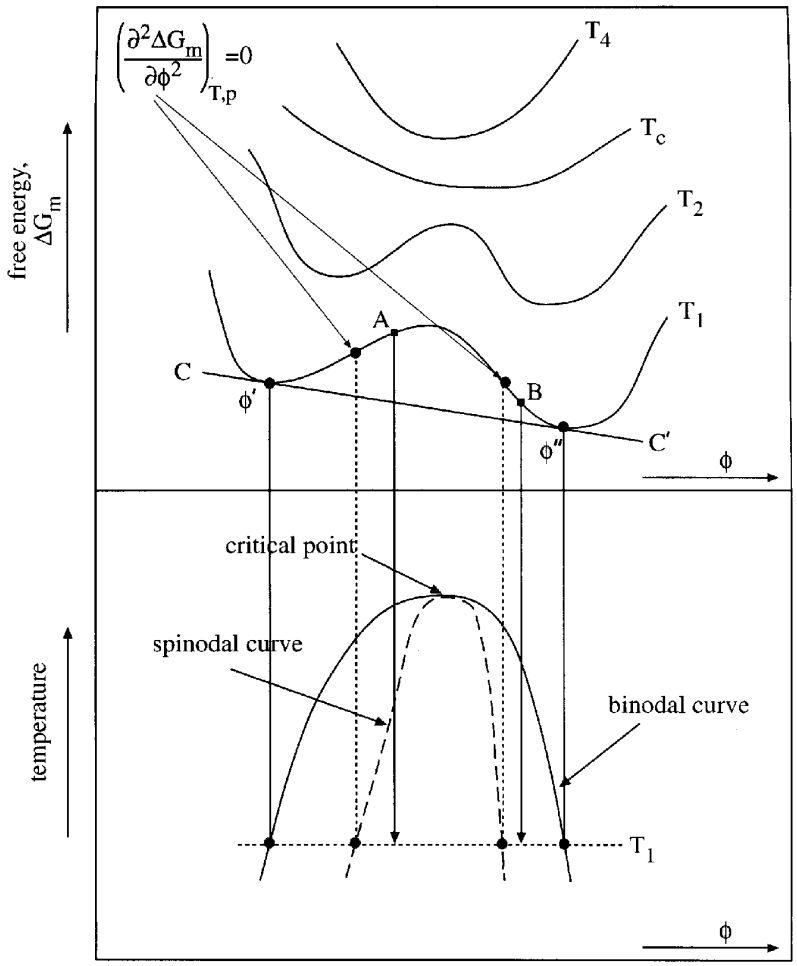

Fig. 1.5: Variation of Gibbs free energy with composition, ϕ ($\phi = \phi_1$ or ϕ_2) at four temperatures. The tie line CC′ defines the compositions on the binodal curve. The locus of points defined by the points of inflection $\left(\frac{\partial^2 G}{\partial \phi^2}\right)_{T,p} = 0$ defines the spinodal curve. At point A (inside the spinodal curve), the mixture will spontaneously phase separate into domains with compositions ϕ' and ϕ'' via spinodal decomposition. However, at point B (outside the spinodal curve), there is an energy barrier to phase separation, which then occurs by nucleation and growth. Adpated from [76]

line (CC') that touches the G curve at these minima. These points define the binodal curve, which narrows to a single point at the critical solution temperature (T_c). The binodal curve represents compositions where two distinct phases coexist in equilibrium. When the mixture's composition lies within the binodal, it is thermodynamically unstable or metastable, meaning that the mixture can spontaneously separate into different phases.

Inside the spinodal region, phase separation occurs spontaneously via spinodal decomposition, initiating coarsening processes. In the region between the spinodal and binodal curves, the phase is metastable. Here, nucleation and growth take place, with small clusters of the minority phase growing and merging over time, minimizing interfacial energy. As this process progresses, Ostwald ripening occurs, where smaller droplets dissolve and redeposit their

material onto larger droplets, resulting in an overall increase in droplet size. The Critical Solution Temperature marks the point where the spinodal and binodal curves converge. The condition

$$\frac{\partial^2 G}{\partial \phi^2} = 0$$

marks the boundary between metastable and unstable regions in a phase diagram.

The Upper Critical Solution Temperature (UCST) occurs at the maximum of the phase diagram, where solubility increases with temperature due to stronger interactions between molecules and solvent, common in non-polar polymer-solvent mixtures. In contrast, in the Lower Critical Solution Temperature (LCST), solubility decreases with temperature as thermal motion disrupts these interactions.

Thermoresponsive phase separation can occur either above or below a critical temperature, depending on the specific balance between enthalpic interactions and entropic contributions in the system[77, 78]. These behaviors are classified as lower critical solution temperature (LCST) and upper critical solution temperature (UCST) transitions, respectively. In some systems, both types of transitions can be observed under different conditions, allowing for tunable phase behavior. UCST-type separation is typically governed by the interplay of entropy and enthalpy of mixing, while LCST behavior is often driven by an entropy gain, such as the release of bound molecules like water.

1.2.2 Understanding Polymer behaviour

Polymers are large molecules characterized by high molecular weight, consisting of repeated subunits known as monomers. These monomers, which are relatively small and simple molecules, are covalently bonded through polymerization. The resulting polymer can exhibit various forms and structures depending on the arrangement and type of monomers. Polymers can be broadly categorized into synthetic polymers, such as plastics, fibres, and glues, and biopolymers, which include natural substances like proteins, polysaccharides, and DNA. The structural diversity of polymers is further enhanced by factors such as branching architectures, stereoisomerism (the spatial arrangement of atoms), and different types of monomer sequences ranging from homopolymers (identical monomers) to alternating and random sequences. Understanding these fundamental aspects of polymers is crucial for exploring their physical properties and applications in various fields.

Ideal Chains

Understanding polymer behaviour can be simplified using models like ideal chains. It assumes that monomers far apart along the chain exhibit no long-range correlation, though short-range correlations between neighboring monomers may still exist. In this model, the

polymer consists of connected monomers that do not interact with each other except when they are adjacent. This allows the chain to behave like a random walk, where each segment can point in any direction without affecting the others. An isolated homopolymer consists of $N+1$ monomers located at positions $r_0, r_1, \ldots, r_n$ in space. It simplifies the representation of polymer chains by treating them as a series of connected monomers or links, each with a fixed length b and no restrictions on the angles between them. A specific case of the ideal chain is the freely jointed chain, which assumes complete independence between bond vectors. In this model, there is no correlation between the orientations of consecutive bonds[79]:

$$\langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle = 0 \quad \text{for } i \neq j,$$

and all bond vectors have the same fixed length b. The total end-to-end vector of the chain is represented as [80]:

$$\mathbf{R} = \sum_{i=1}^{n} \mathbf{b}_i,$$

where n is the number of monomers. The mean square end-to-end distance provides a measure of the chain's size and scales linearly with the number of bonds:

$$\langle R^2 \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle = nb^2.$$

This key result demonstrates that the size of the chain grows with the square root of the number of bonds:

$$R \propto n^{1/2}.$$

This scaling behaviour is a fundamental characteristic of random walk dynamics and offers essential insights into the statistical mechanics governing polymer conformations.

Chain Conformations

The conformation of polymer chains is a key factor that affects their physical properties and behaviours[81, 82, 83]. Polymers are long molecules made up of repeating units called monomers. They are flexible and can take on different shapes. The main conformations include extended, coiled, and folded structures. The specific shape a polymer chain adopts depends on several factors, including steric hindrance, torsional angles, and intermolecular interactions. Understanding these conformations is essential for predicting the behaviour of polymers in various applications. Steric hindrance refers to the arrangement of atoms in a polymer chain that can limit the rotation around bonds, affecting its shape; larger side groups can increase steric hindrance, leading to more compact conformations. Torsional angles are the angles between adjacent bonds in the polymer backbone that influence the overall shape; each bond can rotate freely within certain limits, resulting in various possible shapes. Intermolecular interactions, such as hydrogen bonding or van der Waals forces, can

stabilize certain shapes over others.

Random Coil Model

In dilute solutions, polymer chains often exist in a random coil conformation[84, 85]. The random coil behavior is well-described by the Gaussian chain model, which assumes that the polymer chain behaves as a random walk in three-dimensional space. This model assumes monomers are connected by flexible bonds, allowing free rotation. The dimensions of a random coil can be characterized by the radius of gyration R_g, which is defined as[82]:

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N} r_i^2},$$

where r_i is the distance of the i-th monomer from the center of mass of the chain, and N is the total number of monomers. The radius of gyration provides insight into the size and compactness of the polymer coil.

Persistence Length

The flexibility of a polymer is characterized by its persistence length (l_p), which measures the distance over which the directional correlation between segments of the chain persists.

The directional correlation between two segments of a polymer chain decreases exponentially with the distance separating them along the contour length. This relationship is described by the equation:

$$\langle \cos \Theta(s) \rangle = \exp(-sb/l_p),$$

where s is the length separating the segments, b is the bond length, and l_p is the persistence length. The persistence length serves as a measure of polymer stiffness. For flexible chains, l_p is comparable to the bond length (b), while for semiflexible chains, such as those described by the worm-like chain model, $l_p > b$.

The persistence length is closely related to the Kuhn length (l_k), which represents an equivalent freely jointed chain. The relationship between these parameters is given by:

$$l_p = \frac{1}{2} l_k = \frac{2b}{\langle \Theta^2 \rangle},$$

where $\langle \Theta^2 \rangle$ is the mean square angle between adjacent bonds.

Influence of Solvent Quality and Scaling Laws in Polymer Physics

The behaviour of polymers in solution is fundamentally governed by the quality of the solvent in which they are dissolved[86, 87, 88, 85]. The properties of polymers in solution are determined by the balance between effective monomer-monomer attractions and excluded volume repulsions. Solvent quality can be categorized into three regimes: good, theta, and poor solvents.

In good solvents, polymer-solvent interactions are energetically favorable compared to monomer-monomer interactions. This causes polymer chains to swell and adopt more extended conformations. The dominance of repulsive excluded volume effects leads to chain expansion. The radius of gyration (R_g) of a polymer in a good solvent scales with the number of monomers (N) according to:

$$R_g \sim N^{0.588},$$

where 0.588 is known as the Flory exponent. As temperature decreases and approaches the theta temperature (T_θ), polymer chains behave like ideal random walks, leading to:

$$R_g \sim N^{1/2}.$$

At this temperature, attractive and repulsive forces between polymer segments balance out, leading to ideal chain behavior where the second virial coefficient (B_{22}) equals zero. The theta condition helps determine unperturbed chain dimensions.

In poor solvents, polymer-polymer interactions are favoured over polymer-solvent interactions. This causes the polymer chains to collapse into dense globules to minimize contact with the solvent. The scaling relationship in poor solvents becomes:

$$R_g \sim N^{1/3}$$

These solvent conditions significantly influence various polymer properties. For instance, the second virial coefficient (B_{22}), which describes the strength of pairwise interactions between polymer molecules in solution, is positive in good solvents, approaches zero at the theta point, and becomes negative in poor solvents. This relationship can be expressed mathematically as:

$$B_{22} = \frac{1}{2M^2} \int_0^\infty \left[1 - e^{-U(r)/k_B T}\right] 4\pi r^2 dr,$$

where M is the molecular weight and $U(r)$ is the interaction potential. The behaviour of polymer solutions is characterized by three key temperatures that provide crucial insights into polymer-solvent interactions and phase behaviour: the theta temperature (T_θ), the Boyle temperature (T_B), and the critical temperature (T_c). At T_θ, polymers behave as ideal chains, meaning that the effects of polymer-polymer interactions and polymer-solvent interactions exactly balance each other. As the polymer length (L) increases, both the Boyle temperature

($T_B(L)$)—the temperature at which higher-order virial coefficients vanish—and the critical temperature ($T_c(L)$) approach the theta temperature (T_θ). For an infinitely long chain, these temperatures converge: $T_B(L) > T_\theta > T_c(L)$, and as $L \to \infty$, $T_B = T_c = T_\theta$. Understanding these solvent conditions influences various properties of polymers and helps predict their behaviour across different environments.

1.2.3 Surface tension

Surface tension is a fundamental physical phenomenon that arises at the interface between different phases, such as liquid-gas or liquid-liquid boundaries. This effect is particularly pronounced at small scales, where the surface-to-volume ratio is large, making interfacial properties critically important in various applications, including fluid dynamics, material science, and biological systems. By applying principles from thermodynamics and statistical mechanics, we can gain insights into how interfaces influence the stability and behaviour of multiphase systems.

Surface energy is defined as the excess free energy per unit area at an interface, measured in Joules per square meter (J/m^2) or milliJoules per square meter (mJ/m^2), while surface tension is defined as a tensile force per unit length acting along the surface, measured in Newtons per meter (N/m) or milliNewtons per meter (mN/m). The microscopic origin of surface energies arises because molecules at the surface experience different interactions compared to those in the bulk due to a lack of neighbouring molecules above them. This results in higher potential energy for surface molecules compared to bulk molecules, leading to excess energy at the interface. The cohesive energy of a liquid is determined by the sum of interaction energies with neighboring molecules within the range of molecular interaction potentials.

Molecular interactions include various types such as Coulombic, dipole-dipole, and van der Waals forces, with interaction potentials showing both attractive and repulsive components.

Molecules in a liquid are in constant thermal motion, with neighbours changing rapidly. The average interaction energy reflects many molecular collisions. Surface tensions represent excess free energies rather than pure binding energies and include random thermal motion and entropy effects. Capillary waves are fluctuations at the liquid surface that contribute to the system's entropy, which become more pronounced at higher temperatures, reducing the overall surface energy.

Surface tension is defined thermodynamically as the mechanical work (δW) required to increase the surface area (dA) of a liquid[89]:

$$dW = \gamma dA$$

where γ is the surface tension. This definition aligns with the microscopic view that surface molecules have excess energy compared to those in the bulk. When increasing surface area while keeping the number of molecules constant, work must be done against molecular interaction forces, and this work is stored as additional surface energy in the total Gibbs free energy (G) of the system. Thus, we can express surface tension as:

$$\gamma = \left(\frac{\partial G}{\partial A}\right)_{N,P,T}$$

The presence of finite macroscopic surface tension explains why droplets assume a spherical shape in the absence of external forces. Spheres minimize surface area for a given volume, making them the equilibrium shape of free drops.

Young-Laplace Equation

The Young-Laplace equation describes the pressure jump across curved liquid surfaces and liquid-liquid interfaces due to surface tension. For spherical surfaces, where curvature is uniform across all directions, the pressure difference due to surface tension simplifies to $\Delta P_L = \frac{2\gamma}{R}$, where R is the radius of curvature of the drop. This excess pressure is known as Laplace pressure, and the mean curvature (κ) is defined as $\kappa = \frac{1}{R}$. For nonspherical shapes, the generalized Young-Laplace equation accounts for two principal radii of curvature (R_1 and R_2)[89]:

$$\Delta P_L = \gamma \left(\frac{1}{R_1} + \frac{1}{R_2}\right).$$

In mechanical equilibrium, a liquid interface maintains a constant Laplace pressure unless external forces act on it. Any variation in curvature creates pressure gradients along the interface, driving fluid flow until equilibrium is restored. For spherical surfaces, where $R_1 = R_2 = R$. For flat surfaces, as $R_1, R_2 \to \infty$, the pressure difference becomes zero ($\Delta P_L = 0$).

Wetting Principles

Wetting describes the interaction between a liquid and a solid surface, specifically the liquid's ability to spread and maintain contact with the surface. The degree of wetting is quantified by the contact angle (θ), defined as the angle at the three-phase contact line where solid, liquid, and vapor meet. A lower contact angle (less than 90°) indicates good wettability, while a higher contact angle (greater than 90°) suggests poor wettability. Young's equation provides a framework for understanding wetting[90]:

$$\gamma_{sv} = \gamma_{sl} + \gamma_{lv}\cos(\theta)$$

Here, γ_{sv} is the solid-vapor interfacial tension, γ_{sl} is the solid-liquid interfacial tension, and γ_{lv} is the liquid-vapor interfacial tension. For complete wetting ($\theta = 0°$), the liquid spreads entirely over the solid surface, making the solid-vapor interface energetically unfavorable compared to the solid-liquid and liquid-vapor interfaces. For no wetting ($\theta = 180°$), the liquid minimizes its contact with the solid surface by forming a nearly spherical droplet.

The thermodynamics of wetting can be analyzed using the spreading parameter S, defined as:

$$S = \gamma_{sv} - (\gamma_{sl} + \gamma_{lv}),$$

which quantifies whether complete wetting is energetically favorable. A positive spreading parameter ($S > 0$) indicates that adhesive forces between the liquid and solid are stronger than cohesive forces within the liquid, resulting in complete wetting.

Partial Wetting: The Young Equation

Partial wetting involves three phases: solid, liquid, and vapour, with all combinations of interfaces present. At the edge of the drop, where these phases meet, is the three-phase contact line. On a flat surface, the angle between the solid-vapor (sv) interface and the solid-liquid (sl) interface is geometrically defined as $180°$. The angle between the solid-liquid and liquid-vapor (lv) interfaces is the contact angle, denoted as Young's angle (θ_Y). The contact angle relates to the interfacial tensions in the system. The total surface energy ($E_{\text{surf}}[A]$) is given by:

$$E_{\text{surf}}[A] = \sum_i A_i \gamma_i,$$

where A_i represents interfacial areas for solid-liquid (γ_{sl}), solid-vapor (γ_{sv}), and liquid-vapor (γ_{lv}) interfaces. On a flat substrate with negligible curvature effects, this simplifies to:

$$E_{\text{surf}}[A] = A_{lv}\gamma + A_{sl}(\gamma_{sl} - \gamma_{sv}).$$

where A_{lv} is the area of the liquid-vapor interface and A_{sl} is the area of the solid-liquid interface. The goal is to find the shape A of the drop that minimizes E_{surf}. Near the contact line, we assume the curvature of the liquid-vapor interface is negligible, allowing us to consider it flat.

In equilibrium, any displacement of the contact line must not change the surface energy. If we displace the contact line by a small amount dx, we convert part of the sv interface into sl interface while generating an additional liquid-vapor interface. The variation in interfacial energy can be expressed as:

$$\delta E_{\text{surf}} = dx(\gamma_{sl} - \gamma_{sv} + \gamma\cos\theta).$$

Setting this variation to zero gives us Young's equation:

$$\cos\theta_Y = \frac{\gamma_{sv} - \gamma_{sl}}{\gamma}.$$

If $\gamma_{sl} = \gamma_{sv}$, then $\theta_Y = 90°$, indicating no preference for either phase. If $\gamma_{sv} > \gamma_{sl}$, then $\theta_Y < 90°$, and vice versa. Young's equation assumes partial wetting conditions. If θ_Y approaches $0°$, it indicates complete wetting; if θ_Y approaches $180°$, it indicates no wetting.

In systems involving another liquid (e.g., oil droplets on water), all three interfaces (oil-water, water-air, and oil-air) are deformable. The balance of surface tension forces at the contact line leads to relationships among angles and interfacial tensions described by Neumann's triangle. Neumann's triangle is a geometric representation that relates the interfacial tensions at the three-phase junction:

$$\frac{\gamma_{oa}}{\sin\theta_w} = \frac{\gamma_{wa}}{\sin\theta_o} = \frac{\gamma_{ow}}{\sin\theta_a}.$$

In Neumann's triangle, γ_{oa} represents oil-air interfacial tension, γ_{wa} represents water-air interfacial tension, and γ_{ow} represents oil-water interfacial tension. The angles θ_w, θ_o, and θ_a correspond to angles opposite each respective tension. This equation highlights several key points: First, the contact angle depends on the relative magnitudes of interfacial tensions; a higher tension can lead to a larger contact angle, indicating poorer wetting. Second, smaller contact angles indicate better wetting; if the sum of two interfacial tensions exceeds the third, it results in a smaller angle. Third, this relationship shows how interactions between different phases affect wetting properties. Understanding Neumann's triangle and contact angles has significant applications. In coating technologies, achieving the desired wetting characteristics improves adhesion and uniformity. In biological systems, controlling wetting properties influences cell behaviour and interactions with biomaterials.

Stability diagram

The stability and morphology of double droplets, where two immiscible liquid phases (A and B) coexist in a solvent liquid, depend on the interplay of interfacial tensions ($\gamma_A, \gamma_B, \gamma_{AB}$) and spreading coefficients[91]. Three equilibrium configurations are possible: (i) non-engulfing, where droplets A and B are separated by the host liquid, (ii) complete engulfing, where one droplet fully encapsulates the other, and (iii) partial engulfing, where droplets A and B share a common interface while being exposed to the host phase[92, 93]. The stability of these configurations is governed by the spreading coefficients S_A, S_B, and S_H, defined as: $S_A = \gamma_B - \gamma_A - \gamma_{AB}$, $S_B = \gamma_A - \gamma_B - \gamma_{AB}$, and $S_H = \gamma_{AB} - \gamma_A - \gamma_B$. Stability conditions are as follows: $S_A > 0$ or $S_B > 0$ indicates complete engulfing, $S_H > 0$ corresponds to non-engulfing, and $S_A, S_B, S_H < 0$ ensures partial engulfing.

The equilibrium configuration of the droplets is determined by the force balance at the

three-phase contact line. This force balance leads to the following equations:

$$\gamma_{AB}\cos\theta_B + \gamma_B + \gamma_A\cos(\theta_A+\theta_B) = 0, \tag{1.2}$$

$$\gamma_{AB}\cos\theta_A + \gamma_B\cos(\theta_A+\theta_B) + \gamma_A = 0, \tag{1.3}$$

where θ_A and θ_B are the contact angles formed at the interfaces. These contact angles are further related to the interfacial tensions through:

$$\cos\theta_A = \frac{\gamma_B^2 - \gamma_A^2 - \gamma_{AB}^2}{2\gamma_A\gamma_{AB}}, \tag{1.4}$$

$$\cos\theta_B = \frac{\gamma_A^2 - \gamma_B^2 - \gamma_{AB}^2}{2\gamma_B\gamma_{AB}}. \tag{1.5}$$

For partial engulfing to be stable, the contact angles must satisfy $\theta_A + \theta_B > \pi$, ensuring a wedge-like geometry at the three-phase contact line. The geometry also depends on the volume ratio $k = V_B/V_A$, where V_B and V_A are the volumes of droplets B and A, respectively. The curvature of the interface between droplets varies between concave and convex depending on both k and the contact angles. The curvature R of the interface between

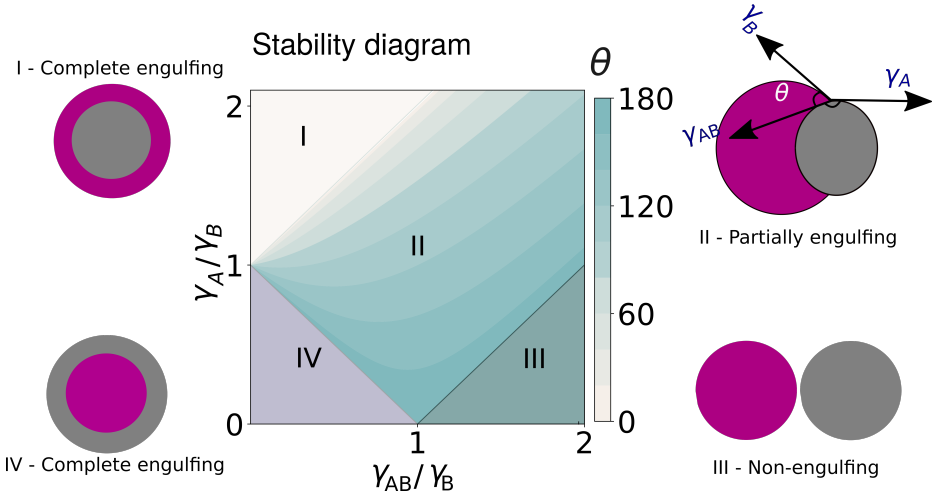

Fig. 1.6: Stability diagram illustrating the possible morphologies of coexisting liquid droplets composed of phases A and B. The equilibrium shapes of the droplets corresponding to 4 distinct morphologies are shown.

droplets A and B is determined by the Laplace pressure balance:

$$\frac{\gamma_B}{R_B} - \frac{\gamma_A}{R_A} = \frac{\gamma_{AB}}{R}, \tag{1.6}$$

where R_A and R_B are the radii of curvature of droplets A and B, respectively, and γ_A, γ_B, and γ_{AB} are the interfacial tensions. This equation reflects the balance of normal forces at the interface between the two immiscible phases. Using the law of sines applied to the Neumann triangle, which describes the equilibrium of interfacial tensions at the three-phase

contact line:

$$\frac{\gamma_B}{\sin\theta_A} = \frac{\gamma_A}{\sin\theta_B} = \frac{\gamma_{AB}}{\sin(\theta_A+\theta_B)}, \tag{1.7}$$

The curvature R can be expressed as:

$$R = \frac{\sin(\theta_A+\theta_B)}{\frac{\sin\theta_B}{R_A} - \frac{\sin\theta_A}{R_B}}. \tag{1.8}$$

In certain limiting cases, the curvature simplifies. When $\theta_A + \theta_B = \pi$, the interface becomes flat, corresponding to infinite curvature ($R \to \infty$). This occurs when the interfacial tensions are perfectly balanced. Conversely, when one droplet forms a perfect sphere (e.g., $\gamma_B/\gamma_{AB} \to 0$), the curvature of the interface becomes infinite.

The stability diagram, constructed using ratios of interfacial tensions, serves as a visual tool for predicting and controlling droplet morphologies (Fig 1.6). The stability diagram for double droplets maps the regions of stable morphologies in terms of normalized interfacial tension ratios, γ_A/γ_{AB} and γ_B/γ_{AB}. Boundaries separating the three configurations are defined by the conditions $\gamma_A = \gamma_B + \gamma_{AB}$, $\gamma_B = \gamma_A + \gamma_{AB}$, and $\gamma_{AB} = \gamma_A + \gamma_B$. Partial engulfing occupies the intermediate region where none of these inequalities hold. This configuration is particularly versatile, as the geometry can be tuned by adjusting interfacial tensions or volume ratios, making it valuable for designing Janus particles, encapsulating substances, and creating synthetic tissues.

Interfaces in Polymer Blends

Interfaces in symmetric binary homopolymer blends are crucial for studying inhomogeneous polymer systems [94, 95]. Intrinsic profiles describe localized, two-dimensional structures of the interface, capturing density and structural variations that are independent of long-range fluctuations. These profiles are centred around a defined "local interface position," essential for understanding polymer interfaces at a microscopic level.

Capillary waves, which are thermal fluctuations along the interface, cause transverse, long-wavelength disturbances that broaden the apparent interfacial width (w). This broadening depends logarithmically on the system size (L) and is mathematically expressed as:

$$w^2 = w_0^2 + \frac{4\gamma k_B T}{\ln(a_0 L)} \tag{1.9}$$

where w_0 is the intrinsic width of the interface, γ is the interfacial tension, a_0 is a coarse-graining length scale, k_B is Boltzmann's constant, and T is the absolute temperature. The capillary wave Hamiltonian (H_{CW}), which quantifies the energy associated with these fluctuations, is given by:

$$H_{CW} = \frac{\gamma}{2}\int dx\, dy\, |\nabla h|^2$$

where $h(x, y)$ is the local interface height. As system size increases, larger capillary waves contribute to a greater apparent width due to their cumulative effect.

Thermodynamic stability at polymer interfaces arises from balancing the minimization of the interfacial area and maximizing the conformational entropy of the polymers. The interfacial energy plays a critical role in determining how polymers behave at interfaces, affecting their adsorption and desorption. The width influenced by capillary waves correlates with mechanical stability; wider interfaces typically have fewer inter-chain entanglements[96]. A broad interface allows polymers to entangle across it, which strengthens the interface by enhancing interactions between the different phases. This entanglement forms a network that improves mechanical stability. In contrast, sharp interfaces rely predominantly on weak van der Waals forces to maintain phase cohesion, resulting in reduced mechanical stability. Therefore, the type of interface, broad or sharp, significantly affects the strength and stability of polymer blends.

The dynamic behaviour of polymers at interfaces differs significantly from that in bulk due to confinement effects. Intrinsic profiles exclude capillary wave effects, providing an "idealized" view of the interface. Understanding these concepts is vital for accurately predicting interfacial behaviour and improving material performance in polymer blends.

1.2.4 Classical Molecular Dynamics - Theory

Computer simulations, particularly molecular dynamics (MD), are crucial for understanding biomolecular systems[97, 98, 99]. These computational methods reveal the dynamic behaviour of biological molecules like proteins and nucleic acids, capturing molecular movements that are invisible in static experimental structures. By applying physical principles and computational algorithms, MD simulations model how molecules move and interact under various conditions, which helps to track complex processes such as protein folding, ligand binding, and conformational changes with unprecedented atomic-level detail. MD simulations bridge the gap between experimental observations and molecular mechanisms, providing insights into biological systems across multiple scientific domains. They are particularly valuable in structural biology, drug discovery, and understanding cellular processes at the molecular level. The technique's effectiveness stems from its ability to predict atom movements using established physical laws, offering a comprehensive view of molecular interactions. As computational power and simulation techniques improve, MD continues to be an essential tool for exploring the intricate world of biomolecular systems. The following section looks deep into how such systems are modeled using physics-based principles.

Classical molecular dynamics (MD) is a computational technique used to simulate the physical movements of particles and molecules over time. This approach is particularly valuable for studying large biomolecular systems. In classical MD, the forces acting on particles are calculated based on their interactions with other particles. The forces are derived from the

potential energy of the system. The core equations are based on Newton's second law:

$$m_i \ddot{\mathbf{r}}_i = \mathbf{f}_i \tag{1.10}$$

where m_i is the mass of particle i, $\ddot{\mathbf{r}}_i$ is its acceleration, and $\mathbf{f_i}$ is the force acting on it. Alternatively, in Hamiltonian mechanics, the equations can be expressed as:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i}, \quad \dot{\mathbf{p}}_i = \mathbf{f}_i \tag{1.11}$$

where p_i is the momentum of particle i. The force is derived from the potential energy $U(r)$:

$$f_i = -\frac{\partial U}{\partial r_i} = -\nabla U \tag{1.12}$$

The MD simulation process involves initializing positions and velocities, calculating forces based on potential energy, and numerically integrating these equations over time. Integrators are based on the Verlet and leapfrog algorithms. In the leapfrog algorithm, velocity and position are updated by:
Acceleration Calculation:

$$\ddot{\mathbf{r}}_i = -\frac{\nabla U(r_i(t))}{m_i}$$

Velocity Update:

$$\dot{\mathbf{r}}_i(t + \frac{1}{2}\Delta t) = \dot{\mathbf{r}}_i(t - \frac{1}{2}\Delta t) + \ddot{\mathbf{r}}_i(t) \cdot \Delta t$$

Position Update:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \dot{\mathbf{r}}_i(t + \frac{1}{2}\Delta t) \cdot \Delta t$$

Choosing an appropriate time step (δt) is important for accuracy. it should be maximised while still capturing essential dynamics and must be small to conserve energy over long simulation times. The leapfrog algorithm requires initial velocities to start a new simulation. These initial velocities are usually generated randomly and follow the Maxwell-Boltzmann distribution. The probability p of a particle having a velocity $\dot{\mathbf{r}}_i$ is given by:

$$p(\dot{r}_i) = \sqrt{\frac{m_i}{2\pi k_B T}} \cdot \exp\left(-\frac{m_i \dot{r}_i^2}{2k_B T}\right)$$

Here, k_B is the Boltzmann constant and T is the absolute temperature.

Interactions in Molecular Dynamics

Non-bonded interactions
Non-bonded interactions refer to the potential energy between particles that are not directly

connected by covalent bonds. The non-bonded potential energy $U_{nb}(r)$ is expressed as[99]:

$$U_{nb}(r) = \sum_i U_1(r_i) + \sum_{i,j>i} U_2(r_i, r_j) + \sum_{i,j,k} U_3(r_i, r_j, r_k) + \dots \tag{1.13}$$

In this equation, $U_1(r_i)$ is the one-body potential for individual particles, $U_2(r_i, r_j)$ is the two-body (pair) potential for interactions between pairs of particles, and $U_3(r_i, r_j, r_k)$ is the three-body potential for interactions among triplets of particles. In simulations of bulk systems, contributions from external fields or container walls are typically ignored. The primary focus is on pair potentials, which dominate non-bonded interactions:

$$E_{non-bond} = E_{vdW} + E_{el} \tag{1.14}$$

Here, E_{vdW} represents van der Waals forces and E_{el} represents electrostatic interactions. The Lennard-Jones is an effective model used to describe how neutral particles, like inert gases, interact with one another. It captures both the attractive and repulsive forces between particles. The Lennard-Jones potential between two particles i and j can be expressed as:

$$V_{LJ}(r_{ij}) = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{1.15}$$

In this equation, ϵ_i indicates how strongly the particles attract each other. The parameter σ_i represents the size of the particles. The variable r_{ij} is the distance between two particles. The parameters ϵ_{ij} and σ_{ij} can either be set individually for each pair of atom types or calculated using averages based on their properties. The attractive force comes from the Van der Waals interactions, represented by the r^{-6} term and the repulsive force is represented by the r^{-12} term. Electrostatic interactions between two particles depend on their partial charges:

$$V_{el}(r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{\varepsilon_r r_{ij}} \tag{1.16}$$

Here, q_i and q_j are the charges of the particles, r_{ij} is the distance between them, ε_0 is a constant (the permittivity of free space). Calculating these non-bonded interactions can be very difficult as the number of atom pairs increases with the number of particles (N) in the system (scaling as $O(N^2)$). To simplify calculations, a cut-off distance is used, to ignore pairs that are farther apart than this distance. To avoid sudden changes in energy at this cut-off point, potentials are adjusted to 0 at this cut-off.

Bonded interactions

In molecular dynamics, bonded interactions describe how particles connected by chemical bonds behave[98]. These interactions are typically modelled using harmonic potentials,

which approximate the energy associated with bonds, angles, and dihedrals:

$$U_{\text{intramolecular}} = \frac{1}{2}\sum_{\text{bonds}} k_r^{ij}(r_{ij} - r_{\text{eq}})^2 + \frac{1}{2}\sum_{\text{bend angles}} k_\theta^{ijk}(\theta_{ijk} - \theta_{\text{eq}})^2 \tag{1.17}$$

$$+ \frac{1}{2}\sum_{\text{torsion angles}}\sum_{m} k_\phi^{ijkl}(1 + \cos(n\phi - \phi_{eq})) \tag{1.18}$$

The first term describes the potential energy V_b of a covalent bond between two particles i and j is given by:

$$V_r(r_{ij}) = \frac{1}{2}k_r^{ij}(r_{ij} - r_{\text{eq}})^2$$

It accounts for the stretching between adjacent particles. Here, k_r^{ij} is the force constant, r_{ij} is the distance between the particles, and r_{eq} is the equilibrium distance. The second term describe the potential energy V_θ formed by three covalently bonded particles i, j, k is given by:

$$V_\theta(\theta_{ijk}) = \frac{1}{2}k_\theta^{ijk}(\theta_{ijk} - \theta_{\text{eq}})^2$$

In this case, k_θ^{ijk} is the force constant for the angle, and θ_{eq} is the equilibrium angle. And it accounts for angles formed by three particles. The third term describes The potential energy V_d for a dihedral angle formed by four particles is periodic and described using:

$$V_d(\phi_{ijkl}) = k_\phi^{ijkl}(1 + \cos(n\phi - \phi_{eq}))$$

Here, k_ϕ^{ijkl} is the force constant for the angle, and ϕ_{eq} is the equilibrium angle. n is the periodicity of the angle. It accounts for rotations around bonds involving four particles.

1.2.5 Sampling in molecular dynamics

Adequate sampling in molecular dynamics (MD) is essential to ensure that simulations effectively capture all relevant conformational states of a system[100, 101]. This is particularly important in complex energy landscapes that have numerous local minima and high activation barriers, which can trap the system in non-functional states. Insufficient sampling can lead to inaccurate characterizations of biomolecular dynamics. To address this challenge, enhanced sampling techniques have been developed. These methods, such as replica-exchange molecular dynamics and metadynamics, facilitate the exploration of phase space. For example, the replica-exchange Molecular Dynamics (REMD) technique involves running multiple simulations (replicas) at different temperatures[102, 103]. Periodically, replicas can exchange configurations, allowing the system to overcome energy barriers and explore a wider range of conformations. These methods enable researchers to study processes that occur over long timescales and better understand the dynamic behaviour and functional properties of biomolecules.

1.2.6 Atomistic and coarse-grained simulations

Atomistic simulations are powerful tools for studying materials at the atomic level[104, 97, 105]. They help analyze specific interactions and molecular dynamics within polymer systems, providing insights into atomic behaviour under various conditions. These simulations effectively characterize interactions, such as the forces within polymer chains and between chains and solvents. However, atomistic models have limitations due to their high computational cost and the short length scales they can probe, typically ranging from 1 to 100 Å and time scales of 1 fs to 100 ns. This restricts their use for studying long-term relaxation processes in polymers with many repeat units. Selecting the appropriate force field is crucial for accurate simulations. Common force fields include AMBER and CHARMM, each with unique parameters optimized for specific conditions[106, 107].

Coarse-Graining in Polymer Science

Coarse-grained (CG) models simplify polymer systems by representing chains as interconnected "beads," each corresponding to a segment of the polymer[108, 109]. This approach allows for the efficient study of large-scale phenomena and employs varying resolutions, such as monomer-level CG, Kuhn-segment models, and mesoscale models[110, 111]. The choice of bead interactions and mappings is crucial for maintaining structural and thermodynamic accuracy while optimizing computational efficiency.

Coarse-graining originated in polymer science and has become widely applied in materials science, addressing the multi-scale nature of polymers that exhibit structures spanning Angstroms to micrometres. At the atomic scale, polymers are analyzed through detailed chemical interactions, whereas mesoscale approaches use simplified models like bead-spring or lattice models based on statistical physics. These models combine multiple real monomers into coarse-grained units, allowing flexibility in their representation.

The choice of how to coarse-grain a polymer depends on the specific properties being studied and the available computational resources. Polymers exhibit universal properties that connect various coarse-grained models, allowing for meaningful connections between various representations. This universality is evident when analyzing polymers at scales larger than the individual monomer level, where distinct models - such as atomistic and coarse-grained representations-demonstrate similar static and dynamic properties [112, 113]. For instance, studies like [112] show that static properties remain consistent across coarse-graining levels, while dynamic properties depend on the degree of simplification. By using a single scaling factor, the study combines mean squared displacement (MSD) data from coarse-grained and atomistic simulations. By applying scaling factors, such research bridges microscopic and macroscopic behaviours, validating CG model's ability to replicate key polymer dynamics. This universality underpins the flexibility of CG approaches, enabling simulations of large systems over extended timescales, which is crucial for investigating

phase transitions, self-assembly, and molecular interactions.

Coarse-graining enhances computational efficiency, allowing the simulation of large systems over longer timescales. By strategically balancing resolution and efficiency, CG modelling has become a cornerstone of polymer science, offering both theoretical insights and practical utility. This is especially useful for studying complex phenomena like polymer phase transitions, self-assembly, and interactions with other molecules. Together, these multi-scale approaches enhance our ability to predict and manipulate polymer behaviour, linking molecular-level insights to macroscopic functionality in fields ranging from biomolecular engineering to industrial material design.

The Stochastic Simulation Algorithm

Biochemical reactions in cells are stochastic due to the discrete nature of molecular interactions and the random timing of reaction events. Reactions occur as discrete events governed by probabilistic interactions between reactants. Randomness arises because of low copy numbers of molecules, with many key regulatory molecules, such as transcription factors, mRNA molecules, and signaling proteins, existing in very small quantities. Molecular motion governed by Brownian motion causes random encounters between reactants, while localized concentration fluctuations further enhance randomness. The timing and sequence of these reactions are not predetermined but follow statistical distributions. In larger molecular populations, stochastic variations average out, making deterministic models sufficient. However, in small systems, fluctuations significantly impact behavior, necessitating stochastic modelling.

Stochastic Chemical Kinetics

Stochastic chemical kinetics models the time evolution of chemical systems by considering discrete nature of molecular populations and probabilistic reaction events[114, 115]. The system consist of N chemical species $S_1, S_2, \ldots, S_N$, and they interact through M reactions $R_1, R_2, \ldots, R_M$. For example, a unimolecular reaction: $S_j \rightarrow Product$, is described by a propensity function $a_j(x)$, which describes the probability of a reaction happening in a time interval dt. For a unimolecular reaction: $a_j(x) = c_j h_j$, where c_j is the rate constant for the reaction, and h_j is the number of molecule of species S_j.

Stochastic models use probability distributions to describe how molecular population changes over time. The chemical Master equation provides a fundamental framework for this. It describes how the probability $P(x, t|x_0, t_0)$ of the system being in state x at time t, given that

it was in state x_0 at time t_0, evolves. The chemical Master equation is expressed as:

$$\frac{\partial P(x,t|x_0,t_0)}{\partial t} = \sum_{j=1}^{M} [a_j(x-\nu_j)P(x-\nu_j,t|x_0,t_0) - a_j(x)P(x,t|x_0,t_0).]$$

This equation takes into account all the ways the system can change state via each of the M reactions. However, solving the chemical Master equation exactly is computationally challenging for systems with many reactions or bimolecular interactions.

Gillespie Algorithm

The Gillespie algorithm is a stochastic simulation method that provides an alternative way to simulate stochastic dynamics[116, 117, 118]. The algorithm leverages the propensity functions $a_j(x)$ to determine the time until the next reaction and which specific reaction will occur. Propensity functions are used to sample random numbers that dictate the waiting. By iteratively applying these steps, the Gillespie algorithm generates sample trajectories that represent the evolution of the system.

1.2.7 Challenges in Studying Biomolecular Condensates

Studying biomolecular condensates presents several challenges despite their importance in cellular processes[13, 119, 120]. One major difficulty is their complexity, as these condensates involve many interacting components, making it hard to fully characterize their composition and behaviour. This complexity is further complicated by the dynamic nature of condensates, which can rapidly form and dissolve, making experimental capture and analysis difficult[121]. Another significant challenge is the vast sequence space of intrinsically disordered regions (IDRs) that contribute to condensate formation. Predicting which sequences will phase separate and under what conditions is complex and often requires time-consuming experimental methods[27]. Additionally, the non-equilibrium nature of condensates in living cells adds to the difficulty; in vivo, these structures undergo constant chemical modifications, such as phosphorylation and dephosphorylation cycles, which can greatly influence their formation, stability, and function[122]. The heterogeneity of biomolecular condensates also complicates research. Different types of condensates can have unique properties and functions due to multiple interacting partners within a single structure, making it challenging to generalize findings across different systems. Moreover, existing experimental techniques may lack the resolution needed to observe condensate dynamics in real time without disrupting normal cellular processes. Finally, computational models must accurately represent the complex interactions within these systems. This requires advancements in modeling approaches and computational power to handle the large amounts of data generated by simulations[123]. Addressing these challenges is essential for improving our understanding of biomolecular condensates and their roles in cellular

processes.

Computational Strategies for Studying Condensates Using MD Simulations

Computational methods and machine learning (ML) provide effective tools to address the challenges of studying biomolecular condensates.

Multiscale Simulation Approaches

Multiscale simulations are useful for studying IDRs and their phase behaviour as they allow exploration across different length and time scales. By combining atomistic models with coarse-grained representations, these simulations capture both detailed molecular interactions and broader conformational dynamics of IDRs. This approach provides insights into the thermodynamic stability of dense phases and the dynamics of condensate formation. Enhanced sampling techniques enable sampling the complex conformational landscapes of biomolecular condensates.

Monte Carlo Methods for Stochastic Simulations

Monte Carlo simulations complement MD simulations by modeling stochastic behaviors under nonequilibrium conditions, capturing fluctuations that deterministic methods may miss.

Integration of Experimental Data

Integrating experimental data into molecular simulations is crucial for refining models, validating predictions, and improving force fields to study biomolecular condensates. This process ensures consistency between computational outputs and experimental observations, enabling accurate representation. Bayesian methods provide a tool for integrating experimental data with molecular simulations by constructing probabilistic models that account for uncertainties in both experimental measurements and simulation predictions.

Machine Learning Applications

Machine learning (ML) is significantly enhancing the analysis of IDRs by offering advantages over traditional molecular dynamics (MD) simulations. Neural networks identify sequence features influencing condensate formation and help to design novel IDRs with desired properties using generative models. ML reduces the need for extensive MD simulations or experiments by efficiently exploring sequence space.[124]

Experimental Validation

Simulations are validated by comparing computational predictions with experimental data from techniques such as fluorescence microscopy or nuclear magnetic resonance spectroscopy.

Despite these advances, predicting protein phase-separation behavior remains difficult due to the vast sequence space and complex sequence–property relationships. MD simulations provide valuable insights but are computationally expensive for large-scale studies.

Machine Learning and Bayesian Optimization for Predicting and Designing Phase-Separating Proteins

Machine learning (ML) offers a promising alternative for modeling these intricate relationships[125]. Recent developments have demonstrated that ML models trained on experimental and simulation data can predict phase separation propensities directly from protein sequences[126, 127].

Active learning frameworks combined with Bayesian Optimization (BO) have emerged as powerful tools for efficiently exploring sequence space and are applied to design novel proteins with desired phase-separation properties. [128]. By identifying the most informative sequences for training, BO optimizes resource allocation while improving predictive accuracy.

These advancements highlight the synergy between physics-based simulations and ML techniques in addressing fundamental questions about protein phase separation.

1.3 Motivation

Gene expression is regulated by RNA Polymerase II (Pol II) condensates, which are crucial for development, health, and disease. These phase-separated condensates fine-tune transcription by creating distinct chemical environments, but their precise nature, dynamics, and sequence-property relationships have remained elusive. The complexity of these condensates, their dynamic nature, and the vast protein sequence space involved have posed significant challenges to their study. This research seeks to address these challenges by integrating multi-scale simulations with experimental observations in C. elegans, employing active learning approaches, and exploring chemically-driven non-equilibrium dynamics. This approach has revealed crucial insights into Pol II condensate behaviour, including their temperature-dependent stability, the formation of distinct chemical environments for different transcription stages, the prediction of peptide interactions using Bayesian optimisations, and the regulatory role of non-equilibrium phosphorylation cycles. By unravelling these mechanisms, this work not only improves our understanding of transcriptional regulation but also provides a robust framework for investigating other phase-separated systems in cells, potentially opening new avenues for therapeutic interventions in diseases associated with dysregulated gene expression.

2

Phase-Separated Promoter and Gene-Body Condensates: A Phosphorylation-Driven Mechanism for Transcription Control

This chapter is based on the preprint 'Promoter and Gene-Body RNA-Polymerase II co-exist in partial demixed condensates, ' which has been incorporated into the thesis with minor modifications for structural consistency. Supplementary texts and figures have been integrated where appropriate. These modifications include restructuring for coherence, adding explanatory details, and omitting redundant sections. Additions are highlighted in blue, while omitted sections are indicated by [...] in blue.

Contributions:

- **Research Hypothesis Development:** Contributed to the formulation of the research hypothesis and objectives, guiding the direction of the work.
- **Simulation setup and execution**: I was responsible for setting up and executing all the simulations for this study.
- **Data analysis**: I conducted all the analysis from the simulation trajectories.
- **Figure preparation**: I created all the plots related to simulation in the paper.
- **Manuscript contribution**: I contributed to writing and revising the manuscript.

Reprint permissions:

Reprinted from " Promoter and Gene-Body RNA-Polymerase II co-exist in partial demixed condensates Arya Changiarath, David Flores-Solis, Jasper J. Michels, Rosa Herrera Rodriguez, Sonya M. Hanson, Friederike Schmid, Markus Zweckstetter, Jan Padeken, Lukas S. Stelzl bioRxiv 2024.03.16.585180; doi: https://doi.org/10.1101/2024.03.16.585180" with the permissions.

This chapter investigates the formation and properties of RNA Polymerase II (Pol II) condensates and their role in regulating transcription. Formed through liquid-liquid phase separation, these condensates create distinct environments that fine-tune Pol II activity during transcription initiation and elongation. Using multi-scale simulations and experimental observations, this study assesses whether the phases for initiation and elongation mix or remain distinct. We also explore how the biophysical properties of Pol II condensate, including their temperature-dependent behaviour and the effects of CTD phosphorylation, influence their function. Our findings reveal that Pol II can form multiple distinct condensates, aiding in the selective recruitment of factors during different transcription phases, thereby enhancing our understanding of transcriptional regulation and the role of biomolecular condensates in cellular processes.

2.1 Introduction

Phase separation is a fundamental mechanism employed by cells to form condensates, enabling precise control of biological processes through localized differences in protein concentrations[129]. This mechanism is crucial for the spatio-temporal regulation of cellular functions, often requiring the presence of multiple, co-existing condensates with distinct, unmixed phases[130].

RNA Polymerase II (Pol II) and its C-Terminal Domain (CTD) exemplify this complexity in transcriptional regulation. The intrinsically disordered CTD plays a crucial role in forming phase-separated condensates during the distinct stages of transcription: initiation, elongation, and termination. These condensates are involved in various processes, from the recruitment of RNA Pol II and initiation of transcription in promoter condensates to later stages of transcription elongation in gene-body condensates. As transcription progresses, the CTD undergoes phosphorylation, impacting its phase behaviour[131, 132, 133].

Several proteins that regulate transcription elongation, such as P-TEFb, localize within these condensates, either directly or indirectly, but also splicing factors. Recent research has proposed a surface condensation model to explain the dynamics of RNA Polymerase II as it transitions between stages, based on super-resolution experiments in zebrafish. Nascent RNA is immediately processed[133]. RNA transcripts are spliced, removing the non-coding regions (introns). Interestingly, splicing condensates remain distinct and demixed from initiation condensates, indicating a complex interplay between different types of transcription-related condensates[134, 135, 136, 137].

While significant progress has been made in understanding these processes, several knowledge gaps persist. The specific factors determining the existence and behaviour of multiple condensates, the effects of CTD phosphorylation on condensate behavior, and the interactions between different transcription-related condensates when in direct contact require further

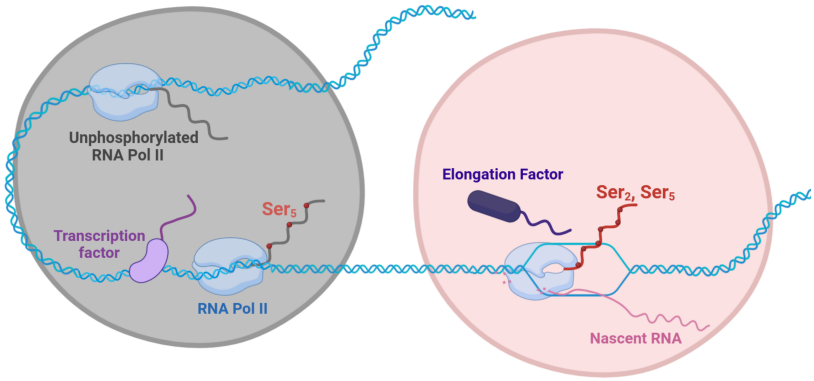

Fig. 2.1: Hypothesized condensates underlying the initiation of gene transcription and the elongation phase. The figure is adapted from Figure 1 of reference [147]

investigation. Current experimental techniques often lack the spatial and temporal resolution needed to observe condensate dynamics in real-time within living cells, and the complexity of cellular environments makes it challenging to isolate and study individual condensates effectively.

To address these challenges, we employed a multifaceted approach combining molecular simulations[138, 139, 140, 141, 142, 10, 9, 143] with experimental observations in *C. elegans* and in vitro. We used both coarse-grained and atomistic molecular dynamics simulations to model the effects of CTD phosphorylation and its interactions with the HRD domain of cyclin T1 on condensate formation[144, 145, 146]. These simulations predicted that CTD forms condensates with lower critical solution temperature (LCST) behaviour, distinguishing them from other nuclear condensates. Additionally, we found that CTD and phosphorylated CTD (pCTD) form distinct phases, with pCTD partially engulfing CTD.

To validate these predictions, our experimental collaborators conducted experiments in *C. elegans* embryos using super-resolution microscopy and in vitro. Their results confirmed the LCST behavior of CTD condensates and the partial engulfment of CTD by pCTD, aligning with our simulation findings. This integrated approach provided new insights into the biophysical properties of Pol II condensates and their potential role in regulating the transcription cycle, contributing to our understanding of complex, multi-phase condensates in diverse cellular contexts.

2.2 Methods

Coarse-grained simulation model

We employed coarse-grained molecular dynamics simulations using the residue-level hydropathy scale (HPS) model, which represents amino acids as beads while considering solvent effects implicitly[139]. This model effectively captures the physical and chemical attributes

of the 20 natural amino acids, resulting in sequence-specific interactions among the beads. To describe these interactions, we considered three key types. First, beads are bonded to each other through harmonic potential energy, represented by, $V_H(r) = k(r - r_0)^2/2$ where $k = 10\,\mathrm{kJ\,\AA^{-2}}$ is the force constant, and $r_0 = 3.8\,\text{Å}$ is the bond length.

Second, electrostatic interactions between charged amino acids are modelled using the Debye–Hückel (DH) electrostatic screening term:

$$E_{ij} = \frac{q_i q_j}{4\pi D r} \exp\left(-\frac{r}{k}\right),$$

where $k = 1\,\text{nm}$ corresponds to a salt concentration of $100\,\text{mM}$ and $D = 80$ is the dielectric constant of the aqueous solvent. A cut-off distance of $r_{\text{cut}} = 3.5\,\text{Å}$ is established; if r exceeds this distance, then E_{ij} is set to zero.

The third type of interaction is represented by the Ashbaugh–Hatch potential for short-range pairwise interactions:

$$\phi(r) = \begin{cases} \phi_{LJ}(r) + (1 - \lambda_{ij})\epsilon, & \text{if } r < 2^{1/6}\sigma_{ij}. \\ \lambda_{ij}\phi_{LJ}(r), & \text{otherwise.} \end{cases}$$

Here, ϕ_{LJ} is defined by the Lennard-Jones potential:

$$\phi_{LJ}(r) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right], & \text{if } r < r_{\text{cut}}. \\ 0, & \text{otherwise.} \end{cases}$$

The parameters λ_{ij} and σ_{ij} are the arithmetic average of the hydrophobicity scale and size of amino acids, respectively. A cut-off distance of $r_{cut} = 2.00$ is used and the potential is shifted accordingly. The residue-specific hydrophobicity values range from 0 to 1 as the hydrophobicity increases. The HPS model also provides parameters for phosphorylated amino acids such as phospho-serine[148], which have been used to model the effect of phosphorylation on the fifth position in the CTD heptad repeat motifs, which we refer to as Ser_5. To study temperature-dependent phase behaviour we used the HPS-T model, a version of the HPS model which has been extended to predict the temperature dependent properties of disordered proteins[149].

Coarse-grained molecular dynamics simulations were run with the HOOMD-Blue package[150]. The system is initialized by placing the protein chains inside an extended rectangular box with a minimum distance between protein chains. The simulations were performed under *NVT* ensemble conditions, using a Langevin thermostat with a friction coefficient of $0.01ps^{-1}$. Equilibration occurred within the initial nanoseconds, resulting in the formation of liquid-dense and dilute phases. We extracted the density profile by analysing protein

concentration along the z-axis, following the methodology described by Tesei et al.[141].

Atomistic molecular dynamics simulations

Two 21mer CTD fragments were simulated in all-atom molecular dynamics with explicit solvent. We employed the Amber99sb-star-ildn-q protein force field[151, 152, 153, 154] and the TIP4P-D water model[155] in the simulations. Equations of motions were integrated with a $2\,\mathrm{fs}$ time step. The *NPT* ensemble was maintained at $300\,\mathrm{K}$ and $1\,\mathrm{bar}$ with the Bussi-Donadio-Parrinello thermostat[156] and the Parrinello-Rahman barostat[157]. Simulations were run in GROMACS[158], with a cumulative simulation duration of $39.8\,\mu\mathrm{s}$.

Simulated systems

Coarse-grained molecular dynamics simulations provide a powerful tool for elucidating the intricate mechanisms underlying gene transcription regulation. To understand how deviations from ideal heptad repeat sequence $\mathrm{Y_1S_2P_3T_4S_5P_6S_7}$ affect phase behaviour, we conducted simulations using chains of equivalent length but with idealized heptad repeat sequence. Additionally, we employed the HPS-T model to investigate the temperature-dependent phase behaviour of *C. elegans* CTD[149].

To study the formation of distinct phases or multiphasic condensates we focus on a simplified system, comprising three key components that play a vital role in the transcription process and may underlie key aspects of phase behaviour in the nucleus. These elements consist of CTD of RNA polymerase II (140 mer with 20 ideal heptad repeats), responsible for the promoter condensate; phosphorylated CTD (140 mer with various degrees of phosphorylation); and Histidine Rich Domains (HRD) of P-TEFb Cyclin T1, contributing to the formation of the gene body condensate (Fig. 2.1A). The coarse-grained nature of our simulations allowed us to explore large-scale and long-term thermodynamic and material properties. This was essential for understanding the stability and structure of multiphasic condensates. We began by studying the formation and regulation of the promoter condensate, which forms from CTD phase separation.

We conducted a series of simulations to elucidate the impact of phosphorylation on CTD phase separation and the co-phase behavior of HRD with CTD and pCTD. The coacervation behavior of HRD and pCTD was analyzed using Flory-Huggins theory. To comprehensively understand how phosphorylation influences multiphase condensate formation, we performed simulations involving CTD, pCTD, and HRD, systematically varying pCTD phosphorylation levels. Furthermore, to investigate the structure and stability of these multiphase condensates, we conducted simulations of condensates in a cubic box. This approach allowed us to gain insights into the spatial organization and interactions within these complex biomolecular assemblies.

Analysis of phase behaviour with a mean-field model

To elucidate the thermodynamic principles governing the co-phase separation of phosphorylated C-terminal domain (pCTD) and Histidine-Rich Domain (HRD), we employed a mean-field Flory-Huggins model to analyse the simulated phase diagram, providing a quantitative framework for understanding the complex interactions and phase behaviour of these biomolecular components[159],

$$\frac{\Delta f}{k_B T} = \frac{\phi_A}{N_A} \ln \phi_A + \frac{\phi_B}{N_B} \ln \phi_B + \phi_S \ln \phi_S + \chi_{AB}\phi_A\phi_B + \chi_{AS}\phi_A\phi_S + \chi_{BS}\phi_B\phi_S, \tag{2.1}$$

Here, ϕ_i and N_i represent the volume fractions and relative degrees of polymerisation and χ_{ij} the binary interaction parameters. In the model, ϕ_S follows from the assumption of incompressibility: $\phi_S = 1 - \phi_A - \phi_B$. Subscripts A, B, and S respectively denote HRD, pCDT, and "solvent". The relative sizes of the polymers, which we fix at $N_A = 70$ and $N_B = 140$, have been normalised by a scaled monomeric volume ($N_S = 1$).

The mean-field phase diagram was obtained in the usual way by demanding that the chemical potential and the osmotic pressure be similar in the dilute and dense phase, using well-known methodologies[160].

Calculation of interfacial tension

Interfacial tension is determined using two methods employing capillary wave theory: first, by assessing interfacial width in the slab simulations, and second, by analysing shape fluctuations of droplets.

For slab geometries, we utilized capillary wave theory (CWT) with the block analysis method to estimate condensate interfacial tension [94]. The interface region was divided into rectangular segments of dimensions $L_x/n_B \times L_y/n_B \times L_z$, where n_B is an integer, where we choose $3 < n_B < 8$. Within each segment, intrinsic profiles are computed based on the density function,

$$\rho(z) = \frac{\rho_A + \rho_B}{2} + \frac{\rho_A - \rho_B}{2} \tanh\left(\frac{z - z_0}{W_0}\right) \tag{2.2}$$

where ρ_A and ρ_B represent the bulk densities of two phases, denoted as A and B, respectively. z_0 is the position of the interface, while W_0 corresponds to the intrinsic width of the system. The apparent interfacial width (W) at a finite temperature is determined by combining the intrinsic width W_0 and the capillary contribution:

$$W^2 = W_0^2 + \frac{k_B T}{2\pi\gamma} \ln\left(\frac{B}{L}\right), \tag{2.3}$$

where k_BT is the thermal energy, γ represents the interfacial tension, and B/L is the ratio of block size to system size.

With CWT, we can determine the interfacial tension at the interface between a dense protein phase with the solvent, as well as at the interface between two distinct protein phases.

We compared the interfacial tension of the homogeneous condensate obtained from the above method with the interfacial tension calculated using the Kirkwood-Buff relation [161],

$$\gamma = \frac{L_z}{2}\left(P_{zz} - \frac{P_{xx}+P_{yy}}{2}\right), \tag{2.4}$$

where, $P_{\alpha\beta}$ are the components of the pressure tensor.

For droplet analysis, the interfacial tension is estimated from the thermal fluctuations of its spherical shape based on the method pioneered by Henderson and Lekner[162], which has recently been applied in the analysis of biological condensates[140]. The potential energy (U) associated with these shape fluctuations is primarily determined by the interfacial tension (γ), which can be mathematically expressed as $U = \gamma\delta A$. Here, δA is the small variation in interfacial area relative to that of a perfect sphere. By analysing the droplet's shape fluctuations at the lowest order using a spherical harmonics expansion, we can derive two independent measures of the interfacial tension:

$$\gamma_{20} = \frac{5k_BT}{16\pi\langle(\delta a+\delta b)^2\rangle} \tag{2.5}$$

and

$$\gamma_{22} = \frac{5k_BT}{16\pi\langle(\delta a-\delta b)^2\rangle} \tag{2.6}$$

Here, the droplet's instantaneous shape is described by a general ellipsoid, with δa and δb representing the lengths of any pair of principal axes of the ellipsoid, while R indicates the average radius of the droplet.

Characterisation of disordered protein conformation

In our study, we investigate the impact of phosphorylation on the conformation and compactness of C-terminal domain (CTD) chains in both dilute and dense phases. A key metric for this analysis is the radius of gyration (R_G), which provides insight into the spatial extent of the protein chains. Additionally, we assess protein anisotropy to identify structural changes resulting from phosphorylation. The radius of gyration of chains is calculated using the following equation:

$$\mathbf{G} = \frac{1}{M}\sum_{i=1}^{N} m_i \Delta\mathbf{r}_i \Delta\mathbf{r}_i^T \tag{2.7}$$

where M represents the total mass of the chain, N is the number of monomers in the chain, and $\Delta\mathbf{r}_i$ is the position vector of monomer i relative to the chain's center-of-mass. The root mean square radius of gyration is derived from the diagonal elements of the **G** matrix: $R_G = \langle Gxx + G_{yy} + G_{zz} \rangle^{1/2}$. To quantify the structural anisotropy of the chains, we employ the following equation[161]:

$$A = \left\langle 1 - 3\frac{G_1G_2 + G_1G_3 + G_2G_3}{(G_1 + G_2 + G_3)^2} \right\rangle \tag{2.8}$$

where G_1, G_2, G_3 represent the three eigenvalues of the gyration tensor **G**. This anisotropy analysis allows us to characterize structural modifications and deviations from symmetric conformations that may arise due to phosphorylation.

Interactions of disordered protein chains in condensates

To quantitatively characterise the interactions between disordered protein chains within condensates, we employed a statistical mechanical approach that calculates the excess of one monomer type in the vicinity of another, providing a robust metric for assessing protein-protein interactions, denoted as n_{ab}^{exc}: where r_1 and r_2 represent the minimum and maximum radial distances from the B monomer, respectively, over which the excess number of monomer A is computed. We specified cutoff values of $r_1 = 0.45, \text{nm}$ and $r_2 = 7.5, \text{nm}$. The correlation function $g_{ba}(r)$ is defined as:

$$g_{ab}(r) = \frac{V}{N_b N_a} \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} \frac{\delta(|\vec{r}_i - \vec{r}_j| - r)}{4\pi r^2} \tag{2.9}$$

where N_a and N_b represent the total number of monomers of type A and B, respectively, and V is the system volume. [...]

2.3 Results and Discussion

Understanding RNA Polymerase II CTD Phase Separation through Molecular Dynamics Simulations

The carboxy-terminal domain (CTD) of RNA polymerase II is essential for regulating transcription in eukaryotic cells. This region, composed of repeating heptad units, is crucial for transcription initiation and RNA processing. Recent research combining experimental methods and computational modelling has improved our understanding of how CTD phase separation works and its implications for gene expression. Residue-level coarse-grained (CG) molecular dynamics (MD) simulations are effective for studying the interactions of CTD. These simulations revealed how phase separation is driven by heptad repeat interactions and

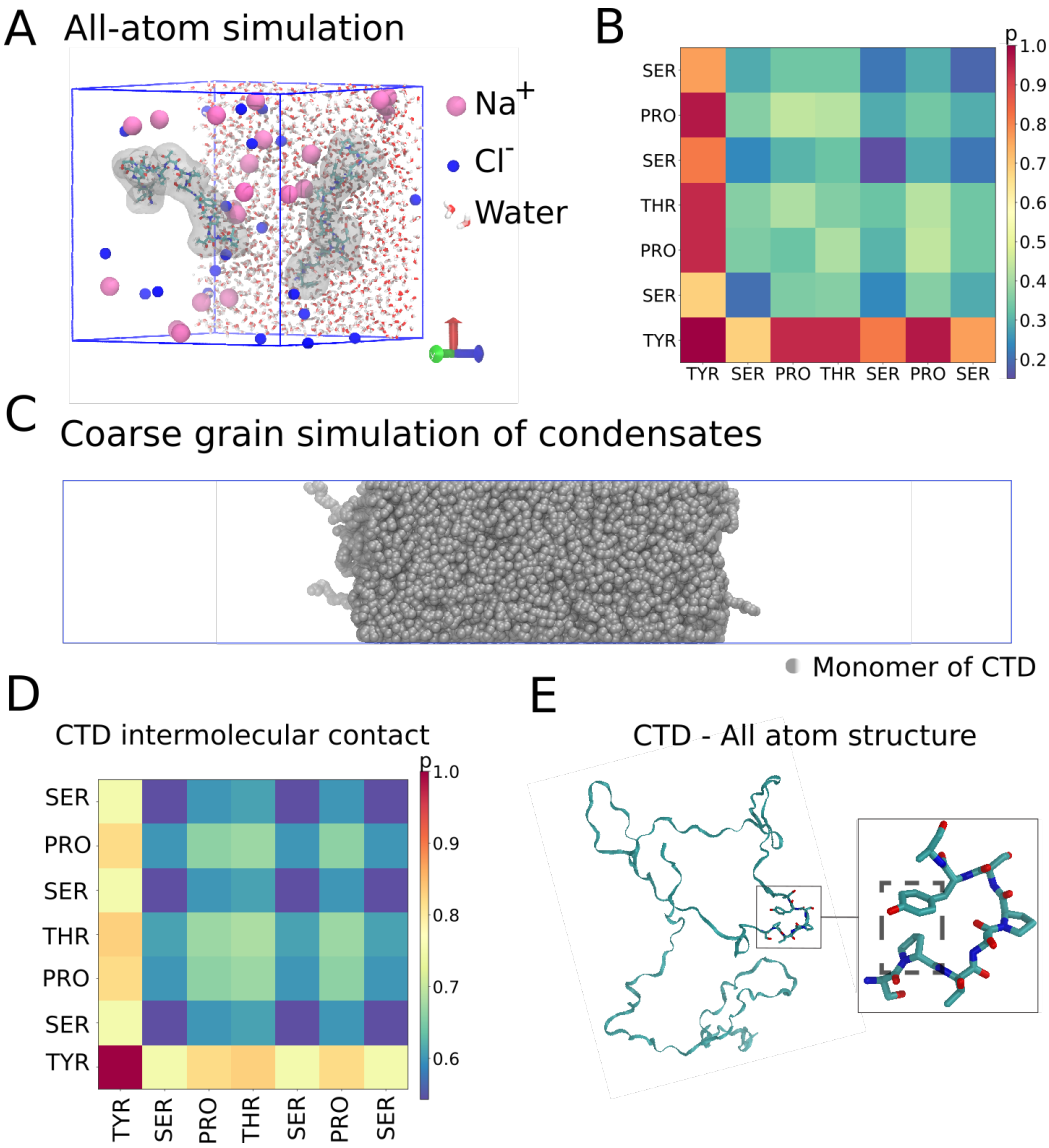

Fig. 2.2: A) All-atom simulation capturing the behaviour of two 21mer CTD peptides. B) Contact map for two all-atom CTD chains showing Tyr-Tyr and Tyr-Pro interactions. A pair of residues is in contact if any atom is below a threshold of $2^{1/6}(\sigma_i + \sigma_j)/2$, where σ is the diameter of the residues. The simulation is $39\,\mu$s long. C) Coarse-grained(CG) simulation featuring two 21mer CTD peptides. D) CG slab simulation depicting CTD condensate. E) Interactions among CTDs in CG simulation revealed through intermolecular contact frequency. F) All-atom model of yeast CTD highlighting Tyr-Pro contacts. Adapted from Figure 1 of reference [147]

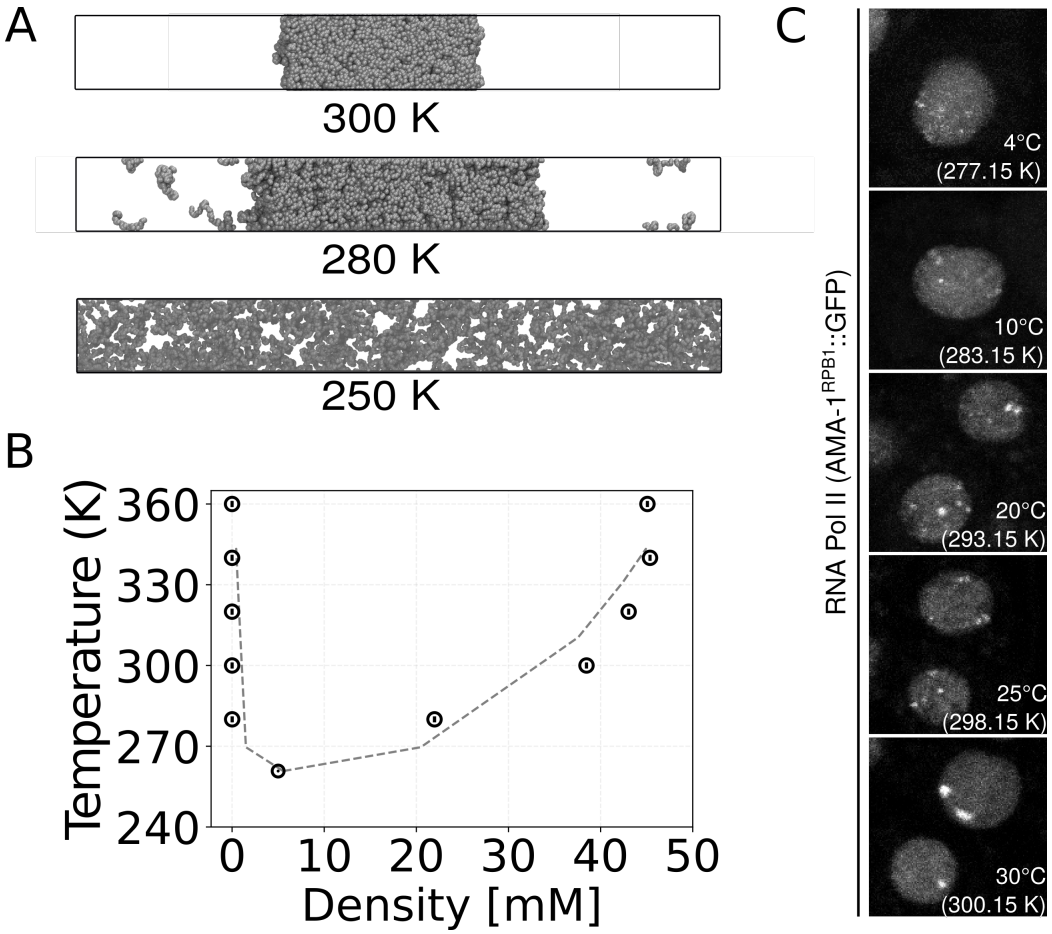

Fig. 2.3: Temperature dependent phase behaviour of RNA Pol II CTD. A) Snapshots from coarse-grained simulations at $280\,\mathrm{K}$ and $300\,\mathrm{K}$. The system phase separates at higher temperatures. B) The density profiles at different temperatures. C) Phase diagram illustrating LCST phase behaviour. The uncertainties are indicated but are smaller than the circle markers. D) RNA Pol II foci in *C. elegans* embryos at different temperatures. As the temperature rises, the number of foci decreases while their size increases. Adapted from the reference [147]

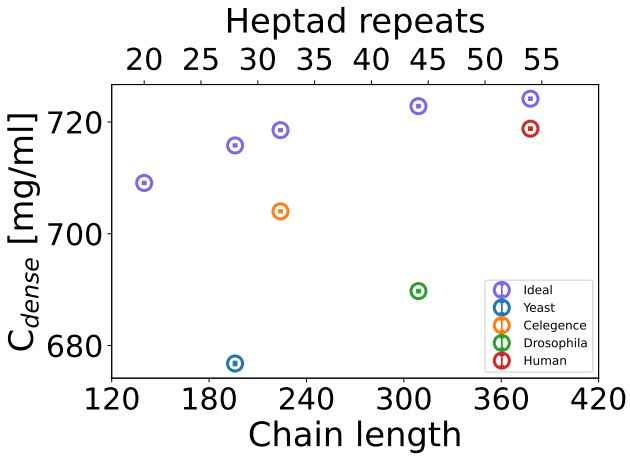

Fig. 2.4: Comparison of the dense concentration of real and ideal sequences from *Drosophilia melanogaster*, *Saccharomyces cerevisiae*, and *Homo sapiens*. The ideal sequences are less prone to phase separation compared to the real sequence. Shorter CTD sequences exhibited reduced phase separation compared to longer ones. The figure is adapted from SI of reference [147]

how deviations from the ideal heptad sequence and truncations influence the CTD's ability to phase separate.

To validate our CG simulation approach, we compared our results to experiments and atomistic reference simulations, focusing on the contact statistics of CTD at various levels: condensates, inter-chain interactions, and intra-chain interactions (Fig 2.2). In our coarse-grained simulations, we found that the CTD (140-mer with 20 ideal heptad repeats) can phase separate on its own. We observed that contacts between aromatic tyrosine (Tyr) residues within the CTD were particularly significant, as shown in our contact analysis (Fig 2.2 B,D). These results support previous experimental findings that highlight the important role of hydrophobic interactions in driving phase separation[163, 164]. Additionally, we noted that Tyr-Tyr contacts were prominent in the interactions of two 21-mer CTD fragments during extensive atomistic reference simulations (100 $\times$ 0.12us) (Fig. 2.2 B). By comparing these atomistic simulations with our coarse-grained simulations, we demonstrated that our approach effectively captures CTD interactions not only qualitatively but also quantitatively. Our contact analysis also highlighted Tyr-Pro interactions (Fig 2.2 B), which have been recently emphasized by nuclear magnetic resonance (NMR) experiments and atomistic modelling. As discussed in Chapter 2, this study generated all-atom structures of yeast CTD that align with experimental data (Fig 2.2E) [164]. Finally, we compare simulations to experiments with CTD truncation, which shows that our modelling captures functional relevant trends in the behaviour of CTD. As anticipated, shorter CTD sequences exhibited less phase separation than longer ones[163] (Fig 2.4). Additionally, we observed that the actual sequences from *Drosophilia melanogaster*, *C. elegans*, *Saccharomyces cerevisiae*, and *Homo sapiens*, which all deviate from the heptad repeat pattern $\mathrm{Y_1S_2P_3T_4S_5P_6S_7}$, are less prone to

phase separate than sequences of equal length consisting only of ideal heptad repeats. In line with experiments, a truncated version of *Drosophilia* (20 repeats) CTD[165], which features only idealised CTD repeats phase separates, about as much as full-length *Drosophilia* real CTD, whereas the full length ideal *Drosophilia* CTD showed a considerably higher tendency to phase separate. [...]

We investigated the temperature-dependent phase behaviour of CTD and found that CTD condensates dissolve at lower temperatures but remain stable at room temperature. To understand this phenomenon better, we compared our predictions about how temperature affects condensate behaviour with observations of RNA Pol II condensates in *C. elegans* embryos exposed to different temperatures (Fig 2.3). To quantify these condensates in living organisms, Dr. Jan Padeken, our experimental collaborator from IMB Mainz, imaged embryos expressing the main RNA Pol II subunit, RPB1 (known as ama-1 in *C. elegans*), tagged with GFP (Fig 2.3 C). Interestingly, our findings revealed that, unlike most other condensates that dissolve at higher temperatures and exhibit an upper critical solution temperature (UCST), both our simulation model and the in vivo observations indicated a lower critical solution temperature (LCST). This means that at lower temperatures, the formation of condensates is reduced, while higher temperatures promote their formation. This LCST behaviour aligns with recent studies showing similar results for RNA Pol II CTD in vitro [164], reinforcing the consistency between our in vivo experiments and simulations. LCST behaviour has been well-studied in synthetic polymers and, more recently in disordered proteins[166, 167, 168, 169, 170]. Many proline-rich protein sequences exhibit this behaviour because water molecules are highly organized around proline residues[171], creating a structured water network that incurs an entropic cost. At higher temperatures, this water network becomes less favourable, allowing proline to interact more readily with other residues. This shift favors compact chain conformations and promotes the formation of protein condensates. In contrast, other nuclear condensates, such as those formed by the histone methyltransferase MET-2[172] and HP1α[173], dissolve as temperatures rise. These differences suggest that multiple distinct condensates coexist within the nucleus, each responding uniquely to temperature changes. [...]

The Influence of Phosphorylation Patterns on RNA Polymerase II CTD Condensates

In addition to the separation of active from inactive regions in the genome, transcription itself occurs in multiple defined steps. The post-transcriptional phosphorylation state of RNA Polymerase II CTD is characteristic of each of these transcriptional stages[131, 133] and has been implicated in condensate formation. Here, we ask how the degree of CTD phosphorylation and the distribution of phosphosites impact condensate formation. To specifically examine how phosphorylation influences phase separation, we conducted simulations using CTD and pCTD chains, each comprising 140-mers (20 repeats of the ideal heptad

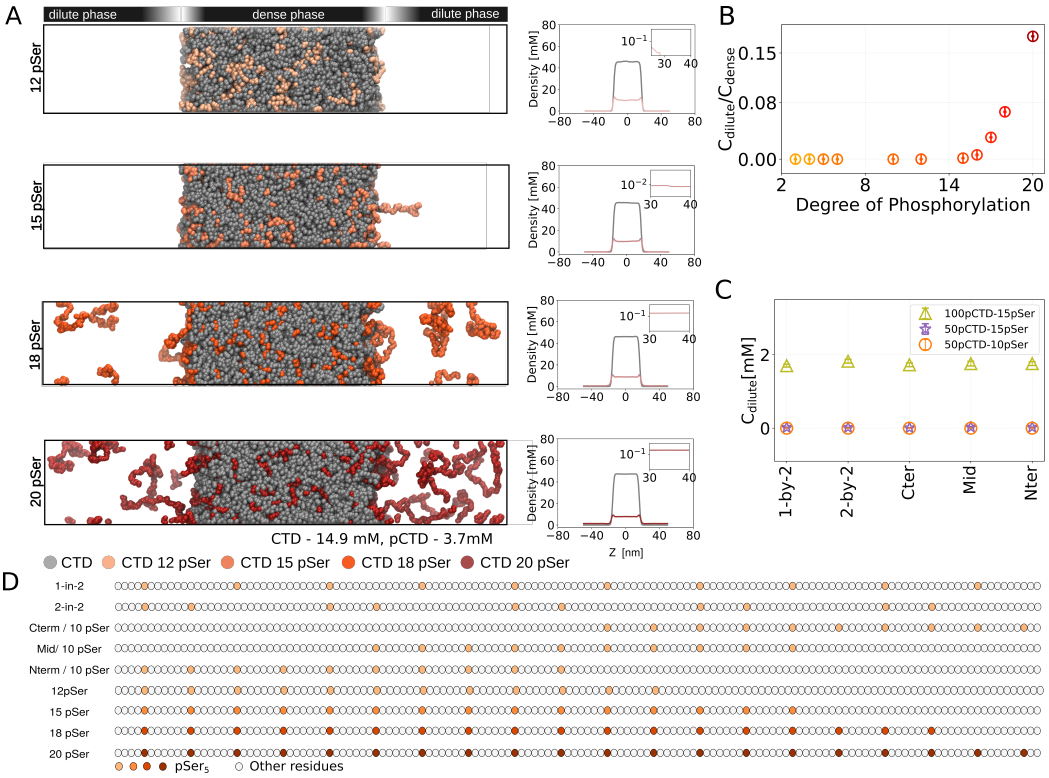

Fig. 2.5: Phase behavior of 100 CTD and 50 pCTD chains. Visualization of CTD and pCTD in molecular dynamics simulations in slab geometry of box dimension $15\,\mathrm{nm}\times15\,\mathrm{nm}\times100\,\mathrm{nm}$. A) The degree of phosphorylation of pCTD chains varies in each simulation. Density plots for simulation with CTD and pCTD with 10pSer, 12pSer, 18pSer, and 20pSer, respectively. B) Quantifying c_{dilute}/c_{dense} of partially phosphorylated pCTD chains as the Ser_5 phosphorylation increases. C) Variation in c_{dilute} of pCTD with pSer residues distributed in five distinct patterns. D) Distribution of $pSer_5$ residues in the 140mer pCTD. Adapted from SI ref [147].

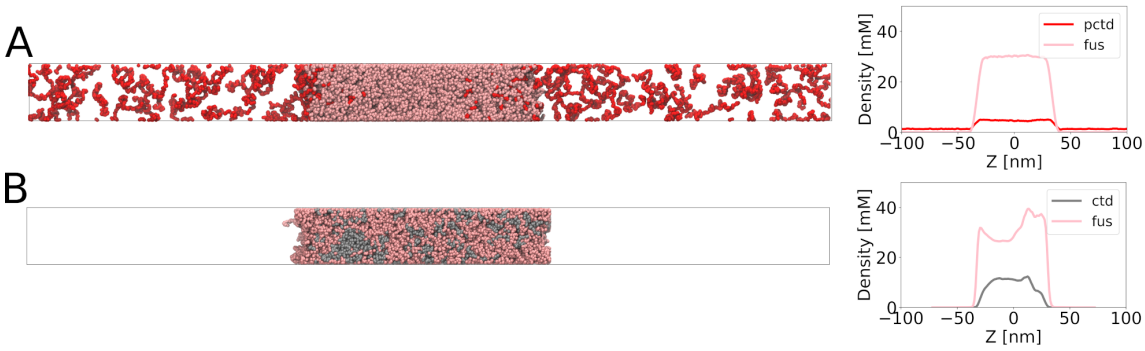

Fig. 2.6: FUS (pink) interaction with A) pCTD (red) and B) CTD (grey). FUS exhibits stronger interaction with CTD compared to pCTD. Recruitment of CTD to the FUS condensates is evident for unphosphorylated CTD, whereas the interaction strength is reduced with pCTD. Adapted from SI of ref [147].

repeats)(Fig 2.5). We incrementally phosphorylated 50 CTDs in our simulations and in accordance with experiments, we find that phosphorylation dissolves CTD condensates in a dose-dependent manner. Phosphorylating up to 15 Ser_5 in the heptad motif out of a total of 20 Ser5 in the 140-mer pCTD still predominantly leads to partitioning into the dense phase of CTD. However, as the phosphorylation level of pCTD increases above the threshold of 15 Ser_5, the dilute concentration of pCTD gradually rises as the electrostatic repulsion between pCTDs in the dense phase dominates and expels the pCTD chains into the dilute phase (Fig 2.5 B). The contact map illustrating pCTD-CTD interactions (Fig 2.5 B) highlights the relatively weak nature of the intermolecular binding between the CTD and pCTD when all 20 Ser_5 residues are phosphorylated, as compared to the intermolecular interactions observed between CTD-CTD. pCTD with each heptad repeat phosphorylated does not phase separate on its own (Fig 2.7B).

To explore the impact of different patterns of distributing an identical number of phosphorylated Ser_5 residues on CTD sequences, we conducted simulations with partially phosphorylated pCTD chains, each carrying different distribution of pSer, in the presence of a CTD condensate and compared the variation in dilute concentration of pCTD (Fig 2.5 C). We considered five distinct distributions: maximally distributed in every second heptad repeat, in adjacent heptad repeats, in the middle, and in N- and C-termini, respectively. For pCTD chains with 10 phosphorylated sites (below the threshold), rarely were any pCTD chains found outside the CTD condensate, regardless of their distribution. With 15 pSer (above the threshold), the dilute concentration remained nearly the same for different distributions. In conclusion, our simulations show a phosphorylation-dependent demixing of unphosphorylated and phosphorylated CTD. The degree of demixing is thereby dependent on the number of phosphorylation sites but not their distribution along the CTD heptad repeats.

Thus, our simulations can explain why in vitro CTD condensates are dissolved over time in the presence of the kinase CDK7 and ATP[163]. However, this also poses a clear contradiction

with all observations made in cells[174, 175], where transcribing RNA Polymerase CTD localises into condensates even when it is highly phosphorylated[176]. A likely explanation for this discrepancy is that in cells, the phosphorylation of CTD can create binding sites for multiple other interacting proteins[174]. We, therefore, wondered if we could simulate the impact of additional factors in our model. As a first indicator, we consider the interactions of CTD and pCTD with FUS condensates. In our simulation, the unphosphorylated CTD binds more strongly to FUS low complexity domain (LCD) condensates than pCTD (Fig 2.6), which is in line with experiments that demonstrated that CTD phosphorylation abrogates the binding of CTD to FUS gels and condensates. Interestingly, while FUS and CTD both localise to a single dense phase, they are not homogenously mixed in the condensate. We further observed that FUS results in a minor enrichment of pCTD in the dense phase. Together, this shows that we can indeed expand our simulation to encompass CTD interaction partners.

Co-phase separation of pCTD and HRD by complex co-acervation

This section examines the role of Cyclin T1 in transcriptional elongation, specifically its interaction with the C-terminal domain (CTD) of RNA Polymerase II. Cyclin T1, a component of the positive transcription elongation factor b (P-TEFb), is crucial for stabilizing elongating RNA Polymerase II and facilitating pCTD condensation both in vitro and potentially in vivo. Simulations corroborate experimental findings [174], demonstrating that phosphorylated CTD (pCTD) co-phase separates with the histidine-rich domain (HRD) of Cyclin T1.

Interestingly, while HRD, which is positively charged due to its histidine side chains, does not interact strongly with unphosphorylated CTD, it forms strong electrostatic interactions with pCTD (Fig 2.7). pSer residues in pCTD interact effectively with positively charged residues (Arg, Lys, and His) in HRD, while tyrosine (Tyr) residues in pCTD make strong contacts with several HRD residues, particularly Lys and Arg (Fig 2.7D). To better understand these interactions, a phase diagram was plotted. This diagram shows that pCTD and HRD form complex coacervates through electrostatic interactions, illustrating a type of phase separation known as cross-interaction-driven phase separation (CIP) [177]. CIP occurs when two species are mutually attracted to each other and cannot phase separate independently, typically resulting in a looped phase diagram as the binodal lines do not intersect with the composition axes. Throughout these simulations, the concentration of HRD remained constant at 3.9 mM while the concentrations of pCTD were systematically increased (Fig 2.7E-H). The relative concentrations of the proteins are critical in determining the extent of phase separation and the location of the coexistence region. Notably, an increase in the concentration of either protein can initiate phase separation, leading to condensate formation.

This phenomenon is illustrated in Fig 2.7I, which shows that at an approximate concentration of 5 mM pCTD, all HRDs in the dilute phase undergo phase separation. Further increasing the pCTD concentration does not affect the dense phase; instead, the dilute concentration of

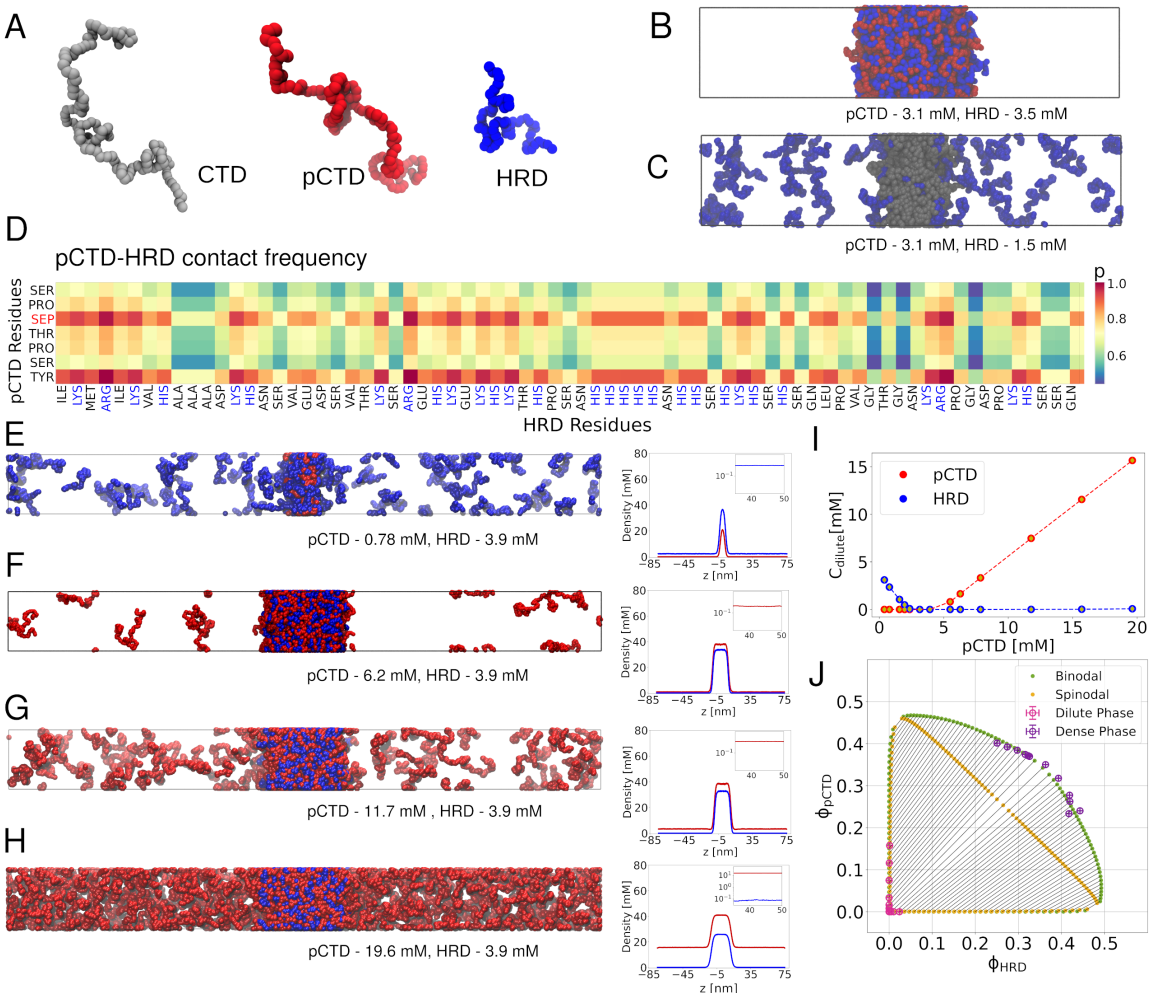

Fig. 2.7: Co-phase separation HRD with CTD and pCTD in slab simulation. A) Single chains of CTD, pCTD, and HRD. B) Simulation snapshot showing pCTD and HRD forming condensate. C) Simulation snapshot of CTD with HRD. While CTD forms a condensate, HRD exhibits no favourable interaction with CTD. D) Contact frequency analysis between pCTD and HRD showing the important electrostatic interactions between positive (blue) and negative (red) charged residues. E-H) Simulation snapshots depicting the effect of increasing pCTD while maintaining a constant concentration of HRD in a slab simulation. Density profiles for pCTD and HRD correspond to the simulations shown on the right side. I) Variation of dilute concentration of pCTD and HRD as the pCTD concentration increases. J) Biphasic diagram of the pCTD-HRD system. Positive tie lines indicate associative phase separation. The volume fractions from molecular dynamic simulations are in magenta and purple. Binodal and spinodal from mean field fit are shown in green and yellow dots. Adapted from ref [147]

pCTD increases (Fig 2.7F-H). The saturation concentration of HRD gradually decreases with increasing pCTD concentration, as HRDs initially present in the dilute phase transition into the dense phase by associating with added pCTD chains. These findings provide a foundation for further analysis. To gain deeper insights into these interactions and quantitatively understand their implications, a mean-field theory analysis was conducted.

Mean-field theory analysis of phase behaviour of pCTD and HRD

Building upon the observations regarding pCTD and HRD co-phase separation, the mean-field theory was applied to analyze their phase behaviour. This approach helps clarify the underlying physical principles driving phase separation. The phase behaviour of binary solutions of biopolymers can often be described using mean-field theory, regardless of the exact nature of the underlying interaction pattern [178]. The ternary phase diagram depicted in Fig 2.7J shows simulated coexisting (binodal) dilute and dense phases represented by pink and purple dots, respectively. The phase separation is associative, evidenced by the positive tilt in tie-lines connecting coexisting phases in composition space. This is further supported by fitting simulated binodal compositions using Flory-Huggins theory (Eq. 2.1), shown by green dots.

The simulated binodal compositions were fitted using Flory-Huggins theory to provide the dimensionless Helmholtz free energy density for the ternary mixture (Eq. 4 in the main text [159]). An excellent fit with the simulated data was achieved despite Flory-Huggins theory not accounting for long-range electrostatic interactions. This suggests significant screening by implicit background salt in simulations, making attractive forces between polymers effectively short-ranged.

The interaction parameters obtained from fitting are: $\chi_{AB} = -2.36$, $\chi_{AS} = 0.55$, and $\chi_{BS} = 0.36$. The highly negative polymer-polymer interaction parameter captures the energy balance associated with the electrostatic attraction between HRD and pCTD, combined with repulsion between like species. Both polymer-solvent interaction parameters are below their critical values ($\chi_c = \frac{1}{2}(\frac{1}{\sqrt{N_{i=A,B}}} + 1)^2$ = 0.63 for HRD and 0.59 for pCTD), which is necessary for a looped miscibility gap. The value of χ_{AS} lies closer to its critical value than χ_{BS}, suggesting that pCTD is somewhat better accommodated by the salt solution than HRD. The fit strengthens interpretations that self-interactions between hyper-phosphorylated CTD and HRD are negligible due to electrostatic repulsion. This observation aligns with findings that the dilute phase lacks polymers of one type but is enriched by polymers of another type as their average volume fraction increases(Fig 2.7J). This study highlighted how pCTD and HRD co-phase separate through electrostatic interactions, forming complex coacervates under varying concentrations. It was established that the phase behavior of these proteins is influenced by their relative concentrations and the presence of electrostatic forces, which can lead to distinct environments within the nucleus.

Co-phase separation of CTD, pCTD, and HRD

The previous section described that unlike CTD condensates, where Tyr-Tyr and Tyr-Pro contacts drive phase separation, pCTD and HRD form co-acervates through electrostatic interactions. This suggests that due to differences in molecular driving forces, CTD and pCTD-HRD phases could potentially constitute distinct environments within the nucleus. This section builds on these findings by examining how increased pCTD phosphorylation affects phase separation dynamics with CTD and HRD. The focus is on how varying phosphorylation levels influence interactions between CTD, pCTD, and HRD, and whether they form a single mixed phase or segregate into distinct dense phases. In separate simulations, increased phosphorylation disrupts the phase separation of pCTD. However, the binding of HRD to pCTD facilitates its phase separation. Yet, within cells, CTD, pCTD, and HRD coexist concurrently, raising the question of whether they form a single mixed phase or segregate into two distinct dense phases. To address this, fixed numbers of CTD, pCTD, and HRD were simulated while increasing pCTD phosphorylation. The results show that partially phosphorylated pCTD (4 $pSer_5$) co-phase separates with CTD, forming a mixed condensate(Fig 2.8A). As phosphorylation increases (8 $pSer_5$), more HRD is recruited to the dense phase by pCTDs, with pCTD and HRD tending to co-localize (Fig 2.8B). Further phosphorylation (10 $pSer_5$) enhances this co-localization and reduces HRD in the dilute phase(Fig 2.8C).

The co-localization of pCTD-HRD is observed to favor the periphery of the condensate border (Fig 2.8B-E). This preference for condensate periphery localization could be attributed to the presence of negative charges on pCTD and positive charges on HRD, potentially facilitating interactions with the solvent compared to the uncharged CTD chains. Earlier simulations employing the same coarse-grained approach have also shown microstructuring in condensates due to electrostatic interactions[161]. With further increasing phosphorylation of pCTD, there is an enhanced tendency for both pCTD and HRD to localize towards the boundaries of the CTD-pCTD-HRD condensate (Fig 2.8C-D). Unphosphorylated CTD is clearly at the centre of the condensate. Adding more phosphates to pCTD results in a core-shell morphology, with the CTD density mostly at the centre of the condensate, with steep decreases in CTD density towards the boundaries. De-mixing is not complete, since some CTD chains co-localize with pCTD-HRD as is apparent from both the visualisation of the simulation and the density profile. Phosphorylating all Ser_5 residues in the 140mer pCTD chains (Fig 2.8E), further enhances the de-mixing.

The size of the largest CTD cluster in simulation shows a sharp transition when about half of the Ser_5 of pCTD chains are phosphorylated, indicating the onset of the separation process into two distinct phases: CTD and pCTD-HRD (Fig 2.8F). The microphase separated domains of the homopolymer CTD phase and the heteropolymer pCTD-HRD phase begin to evolve as the degree of phosphorylation increases. This change becomes more pronounced with increasing phosphorylation, leading to an expansion in the size of these domains. The excess of CTD in proximity to CTD and pCTD (n^{exc}) also shows a similar transition as

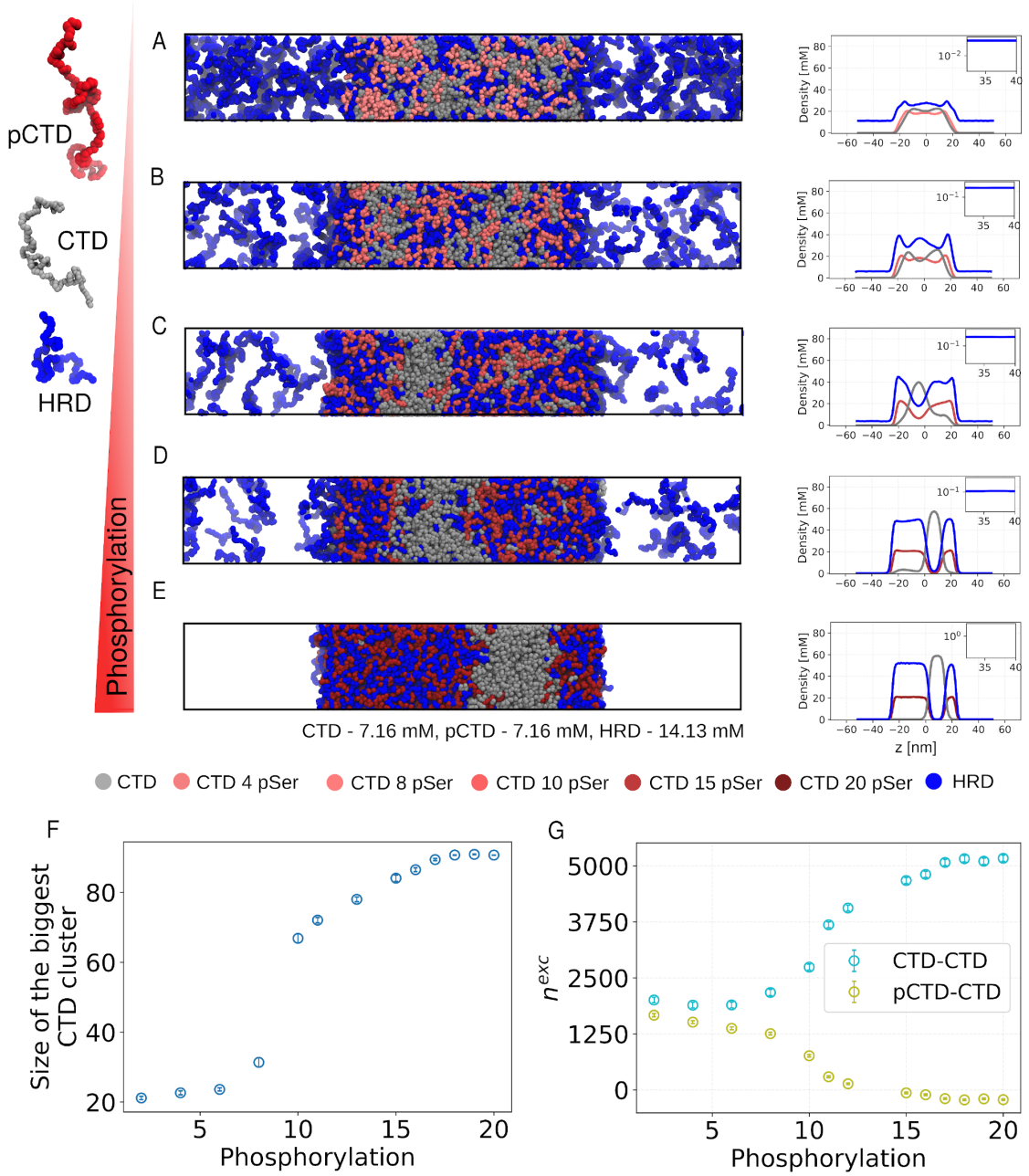

Fig. 2.8: Mixing and de-mixing of the dense phase of 100 CTD, 100 pCTD, and 250 HRD chains, with increasing the degree of pCTD phosphorylation. A) 4pSer B) 8pSer C) 10pSer D) 15pSer E) 20 pSer of pCTD are phosphorylated. F) The size of the biggest CTD cluster as the phosphorylation of pCTD varies in the simulation. G) Comparison of the excess of CTD monomers surrounding a phosphorylated pCTD (CTD-pCTD) and surrounding another CTD (CTD-CTD), with increasing phosphorylation levels, denoted as n^{exc}. This metric quantifies the number of protein-protein contacts and is calculated using Equation 2.9. Adapted from ref [147]

phosphorylation increases (Fig 2.8G). The parameter n^{exc} reveals how the number of nearby pCTD and CTD chains within a specific distance from a CTD chain changes with increasing phosphorylation of pCTD. It shows changes in local environments, providing insights into interaction frequency and strength between CTD and pCTD chains as phosphorylation varies. At lower phosphorylation levels, a CTD chain is surrounded by both pCTD and CTD, with consistent numbers for both types. However, noticeable deviations occur as phosphorylation exceeds 10 $pSer_5$. The findings from this study elucidate the intricate dynamics of phase separation among CTD, pCTD, and HRD, particularly highlighting the significant influence of pCTD phosphorylation. As phosphorylation levels increase, a transition from a mixed phase to distinct core-shell morphologies is observed, with unphosphorylated CTD central and phosphorylated pCTD and HRD localising to the periphery. [...]

This leads us to the next section, where we will examine the core-shell arrangement of these condensates, exploring how varying interfacial tensions influence their morphological characteristics and the implications for protein interactions and stability.

Morphology of CTD and pCTD condensates in simulations, in vitro, and *C. elegans* embryos

To further understand the implications of the phase separation dynamics observed earlier, we now examine the core-shell arrangement of the condensates formed by CTD, pCTD, and HRD. In this section, the potential morphologies of these condensates will be explored, specifically how interfacial tensions influence their stability and interactions within cellular environments.

Our molecular dynamics simulations and calculations of interfacial tensions [179, 140] provide insight into the potential morphologies of condensates formed by CTD and pCTD-HRD. When these two immiscible droplets come into contact, four distinct morphological configurations can emerge[180, 181]. These configurations are visually represented using a stability diagram generated from the contact angle equation, as depicted in Fig 2.9D. The contact angle equation is given by:

$$\theta_{\text{[pCTD-HRD]}} = \arccos\left(\frac{\gamma^2_{\text{[CTD]}} - \gamma^2_{\text{[pCTD-HRD]}} - \gamma^2_{\text{[pCTD-HRD]CTD}}}{2\gamma_{\text{[pCTD-HRD]}}\gamma_{\text{[pCTD-HRD]CTD}}}\right) \tag{2.10}$$

.

In Figure 2.9D, a schematic depicts partially engulfing CTD and pCTD-HRD phases, illustrating the contact angles between these phases. The stability diagram describes the four different morphologies attainable by adjusting relative interfacial tensions, and these variations can be characterised through the corresponding contact angles. In Region I, the CTD phase is entirely enveloped by the pCTD-HRD phase, with the condition $\gamma_{\text{[CTD]}} > \gamma_{\text{[pCTD-HRD]}} + \gamma_{\text{[pCTD-HRD]CTD}}$.

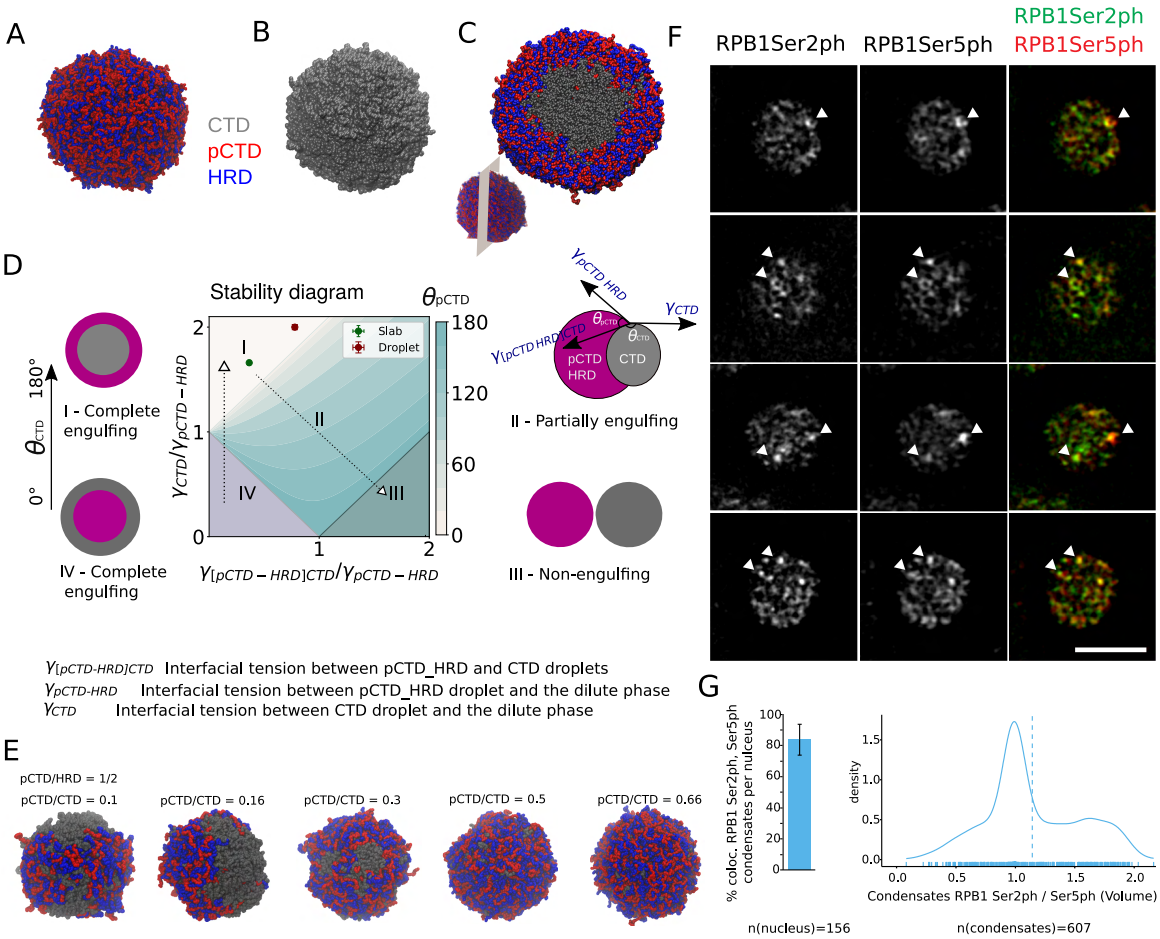

Fig. 2.9: Morphology of the multiphasic condensate A) Formation of a droplet composed of 900 pCTD and 1800 HRD chains. B) Droplet comprising 600 CTD chains. C) Multiphasic droplet with a core-shell morphology, constituted by 900 pCTD, 1800 HRD, and 600 CTD chains. D) Stability diagram illustrating four distinct morphologies arising from the interaction between CTD and pCTD-HRD droplets. Interfacial tension analysis from slab and droplet simulations indicates the core-shell morphology. Sketch illustrating a partially engulfing CTD and pCTD-HRD phases indicating the contact angles. E) CTD condensate partially enveloped by pCTD and HRD condensate. Sequential images depict the progression from partially engulfed to fully engulfed morphology as the ratio of pCTD/CTD chains increases. F) Super-resolution microscopy image of *C. elegans* embryos stained with antibodies targeting phospho-Serine 5 (pSer5, indicating transcriptional initiation) and phospho-Ser2 (pSer2, indicating transcriptional elongation) phosphorylated RPB1. RPB1 labeled with either PTM is localized in prominent subnuclear foci. G)The segmentation and quantification of these foci demonstrate that both types of foci colocalize, exhibiting only partial overlap. Adapted from ref [147]

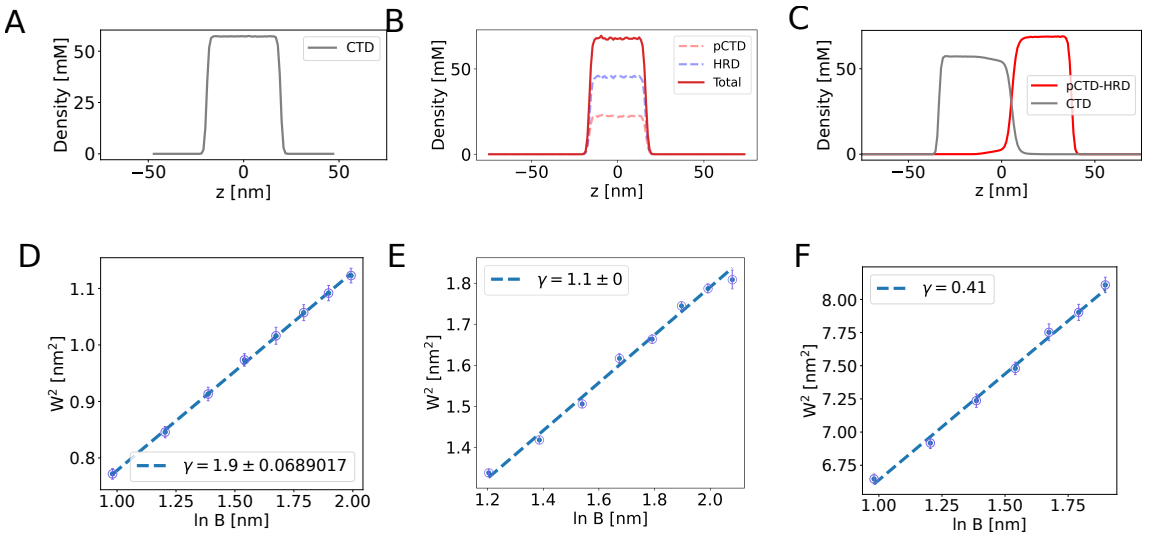

Fig. 2.10: Analysis of interfacial tension from slab simulations using Capillary Wave Theory (CWT) with the block analysis method. Panels A-C display density profiles for CTD, pCTD-HRD, and CTD-pCTD-HRD systems. Panels D-F illustrate the apparent interfacial width W^2 with block size B. The interface region is divided into rectangular segments. Interfacial tension is determined using the equation: $W_0^2 + \frac{K_B T}{2\pi\gamma} \ln\left(\frac{L}{B}\right)$. Adapted from SI of ref [147]

In Region II, the CTD and pCTD-HRD phases are partially engulfed, sharing an interface while also being exposed to the dilute phase. The transition from complete engulfment to non-engulfing phases is gradual as $\theta_{[\text{pCTD-HRD}]}$ increases. Here, the three interfacial tensions are comparable: $\gamma_{[\text{CTD}]} \sim \gamma_{[\text{pCTD-HRD}]} \sim \gamma_{[\text{pCTD-HRD}]\text{CTD}}$. In Region III, the significant interfacial tension between the CTD and pCTD-HRD droplets leads to a non-engulfing morphology, defined by $\gamma_{[\text{pCTD-HRD}]\text{CTD}} > \gamma_{[\text{CTD}]} + \gamma_{[\text{pCTD-HRD}]}$. In Region IV, the CTD phase engulfs the pCTD-HRD phase. The transition from Region 4 to Region I is marked by a black arrow, signifying an increase in $\theta_{[\text{CTD}]}$.

Full engulfment of CTD but pCTD-HRD is the most favourable morphology, but small changes in the system would enable partial engulfment. Capillary wave theory (CWT)[179] was utilized to compute interfacial tensions of CTD and pCTD-HRD using slab geometry with each Ser_5 in the heptad repeats of pCTD phosphorylated. Three sets of slab simulations were performed to determine interfacial tension at different interfaces (Fig. 2.10). The first two simulations involved separate simulations of CTD and pCTD-HRD forming their respective condensates, allowing for the calculation of interfacial tension at the interface with the dilute phase. In the third simulation, both CTD and pCTD-HRD were simulated together to form a multiphasic condensate, enabling the computation of interfacial tension at their interface. Evaluating the ratio of interfacial tension for different interfaces, specifically $\gamma_{[\text{CTD}]}/\gamma_{[\text{pCTD-HRD}]}$ and $\gamma_{[\text{pCTD-HRD}]\text{CTD}}/\gamma_{[\text{pCTD-HRD}]}$, confirmed that the observed core-shell morphologies are consistent with the theory of liquid interfaces (depicted by the green dot in Fig 2.9E). This supports the conclusion that core-shell morphology is a stable state in simulations and not an artifact of insufficiently long simulations.

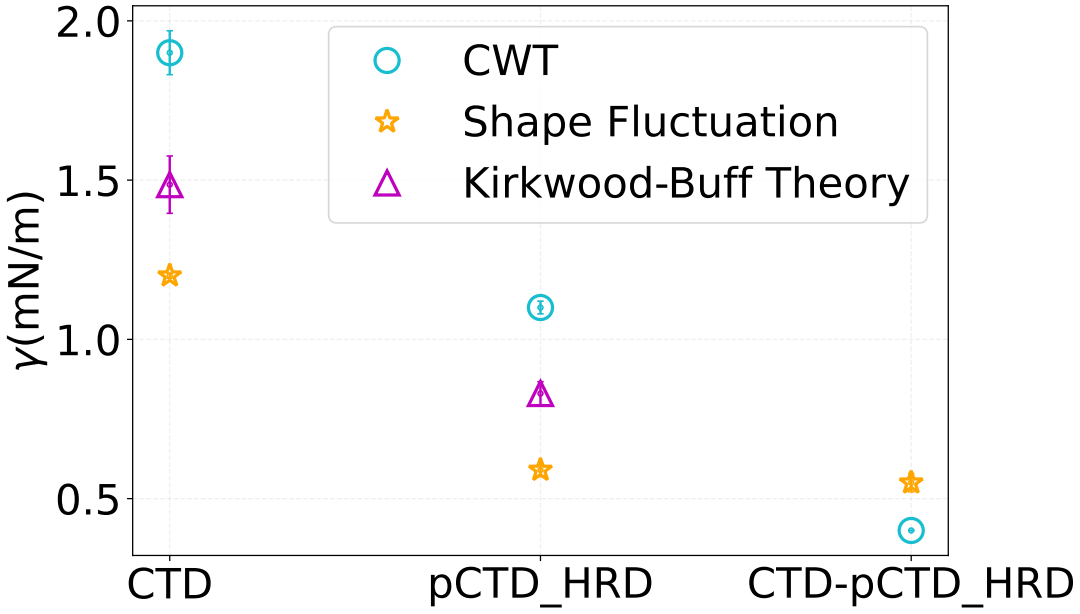

Fig. 2.11: Comparison of interfacial tension obtained from different methods: Capillary Wave Theory (CWT) with interfacial width calculation, Shape Fluctuation of Droplet, and Kirkwood-Buff Theory. Comparison among CTD, pCTD-HRD condensates, and the interface between CTD and pCTD-HRD. Adapted from SI of ref [147]

To rule out effects from box geometry, simulations were repeated in larger cubic boxes measuring $150\,\mathrm{nm}\times150\,\mathrm{nm}\times150\,\mathrm{nm}$ protein chains forming spherical droplets (Fig 2.9A-C. The ratios of interfacial tension between different interfaces, $\gamma_{[\mathrm{CTD}]}/\gamma_{[\mathrm{pCTD\text{-}HRD}]}$ and $\gamma_{[\mathrm{pCTD\text{-}HRD]CTD}}/\gamma_{[\mathrm{pCTD\text{-}HRD}]}$, were computed from the shape fluctuations method [140] (indicated by the red dot in Fig 2.9D). These results further confirm that core-shell morphology is indeed favorable for the CTD-pCTD-HRD system in simulations. A comparison of interfacial tension at different interfaces using various methods is illustrated in Fig 2.11, demonstrating consistent values across approaches.

The computed ratios of surface tensions are closer to the boundary between full and partial engulfment. Simulations also reveal partial engulfment of CTD when there are insufficient pCTD and HRD chains to fully cover its surface (Fig 2.9E). Simulations clearly show that the condensate of pCTD-HRD either fully or partially surrounds the CTD condensate (Fig 2.9C, E). At lower concentrations of pCTD and HRD, they undergo coacervation and form a small patch on the surface of the CTD condensate rather than spreading evenly across the surface, resulting in a partially engulfed structure. With an increase in the concentration of pCTD and HRD, the degree of engulfment also increases, indicating that higher concentrations of pCTD and HRD lead to a decrease in interfacial tension between the pCTD-HRD complex and CTD[182].

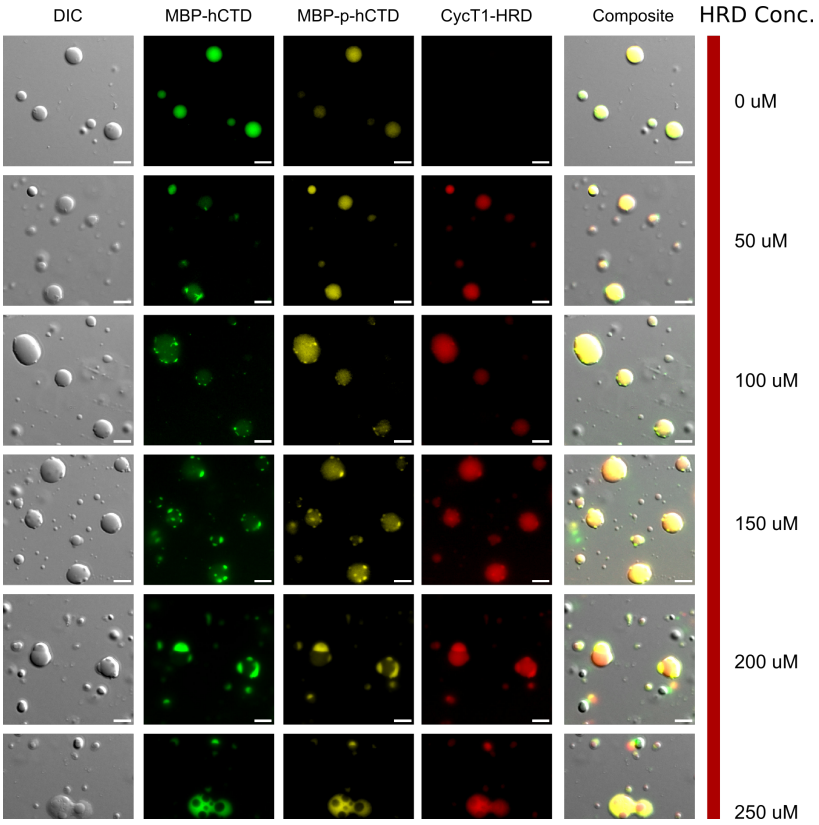

Fig. 2.12: Droplet compartments induced by HRD. Differential interference contrast (DIC) shows MBP-hCTD (AF488; green), MBP-p-hCTD$_{\mathrm{partial}}$ (AF594; yellow) and HRD (AF647; red) forming different protein condensates upon the addition of HRD. Composite pictures show droplets formed by different components in the function of HRD's concentration. Adapted from SI of ref [147]

Experimental Validation of Simulation Predictions

In Vitro Observations

Collaboration with Markus Zweckstetter's lab at the German Center for Neurodegenerative Diseases (DZNE) investigated composition-dependent partial demixing of CTD, pCTD, and HRD. CDK9 addition to human CTD produced partially phosphorylated CTD (pCTDpartial) with varying phosphorylation levels. pCTD$_{\mathrm{partial}}$ mixed well with CTD (Fig 2.12F, top row). Adding purified HRD (50 M) to 25 M mixtures of CTD and pCTDpartial created distinct CTD foci within mixed condensates. More phosphorylated pCTDpartial chains preferred to reside outside these foci, while less phosphorylated chains interacted more readily with them. CTD foci prominence increased with HRD concentration; at $300\,\mu\mathrm{m}$ and $150\,\mu\mathrm{m}$, foci became more pronounced. At 150 M, HRD was clearly present within foci alongside CTD-pCTDpartial-HRD condensates. This trend continued with increasing HRD concentration. At $200\,\mu\mathrm{m}$, the effect was more pronounced, and at $250\,\mu\mathrm{m}$, distinct HRD droplets emerged with CTDs and pCTDspartial partially covering them. Most droplets were mixtures of all three components. These observations confirm that partial demixing aligns with simulation predictions, demonstrating simulated behavior replication in purified protein experiments.

In Vivo Validation

Dr. Jan Padeken at the Institute of Molecular Biology performed super-resolution microscopy on C. elegans embryos to test RNA Polymerase II condensate segregation. Embryos were stained with antibodies against phospho-Serine 5 ($pSer_5$, transcriptional initiation) and phospho-Ser2 ($pSer_2$, transcriptional elongation) of RPB1. Large subnuclear foci marked by either post-translational modification were identified (Fig 2.9F). Quantification showed both foci types colocalize but only partially overlap (Fig 2.9G), consistent with partial or full engulfment in simulations. Some foci formed hollow spheres or half-moon-like structures, while total RPB1::GFP formed solid round foci (Fig 2E and Ref [183]). This suggests hollow spheres are part of larger RNA Pol II condensates, highlighting partial and full engulfment of CTD by pCTD. CTD condensates remained stable in vivo even when Cyclin T1 was knocked down via RNA interference, abolishing Ser_2 phosphorylation [184]. This indicates that CTD condensates can form independently of Ser_2 phosphorylation, as shown in simulations. These condensates were more regularly round compared to those formed with Cyclin T1, aligning with differences between pure CTDs and multi-component condensates. These findings demonstrate significant agreement between experimental data and simulation results regarding partial and full engulfment patterns of RNA polymerase II across different phosphorylation states, validating molecular dynamics simulation predictions.

These results emphasize how the different phases within the condensate significantly influence the shapes of CTD and pCTD chains. The unique properties of the CTD and pCTD-HRD phases suggest they may play distinct roles in regulating protein behavior and interactions in a biphasic environment. In the next section, we will examine how these conformational changes occur and their implications for molecular interactions, which are crucial for understanding the mechanisms of transcriptional regulation by RNA Polymerase II.

The conformations of the CTD and pCTD chains are modulated by the two condensates phases

This section examines how the conformations of CTD and pCTD chains are influenced by the distinct phases and interfaces within the condensates. The CTD and pCTD-HRD phases exhibit different properties, which can affect their conformations (Fig 2.13A, B). These differences may underlie unique regulatory processes mediated by the two phases in a biphasic condensate. Recent theoretical and experimental advances have demonstrated the significant impact of condensates on protein conformations, including their orientations and extensions [185, 186].

The conformation of a protein chain within a phase can be understood through the concept of solvent quality from polymer science [187, 161]. In a “good” solvent, compatible with the chain, the chain remains extended, regardless of whether it is in a dilute solution or a dense

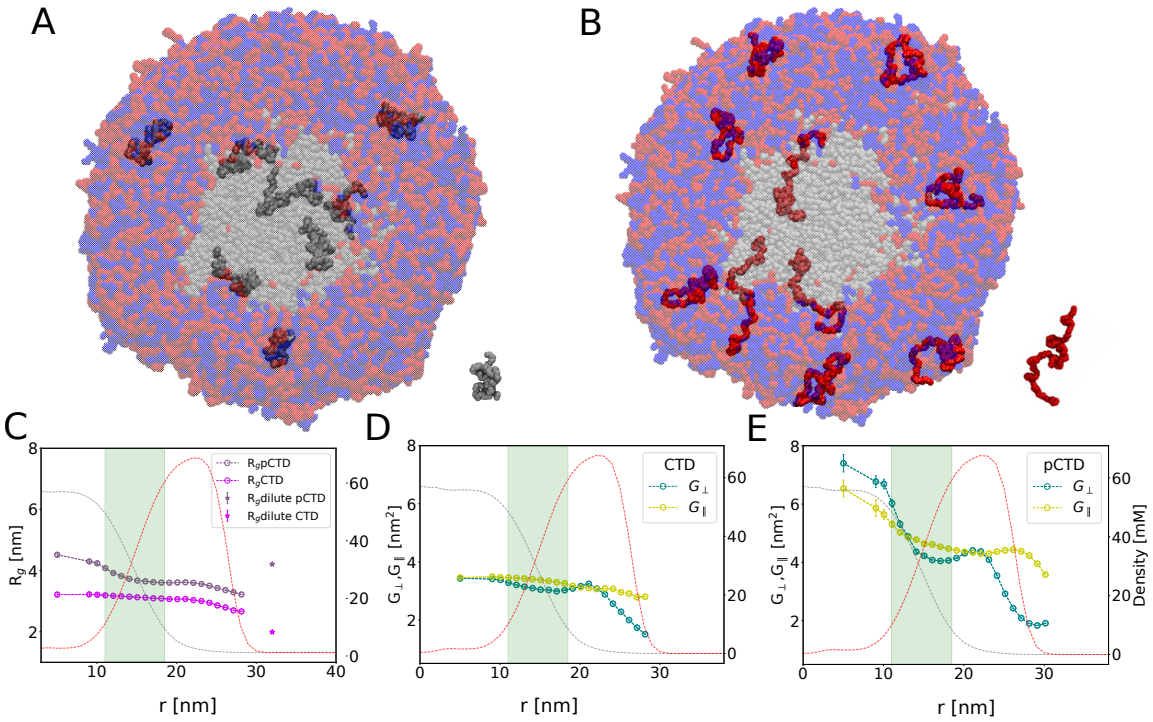

Fig. 2.13: Conformation of CTD and pCTD within the droplet. A) Snapshot capturing CTD conformations, where the R_G of CTD chains reduces slightly in the pCTD-HRD phase. In the dilute phase, the chains become more compact. B) Snapshot showing pCTD conformations. The chains are more elongated in the CTD phase and compact in the pCTD-HRD phase. The chains are elongated in the dilute phase. C) Comparison of the Variation in R_g (radius of gyration) between CTD and pCTD Conformations, Analysed from the Center of the Condensate. D) Changes in R_G and diagonal components of the gyration tensor for CTD chains, measured from the droplet's center. E) Variation in R_G and diagonal components of the gyration tensor for pCTD chains, measured from the droplet's center. Adapted from SI of ref [147]

phase. Conversely, in a "poor" solvent, which is incompatible with the chain, it tends to collapse. In this context, CTD is more expanded in the condensate than in a dilute aqueous solution, as indicated by its radius of gyration(R_g) (Fig 2.13C). The aqueous phase is a poor solvent for CTD chains since the ideal YSPTSPS heptads are devoid of charged groups and feature aromatic and non-polar residues.

To analyze chain conformations in the condensate, parallel and perpendicular components of the gyration tensor were computed. The components $\langle G_{\parallel} \rangle$ align parallel to the interface, while $\langle G_{\perp} \rangle$ represents the perpendicular component. In the center of the CTD phase, extended CTD chains are neither aligned parallel nor perpendicular to the condensate interface; both components of the radius of gyration tensor adopt similar values (Fig 2.13D). At the interface of the CTD phase, CTD chains tend to align parallel to the interface, with $\langle G_{\perp} \rangle$ being smaller than $\langle G_{\parallel} \rangle$ (Fig 2.13A)[188]. Occasionally, some CTD chains enter the pCTD-HRD phase; however, their R_G remains largely unchanged despite this phase being enriched with positive and negative charges, which reduces solvent quality for CTD. Near the interface between pCTD-HRD and the dilute aqueous phase, CTD chains collapse as they preferentially interact with pCTD-HRD rather than with the dilute phase. The conformation of pCTD is influenced differently by these phases. Hyperphosphorylated pCTD is more extended in dilute aqueous solutions than in condensates (Fig 2.13B, E). Phosphorylation significantly affects pCTD's conformation in solution; however, its conformation changes less in dense phases due to interactions with positively charged HRD. While CTD shows minimal conformational changes when transitioning from its parent phase to pCTD-HRD phase, pCTD expands further when moving from its parent phase into the CTD phase. Within the CTD phase, pCTD extends even more; $\langle G_{\parallel} \rangle$ decreases compared to $\langle G_{\perp} \rangle$, indicating a slight preference for alignment perpendicular to the interface. This alignment enhances interactions with positively charged HRD residues in the core. At the interface with the dilute phase, pCTD chains are squeezed; here, $\langle G_{\parallel} \rangle$ is larger than $\langle G_{\perp} \rangle$, as chains align parallel to the interface. These observations show that the distinct chemical environments of each phase significantly influence chain conformation. This implies that the different phases within condensates can recruit molecules in unique ways and play critical roles in regulatory processes. With the understanding that the CTD and pCTD-HRD phases represent distinct chemical environments, we investigated their differential interactions with other molecules, which may enable each phase to support different stages of transcription. In this section, we will examine the recruitment of FUS and partially phosphorylated CTD to multiphasic condensates and how these interactions affect transcriptional processes.

Recruitment of FUS and partially phosphorylated CTD to multiphasic condensates

After establishing that the CTD, pCTD, and HRD condensates consist of distinct chemical environments, we investigated whether these phases interact differently with other molecules,

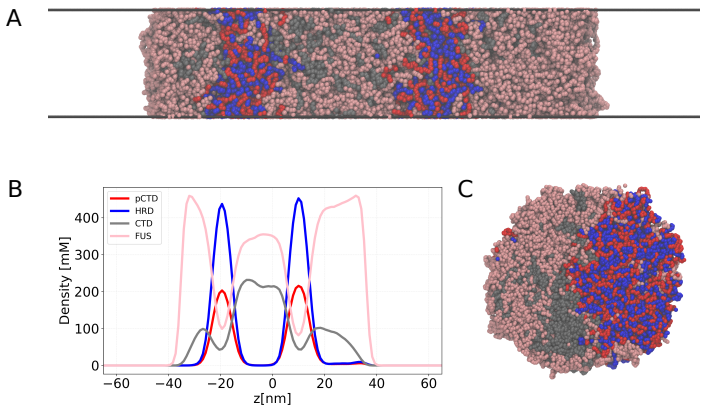

Fig. 2.14: A) Simulation of pCTD, CTD, HRD, and FUS (Pink). B) Density profile showing FUS interactions with CTD. C) Cross-section of the droplet. The droplet has partial engulfment morphology. Adapted from SI of ref [147]

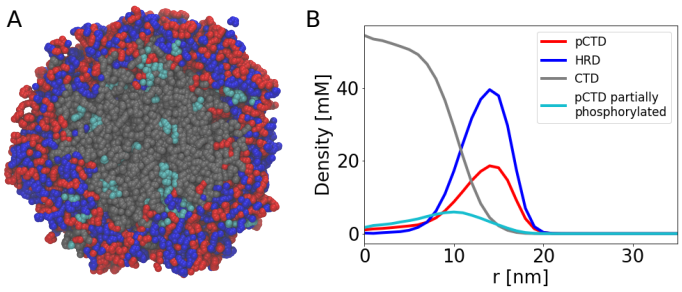

Fig. 2.15: Interactions of partially phosphorylated CTD with the CTD and pCTD-HRD phases. A) Picture shows cut through a droplet of partially phosphorylated (cyan), fully phosphorylated CTD (red), unphosphorylated CTD, and HRD from coarse-grained molecular dynamics. In the partially phosphorylated CTD, Ser_5 within the initial 10 heptad is phosphorylated. B) Density profile of the system. Adapted from SI of ref [147]

potentially enabling each phase to support various stages of transcription. Contrasting interaction patterns were observed between CTD and pCTD-HRD with phase-separated LCD FUS. Specifically, CTD interacts with phase-separated FUS, as shown in snapshots of simulation systems involving CTD, pCTD-HRD, and FUS (Fig 2.14). Regions enriched in pCTD-HRD are relatively depleted of FUS.

This preference for the CTD phase over the pCTD-HRD phase aligns with the proposed roles of CTD chains and condensates in facilitating transcription initiation by interacting with transcription factors to form promoter condensates. Experimental evidence also supports a preference for FUS LCD to interact with CTD rather than pCTD. Additionally, the partitioning of partially phosphorylated pCTD within multiphasic condensates was examined, revealing that it can interact with both the CTD and pCTD-HRD phases (Fig 2.15). Notably, partially phosphorylated CTD is relatively depleted at the peaks of the pCTD and HRD concentration profiles. This suggests that partially phosphorylated CTD can reside in the condensate phase associated with transcription initiation, while RNA Polymerase II would switch to the other phase upon further phosphorylation.

2.4 Conclusion

This chapter demonstrates that phosphorylation of the RNA Polymerase II CTD leads to the formation of distinct phase-separated condensates that regulate transcription through interactions with intrinsically disordered regions and other proteins. The findings underscore the importance of these condensates in facilitating transcription initiation and elongation, highlighting their role in fine-tuning gene expression.

Intrinsically disordered regions (IDRs) are essential for the formation of transcription condensates, as perturbations in these structures can significantly impact transcriptional activity[165, 134]. Experimental evidence suggests that transcription initiation and elongation are spatiotemporally regulated by distinct phase-separated condensates[183, 176]. Our results suggest that, in *C. elegans*, CTD phase-separated condensates facilitate chromatin target searching despite the structured and featureless nature of IDRs. Molecular dynamics simulations in this study have provided insights into the interactions between CTD, phosphorylated CTD (pCTD), and proteins such as Cyclin T1's HRD, revealing distinct phases influenced by hydrophobic and electrostatic interactions. Increased phosphorylation levels drastically alter CTD-pCTD interactions, and experimental validation through partial demixing of CTD and p-hCTD with HRD supports these simulation findings, demonstrating effective phase separation. Furthermore, thesis simulations indicate that CTD can be fully or partially engulfed by pCTD-HRD, with interfacial tensions determining favorable morphologies. This approach allows for the prediction of morphologies for various protein mixtures. The coarse-grained model employed in these simulations accurately reproduces experimental truncation and sequence effects, capturing CTD's contact statistics from both NMR and atomistic molecular

dynamics data. Super-resolution microscopy in *C. elegans* embryos further confirms the simulation results regarding CTD engulfment by pCTD. Distinct phases have been shown to recruit different proteins, with FUS exhibiting a stronger binding affinity for CTD compared to pCTD-HRD. Simulations and cellular observations suggest core-shell morphologies, where promoter condensates may be engulfed by gene-body condensates, consistent with findings related to Ser5-phosphorylated CTD. Additionally, CTD demonstrates temperature-dependent behavior, forming condensates that enhance transcription at elevated temperatures, thereby fine-tuning gene expression rather than triggering a classic heat shock response. This phase behavior is distinct from heterochromatin proteins like MET-2 and HP1α, which dissolve at higher temperatures, underscoring functional distinctions that are essential for effective transcription regulation. This chapter highlights the intricate relationship between phosphorylation, phase separation, and transcription regulation, providing a foundation for further exploration into the molecular mechanisms governing gene expression.[...]

3 Modelling Intrinsically Disordered Proteins: The Hierarchical Chain Growth Approach

3.1 Introduction

RNA Polymerase II (Pol II) plays a fundamental role in gene expression, with its carboxy-terminal domain (CTD) serving as a critical regulatory element in transcription [191, 51]. A key feature of Pol II is CTD, an intrinsically disordered region composed of heptad repeats with the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$ [192]. It is crucial for transcription initiation, elongation, and co-transcriptional RNA processing. Recent research shows that the CTD's ability to undergo phase separation is vital for forming transcriptional hubs— dynamic clusters of Pol II at active genes that enhance gene expression efficiency.

IDPs like CTD exhibit significant structural variation and are more often compact than expected for chemically denatured proteins. Unlike denatured proteins, which have lost their native structure due to factors such as external stresses, IDPs naturally lack a stable three-dimensional structure but retain functional capabilities. The structural properties of IDPs, including their degree of compaction, vary widely and are influenced by their unique sequence characteristics.

Recent research has revealed that proline content and net charge are primary determinants of compaction in IDPs, while hydrophobic contacts play a limited role. This is particularly relevant for the CTD, which is rich in proline residues [193]. The hydrodynamic radius (Rh) serves as a key measure of IDP compaction, calculated from the translational diffusion coefficient.

Experimental studies (3.1B), the interactions within the CTD were examined using nuclear magnetic resonance (NMR) experiments [190]. Two-dimensional spectra of ^{15}N-labeled yCTD in a dilute, non-phase-separated state showed low chemical shift dispersion, suggesting the absence of well-defined 3D structures such as secondary or tertiary structures. A comparison of the spectra revealed five cross-peaks with identical chemical shifts, indicating similar chemical environments. These residues corresponded to the conserved heptad repeats (Y1, S2, T4, S5, S7). Weaker peaks were attributed to residues in non-conserved repeats or due to changes in proline isomerization, implying that these residues may exhibit more flexible or dynamic behaviours compared to the conserved ones. Overall, yCTD exhibits dynamic,

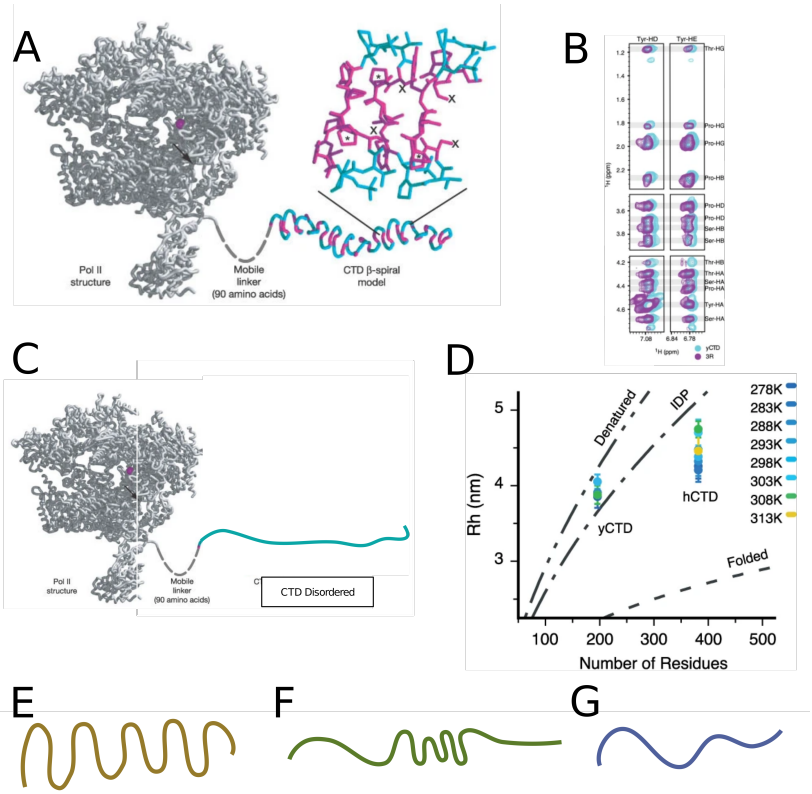

Fig. 3.1: A) CTD β-turn conformation and modeling. The 12-subunit structure of Pol II is shown in silver, with the active site indicated by a magenta sphere and an arrow for RNA exit. The CTD is represented as a β-spiral model, with β-turns (cyan) and extended regions (pink) derived from observed CTD peptide structures. The model assumes all repeats adopt this conformation, resulting in a compact left-handed β-spiral. Adapted from [189]. B) NOE contacts between Tyr ring protons and aliphatic proline protons in 2D ^{1}H-^{1}H NOESY spectra of 3R-CTD (purple) and yCTD (cyan) at 5 °C. Additional Tyr-Thr and Tyr-Ser cross-peaks were identified in the NOESY spectrum of 3R-CTD. C) CTD with flexible disordered structure. D) Comparative analysis of the hydrodynamic radii (Rh) for human CTD (hCTD) and yeast CTD (yCTD) measured across a range of temperatures under dilute conditions (25 M concentration). The plot includes theoretical curves depicting the expected Rh trends for fully denatured proteins, intrinsically disordered proteins (IDPs), and folded proteins as a function of residue count. Adapted from [190] E) Peptide exhibiting well-defined secondary structure elements F) Peptide with localized structural characteristics with regions of disorder. Exhibits a mix of structured and unstructured regions. Some segments form secondary structure elements, while others remain flexible. G) Fully disordered peptide. This peptide lacks any stable secondary structure. It exists as a highly flexible, dynamic ensemble of conformations, characteristic of IDPs or protein regions (IDRs).

predominantly random coil behaviour, with a slight tendency to form turns, particularly at the Y1 heptad position.

Complementary data in figure 3.1D shows the hydrodynamic radii (Rh) for human CTD (hCTD) and yeast CTD (yCTD) across a range of temperatures in dilute conditions (25 M concentration) [190]. The graph includes theoretical curves representing the expected Rh trends for three distinct protein categories: fully denatured, intrinsically disordered (IDP), and folded proteins, each plotted as a function of residue count. What is interesting is that R_h fall within a range intermediate to those typically observed for IDPs and denatured proteins in the yeast CTD. This suggests that the CTD has unique structural properties—it is less compact than typical IDPs but more compact than fully unfolded proteins(Fig 3.1E-G).

At the same time, earlier structural studies have presented conflicting views on the structure of CTD. Structural modelling based on conformations resolved by X-ray crystallography [189] suggest that the CTD can adopt a compact β-spiral conformation, which aids its incorporation into transcription initiation complexes (Fig 3.1A). Newer NMR data would at first glance also support well-defined and compact conformations of the CTD (Fig 3.1B). In NMR spectra, Nuclear Overhauser enhancements (NOEs) are apparent, which are observed for nuclei in close spatial proximity. However, a well-defined 3D structure would not be accepted, considering the low chemical shift dispersion in the NMR experiments. This structural ambiguity has made it challenging to understand the CTD's precise role in transcriptional regulation.

This structural ambiguity in the CTD hinders our understanding of the mechanism by which the CTD facilitates phase separation and the formation of transcriptional condensates. Clarifying whether the CTD functions primarily as an IDP (Fig 3.1B) or adopts specific conformation is crucial for understanding its biological functions. This would help elucidate transcriptional regulation and explore how variations in CTD structure impact the gene transcription and cellular processes.

To resolve this ambiguity, it is essential to integrate experimental and computational approaches that can capture the full spectrum of conformational states characteristic of IDPs. Experimental techniques, such as nuclear magnetic resonance (NMR) spectroscopy and small-angle X-ray scattering (SAXS), provide valuable insights into protein conformations and interactions in physiological contexts. However, these methods have limitations in resolving the full range of structural states. Computational modelling, in contrast, allows systematic exploration of the conformational landscape, revealing dynamic features and interactions that may be inaccessible to experimental observation.

Recent advances in ensemble generation techniques, such as hierarchical chain growth (HCG)[194] provide frameworks for exploring the conformational ensembles and interactions of IDPs like the CTD. This methods integrate experimental data to construct ensembles consistent with experimental observations while sampling the vast conformational space

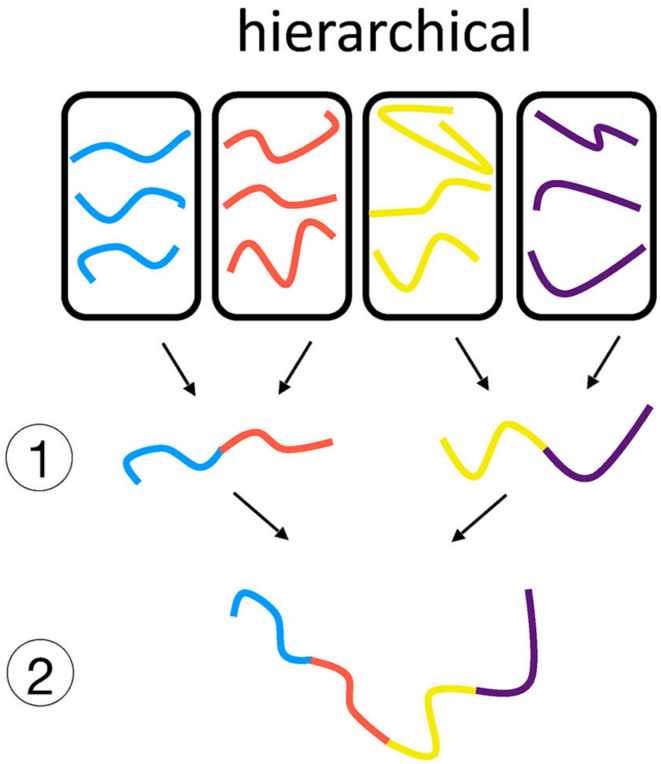

Fig. 3.2: Overview of the Hierarchical Chain Growth (HCG) Algorithm. The HCG algorithm efficiently constructs polymeric molecules by recursively combining fragments. It starts with monomers, which are combined into dimers, then dimers into tetramers, and so on, as long as there are no steric clashes. For chains made up of $N = 2^M$ fragments, the algorithm has $M = \log_2 N$ assembly levels, highlighting its systematic approach to building complex structures. Adapted from [194].

accessible to IDPs.

In this work, we apply HCG and RHCG to investigate the conformational ensemble of the CTD and its potential role in mediating phase separation and protein–protein interactions relevant to transcription. These models provide insight into the dynamic, heterogeneous nature of the CTD and help elucidate how its structural properties contribute to transcriptional condensate formation.

In this chapter, we detail the methodology of the HCG approach, emphasizing its application to IDPs such as the CTD.

3.2 Methods

We employed the Hierarchical Chain Growth (HCG) algorithm [195] to generate initial conformational ensembles of IDPs. Chain growth algorithms can be advantageous for modelling biomolecules with intrinsically disordered regions. The weak coupling between different segments of IDPs allows for simplified sampling. Instead of simulating the entire protein at once, individual fragments can be simulated independently to efficiently explore their conformational space.

Hierarchical Monte Carlo Chain-Growth Algorithm

The Hierarchical Monte Carlo Chain-Growth (HCG) algorithm was used to generate initial conformational ensembles of IDPs. These algorithms begin by dividing biomolecules into small fragments, creating a pool of structural fragments (Fig 3.2). The Hierarchical Chain Growth (HCG) algorithm leverages the hierarchical organization of IDPs by dividing protein chains into overlapping sequence fragments [195]. The details of the method is described below:

Replica Exchange Molecular Dynamics (REMD) Simulations

To efficiently sample the conformational space of intrinsically disordered protein (IDP) fragments, Replica Exchange Molecular Dynamics (REMD) simulations were employed. REMD enhances conformational sampling by running multiple replicas at different temperatures, allowing exchanges between neighbouring replicas. This helps overcome energy barriers and accelerates the sampling of rare events like proline isomerization [196, 195]. These simulations were run independently for each fragment. The `Amber99sb-star-ildn-q` force field was used for the protein structures, and the simulations were performed in an explicit solvent with the `TIP3P` water model. The REMD simulations spanned across 32 replicas, with temperatures ranging from 300 K to 540 K. Exchange attempts were made every 1 ps to promote transitions between energy states, and each fragment was simulated for a total time of 1.9 μs, ensuring thorough sampling of the conformational space.

Assembly of Full-Length yCTD Chains

A library of 10,000 fragment structures for each peptide segment was created using the REMD data. Structures are extracted from the 300 K ensemble. These fragment structures were assembled into full-length yCTD chains using the HCG algorithm. During the assembly process, steric exclusion was enforced by preventing atomic overlaps between fragments, ensuring physically realistic structures. This was done using distance-based cutoffs to maintain proper spatial arrangement and avoid unrealistic overlaps.

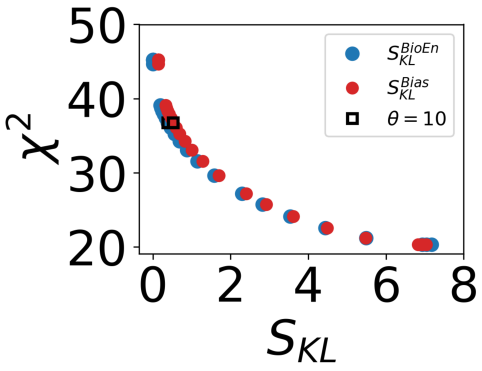

Fig. 3.3: (A) L-curve analysis for the global BioEn KL divergence S_{KL}^{BioEn} from the initial weights (blue) and the KL divergence S_{KL}^{bias} from uniform weights in ideal importance sampling (pink). The black circles indicate the KL divergences for the global BioEn with $\theta = 10$.

Refinement and Validation Methods

Reweighted Hierarchical Chain Growth (RHCG)

The Reweighted Hierarchical Chain Growth (RHCG) method was employed to refine the initial structural ensemble using experimental data. Bayesian inference was applied to combine the HCG-generated ensemble with experimental data, enhancing the agreement between computational models and experimental observations.

Chemical Shift Refinement

To refine the structures further, SPARTA+ and POTENCI tools were used to predict and analyze chemical shifts. SPARTA+ employs a machine-learning approach to predict chemical shifts based on known structures, while POTENCI analyzes secondary shifts. BioEn reweighting was applied, with a confidence parameter (θ_f) of 20, to adjust the structural ensemble to match experimental chemical shift data for Cα, Cβ, N, and HN chemical shifts.

Global Reweighting

The final structural ensemble was further refined through global reweighting using BioEn with the parameter θ set to 10, determined by L-curve analysis(Fig. 3.3). This optimization ensured that the structural ensemble closely matched experimental constraints while maintaining its physical validity.

Hydrodynamic Radius (Rh) Refinement

The hydrodynamic radius (Rh) was calculated for each structure in the ensemble. The calculated Rh values were compared to experimentally measured Rh. The refinement process adjusted the structures to match these experimental values, ensuring that the final ensemble was consistent with experimental measurements of size.

NOE Contact Analysis

For further validation, 21 residual fragments corresponding to YSPTSPS repeats were extracted from the yCTD ensemble. Nuclear Overhauser Effect (NOE) contacts were computed and compared with experimental data for the 3R-CTD construct. We utilized the BME2

(https://github.com/sbottaro/BME2) library to match the upper bound distances from the NOE measurements of 3R-CTD [197].

Residual Dipolar Coupling (RDC) Analysis

Residual Dipolar Coupling (RDC) analysis was performed using the PALES method to assess the global alignment and accuracy of the structural ensemble [198]. RDC values were calculated for each structure in the ensemble and compared with experimental RDC data, further validating the long-range structural properties and global alignment of the protein.

Statistical Independence and Detailed Balance

The HCG algorithm preserves detailed balance, ensuring that the final conformational ensemble is independent of the order in which the fragments are assembled. This ensures unbiased sampling of the IDP conformational space, allowing the method to explore a broad range of potential conformations.

Software and Computational Setup

All simulations were performed using GROMACS. The HCG and RHCG algorithms were implemented using a publicly available framework (https://github.com/bio-phys/hierarchical-chain-growth), facilitating reproducibility and further methodological improvements. The framework allows for easy integration of new experimental data to improve the refinement process and expand the scope of the algorithm for other IDPs.

Refining IDP Ensembles Using Bayesian Methods

The Hierarchical Chain Growth (HCG) algorithm is designed to efficiently generate diverse structural ensembles of IDPs by assembling fragments sampled from molecular dynamics (MD) simulations. This hierarchical process, which scales logarithmically with protein length, produces a wide array of conformations. To ensure these ensembles align with experimental observations, Bayesian Inference of Ensembles (BioEn) is applied.

BioEn refines the initial ensemble by adjusting the weights of each conformation to optimize the fit with experimental data, such as NMR chemical shifts. The weight vector $\mathbf{w} = (w_1, ..., w_C)$ contains the normalized weights of the individual chains in the ensemble. The refinement uses a posterior probability distribution $P(\mathbf{w}|\text{data}, I)$, which depends on the weights $\mathbf{w}$, the experimental data, and any prior information I:

$$P(\mathbf{w}|\text{data}, I) \propto P_0(\mathbf{w}|I) \cdot P(\text{data}|\mathbf{w}, I)$$

In the BioEn method, the prior probability distribution $P_0(\mathbf{w}|I)$ over the ensemble weights $\mathbf{w}$

is formulated as a maximum-entropy distribution given by:

$$P_0(\mathbf{w}|I) \propto \exp\left(-\theta S_{\mathrm{KL}}\right) = \prod_{c=1}^{C} \left(\frac{w_c^0}{w_c}\right)^{\theta w_c},$$

where θ is a hyperparameter controlling entropy regularization, and S_{KL} is the Kullback-Leibler (KL) divergence:

$$S_{\mathrm{KL}} = \sum_{c=1}^{C} w_c \ln\left(\frac{w_c}{w_c^0}\right),$$

measuring the deviation of the refined weights w_c from the initial weights.

The Reweighted Hierarchical Chain Growth (RHCG) method builds on the HCG algorithm, designed to integrate experimental data into the process of ensemble generation [199, 195]. RHCG improves on HCG by assigning probabilistic weights to fragment conformations rather than treating all fragments equally. This weighting corrects potential biases in the fragment pool, such as those arising from force-field inaccuracies that could skew the conformer distribution (e.g., favouring extended over helical structures). Without such weighting on the level of fragments, larger chains could become dominated by only a few conformational types when refining such ensembles against experimental data [200]. To resolve this, RHCG applies Bayesian inference to refine the weights, ensuring that the generated ensemble not only captures a wide range of structural possibilities but also remains consistent with experimental data.

The RHCG focuses on fragment-based chain growth. At each position n along the chain, there are F fragments in the library to choose from. Instead of picking fragments with equal weights as in HCG, RHCG picks fragments according to a set of weights w_n^f, where the weights are normalized such that the sum over all fragments f at position n is $\sum_f w_n^f = 1$. The probability of selecting a particular chain c is the product of the weights for its fragments: $p[f^c] = \prod_{n=1}^{F} w_n^{f_n^c}$. The bias introduced during this process is later corrected in a reweighting step to match experimental data.

The fragment library is generated from MD simulations of short overlapping peptides. To bias the fragments, NMR chemical shifts are used. SPARTA+ predicts backbone chemical shifts for proteins based on their three-dimensional structure, focusing on atoms like Cα, Cβ, N, and H$_N$ [201]. It provides error estimates (0.92 ppm for Cα, 1.13 ppm for Cβ, 2.45 ppm for N, and 0.49 ppm for H$_N$) that help refine protein structures, and its predictions can be combined with methods like POTENCI to determine secondary shifts [202].

In an ideal importance sampling framework, chains are grown with equal probabilities, ensuring that each generated chain is an equally representative member of the target ensemble. This results in uniform weights after reweighting with Bayesian Inference of Ensembles (BioEn). The effectiveness of the bias during chain growth is evaluated using the

KL divergence of the BioEn-optimized weights w_c from ideal importance sampling:

$$S_{\mathrm{KL}}^{\mathrm{bias}} = \sum_{c=1}^{C} w_c \ln\left(C w_c\right)$$

If $S_{\mathrm{KL}}^{\mathrm{bias}} \ll 1$, it indicates that, there is a high overlap between the biased ensemble and the reweighted ensemble, meaning the bias introduced during chain growth was effective and close to ideal importance sampling. Conversely, if $S_{\mathrm{KL}}^{\mathrm{bias}} \gg 1$, it suggests the overlap is small and that the chain growth method needs further optimization.

Finally, $S_{\mathrm{KL}}^{\mathrm{bias}}$ is used to determine the optimal confidence parameter, θ_f, which controls the strength of bias during fragment selection. The value of θ_f is chosen to minimize $S_{\mathrm{KL}}^{\mathrm{bias}}$, striking a balance between guiding chain growth toward the desired distribution and avoiding overfitting to experimental data. This refinement process was implemented using the BioEn library ([https://github.com/bio-phys/BioEn]).

Hydrodynamic Radius Calculation

The hydrodynamic radius (R_h), or Stokes radius, represents the effective radius of a theoretical hard sphere with the same translational diffusion coefficient as the protein of interest [87]. The translational diffusion coefficient (D_t) can be measured experimentally through methods such as pulsed-field gradient NMR, fluorescence correlation spectroscopy, and dynamic light scattering. The Stokes-Einstein equation relates D_t to R_h:

$$D_t = \frac{k_B T}{6\pi\eta R_h} \tag{3.1}$$

where k_B is the Boltzmann constant, T is temperature, and η is solvent viscosity. The hydrodynamic radius (R_h) is calculated using the Stokes-Einstein equation. An empirical parameterization relates R_h to the radius of gyration (R_g) and number of residues (N):

$$\frac{R_g}{R_h}(N, R_g) = \alpha_1 R_g - \alpha_2 N^{0.33} \cdot (N^{0.60} - N^{0.33}) + \alpha_3 \tag{3.2}$$

where α_1, α_2, and α_3 are empirically determined parameters. The radius of gyration (R_g) is calculated for each structure, and these values are converted to R_h using an empirical parameterization.

NMR Residual Dipolar Couplings (RDCs)

Residual dipolar couplings (RDCs) provide insights into the orientation and dynamics of molecular structures in solution[203]. In this study, RDCs were calculated from ensembles of full-length structures using the PALES software in steric alignment mode. For each residue i, the backbone amide RDC value $D_{\mathrm{HN}}(i)$ was determined by aligning each chain c in the

ensemble and averaging the results across all structures. This process involves calculating the alignment tensor, which captures the average orientation of the molecule relative to the external magnetic field. The RDC constant is given by:

$$D_{\text{HN}} = \frac{\mu_0}{4\pi} \cdot \frac{\hbar \gamma_{\text{H}} \gamma_{\text{N}}}{r_{\text{HN}}^3} \cdot (3\cos^2\theta - 1) \cdot S$$

where μ_0 is the permeability of free space, $\hbar$ is the reduced Planck constant, γ_1 and γ_2 are the gyromagnetic ratios of the nuclei, r is the distance between the nuclei, θ is the angle between the internuclear vector and the magnetic field, and S is the alignment factor derived from the molecular alignment tensor. RDCs provide long-range orientational information, complementing traditional NMR techniques such as Nuclear Overhauser Effects (NOEs), which focus on short-range distance constraints.

Nuclear Overhauser Effect (NOE) Calculations

NOEs provide crucial short-range distance constraints in biomolecular structure determination [204]. The average effective distance (r) from NOE data is calculated using the equation:

$$r = \langle r_i^{-6} \rangle^{-1/6}$$

where r_i is the distance in structure i. This equation reflects the r^{-6} dependence of NOE intensities, emphasizing shorter distances.

3.3 Results and Discussions

Conformational Ensemble Generation

To explore the conformational landscape of the yeast RNA polymerase II C-terminal domain (yCTD), we employed a hierarchical Monte Carlo chain-growth algorithm combined with Replica Exchange Molecular Dynamics (REMD) simulations. Given the dynamic and flexible nature of CTD, accurately modelling their conformational transitions, especially for proline isomerization, is crucial for understanding their function in cellular processes.

REMD can generate libraries of fragment conformations that sample both cis and trans states 3.5B. Figure 3.5 A and B show the cis and trans isomers of a fragment, respectively. REMD accelerates sampling by allowing exchanges between replicas at different temperatures, enabling transitions like proline cis-trans isomerization that occur on longer timescales. We observed around 0.3 per cent of cis isomers at 300K from the REMD simulations. After 900 ns we started to observe cis isomers.

Proline isomerization plays a critical role in protein folding, dynamics, and function [205]. Proline is unique among amino acids because of its cyclic structure, which limits

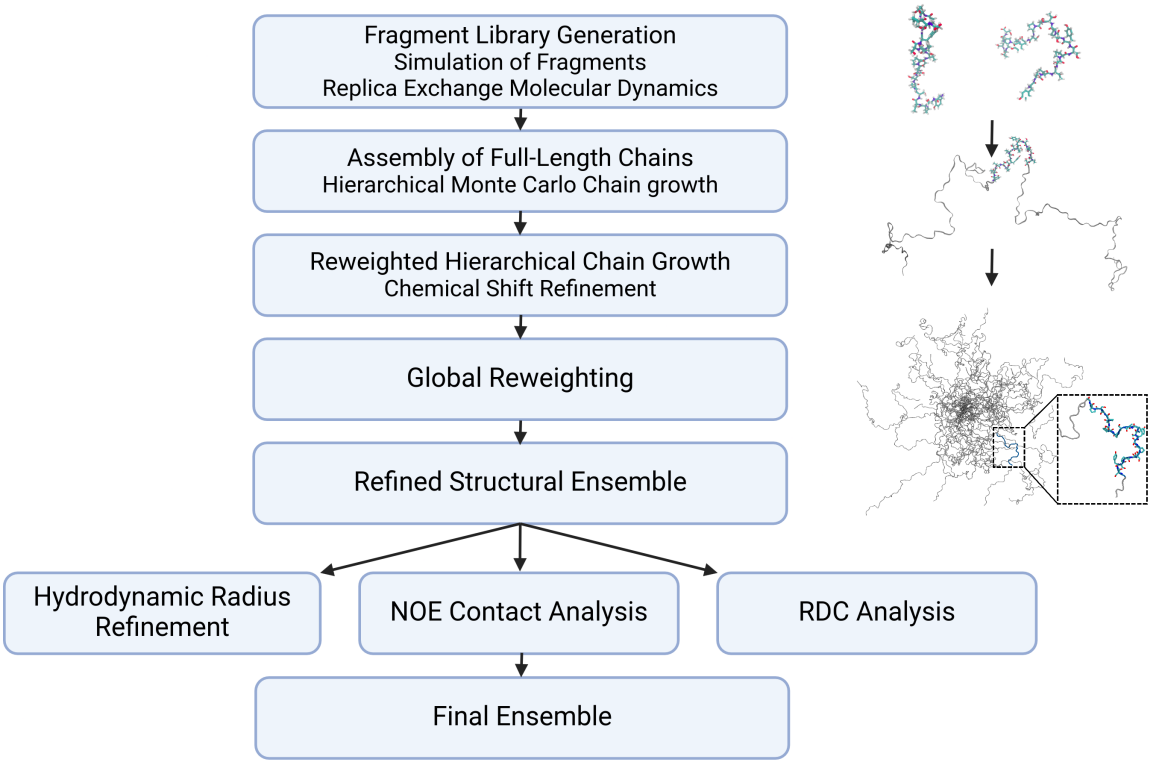

Fig. 3.4: Workflow for Generating yCTD Conformational Ensembles.

the flexibility of its backbone. As a result, peptide bonds involving proline, particularly Xaa-Pro bonds, have a greater tendency to adopt the cis conformation compared to other amino acids. Under physiological conditions, about 10 % of Xaa-Pro bonds exist in the cis state. This isomerization process is energetically demanding and occurs on a much slower timescale than protein folding, often serving as a rate-limiting step in folding pathways. Accurately capturing proline cis-trans isomerization in molecular simulations poses significant challenges. Advanced techniques like replica exchange molecular dynamics and the use of precise force fields are necessary to overcome sampling inefficiencies and ensure reliable modelling of proline's conformational dynamics.

Assembly of Full-Length yCTD Chains from HCG Ensembles

We used structures from the T = 300 K ensemble to assemble chains. we created a library of 10000 fragment structures for each peptide segment at the temperature of interest (Fig. 3.6A). This extensive library allows us to capture a wide range of conformations for yCTD. The algorithm grows the full-chain by assembling possible structures of dimers, quadruplets, and so on until the entire chain is constructed. At each step of the assembly, steric exclusion is imposed to prevent atom overlaps. To evaluate the accuracy of our models, we compared them to full-length CTD structures generated from the HCG ensemble with the NMR structure

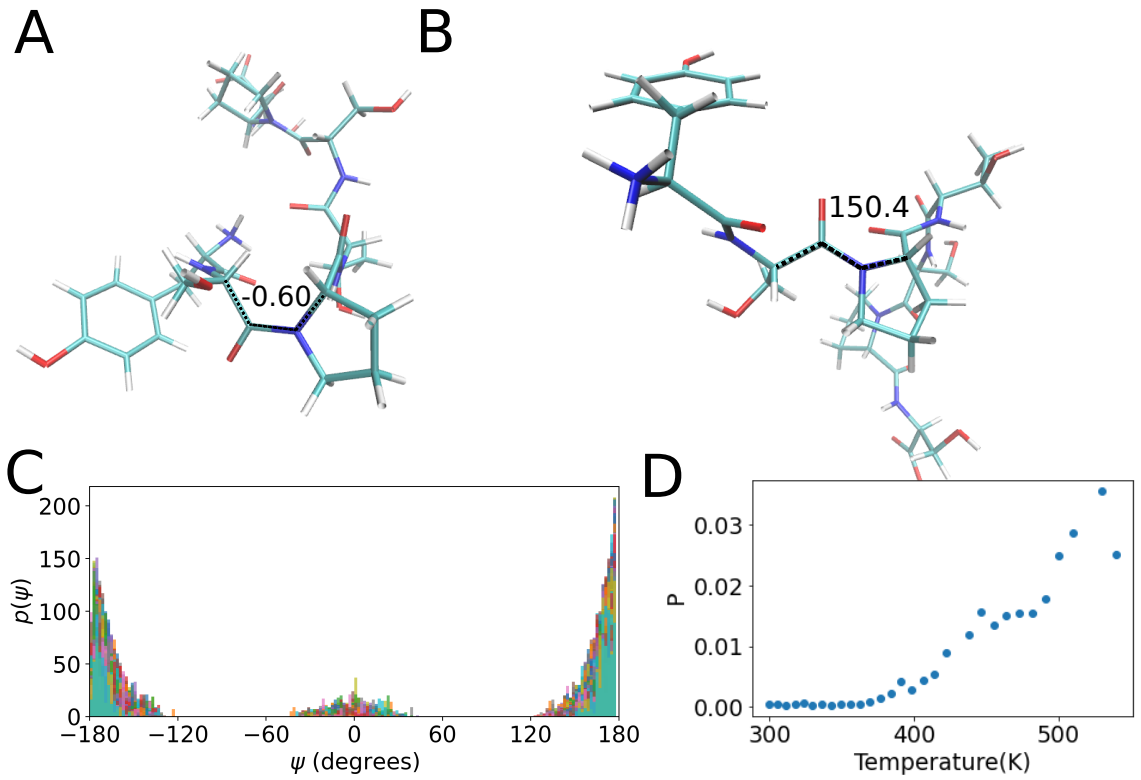

Fig. 3.5: (A) Representative structure of fragment with a cis proline isomer, illustrating the local backbone conformation. (B) Representative structure of fragment with a trans proline isomer, showing the structural differences compared to the cis form. C)Distribution of the ψ dihedral angle across different fragments, indicating the presence of both cis and trans isomers. Different colours represent distinct sequence fragments analyzed in the simulation. d) Probability of the cis proline conformation of one fragment at different temperatures, demonstrating the temperature-dependent shift in isomerization equilibrium.

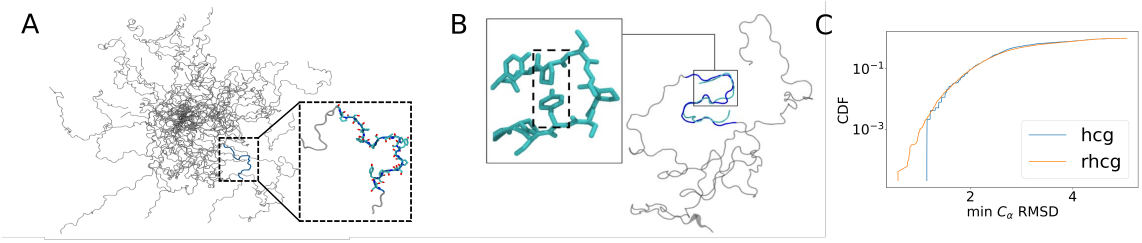

Fig. 3.6: Hierarchical chain growth ensemble. (A) Integrative models of RNA polymerase II CTD structural ensembles, with the 21-mer highlighted. (B) Comparison of full-length CTD models from the hierarchical chain-growth (HCG) ensemble computation with the NMR structure of three heptad YSPTSPS CTD-repeats (cyan). The full-length CTD model from the hierarchical chain growth ensemble is shown in grey, with the aligned region highlighted in blue. C) Minimum Cα root mean square deviation (RMSD) between conformations from HCG and RHCG and equivalent regions from the NMR-data based structural model for the three heptad repeats.

of three heptad YSPTSPS CTD repeats (depicted in cyan). The NMR structure was derived using Rosetta [206], incorporating data from chemical shifts, NOE contacts, and 1H-15N residual dipolar couplings. In the Rosetta calculation, one assumes that each structure needs to fulfill all the experimental restraints. This assumption is typically not justified [207] but is the basis of determining structures from NMR data. In practice, this structure can be considered as one that best matches experimental data without accounting for conformational heterogeneity. Conformational heterogeneity often results in multiple distinct conformations that better represent molecular-level behavior than a single average structure. For instance, the NMR structure of a functionalised DNA molecule computed by such an approach matches less well with the raw experimental data than the MD simulations as the simulations show averaging between two different conformations, which both give rise to distinct NMR signatures which also found in the experimental raw data [204]. Two representative CTD models from the HCG ensemble are illustrated in grey, with the aligned region highlighted in blue (Fig. 3.6B). A considerable fraction of the conformations of HCG and RHCG structures are very close to the NMR-based Rosetta structure, as is apparent by computing the Cα RMSD to the Rosetta structure. The yCTD ensembles generated through hierarchical chain growth (HCG) show a strong correlation with NMR experiments that probe chemical shifts (Fig 3.7).

Refinement Against Chemical Shifts

After generating conformational ensembles, we refined yCTD fragments using experimental chemical shift data. Chemical shifts are especially informative for IDPs because they provide insights into transient structural features and dynamic interactions that are difficult to capture using traditional structural techniques. Incorporating them into our model helps ensure that the conformational ensemble aligns closely with what is observed in biological systems.

HCG provides a systematic approach to generating such ensembles. Once the initial ensemble is generated, the reweighting technique RHCG is employed to better fit experimental data, making the approach highly adaptable to integrating with experimental measurements. In the Reweighted Hierarchical Chain Growth (RHCG) method, we tailored fragment selection based on $C\alpha$,$C\beta$,N, and HN chemical shifts obtained from NMR measurements. We used SPARTA+ and POTENCI tools for this refinement.

For each fragment position, we performed independent BioEn reweighting using the chemical shift data. We set a confidence parameter θ_f of 20 to balance the influence of experimental data with prior knowledge about the system. This parameter determines the weight given to experimental chemical shifts versus the predictions from our computational models. Applying a high confidence parameter increases the impact of the chemical shift data, ensuring that conformations closely matching the experimental shifts are assigned higher weights in our ensemble. A lower value would prioritize prediction. By setting a high confidence threshold, we reduce noise and variability in the data, resulting in a more consistent representation of

the local structural features across all fragments. Here, by choosing $\theta_f = 20$, we aimed to better capture the experimental insights into the prior distribution so that both contribute effectively to the refinement process.

After chemical shift refinement, we performed a global reweighting step(Fig. 3.4). The bias in the assembly of the RHCG ensembles was corrected through final BioEn refinement. This step is crucial as it ensures that the entire ensemble is consistent with all available experimental data while correcting any biases introduced during fragment-based simulations. The global reweighting helps mitigate discrepancies and enhances the accuracy of the ensemble representation. With θ set at 10, our results indicate that $S_{\mathrm{KL}}^{\mathrm{BioEn}}$ is less than 1, and $S_{\mathrm{KL}}^{\mathrm{bias}}$ is also around 1 in Fig 3.3. This suggests that the fragments generated from the REMD simulations and the reweighting on the fragment level are already well aligned with the NMR chemical shift data, which helped guide the process by balancing the effect of experimental data with the initial ensemble distribution.

Hydrodynamic Radius Refinement

Following the initial ensemble generation, we observed that our structures did not match the experimental hydrodynamic radius (Rh) values well. This mismatch necessitated a second refinement step focused on aligning our yCTD ensembles with experimentally measured Rh. While chemical shift refinement ensures local structural accuracy, hydrodynamic radius refinement provides validation of overall protein size and shape in solution. Rh, or the Stokes radius, measures protein expansion based on translational diffusion, determined experimentally via techniques like NMR or light scattering. The hydrodynamic radius provides information about the overall size and shape of the protein in solution, which is particularly important for IDPs that can adopt various conformations depending on their environment. It helps validate our models against experimental data. For each structure in the ensemble, we calculated the hydrodynamic radius [87], ensuring that the refined ensembles accurately reflected the experimental data. This step enhances our confidence in the biological relevance of our generated ensembles by ensuring they closely match experimentally observed values. This combined approach of using RHCG with chemical shift-guided fragment selection, BioEn reweighting, and hydrodynamic radius refinement enabled us to generate ensembles that closely match experimental NMR data, providing a reliable representation of the yCTD structure in solution. The secondary chemical shifts ($C\alpha$,$C\beta$,N, and HN) are calculated from the yCTD ensemble, obtained using RHCG and refined against R_H, and were compared with experimental NMR shifts (Figure 3.7). The RHCG ensemble generally reproduced the overall trend of experimental shifts. The models consistently produced slightly larger shifts than those measured by NMR. Quantitative measures of agreement were calculated to assess the accuracy of the predictions. The Pearson correlation coefficients (r) between predicted and experimental shifts were calculated.

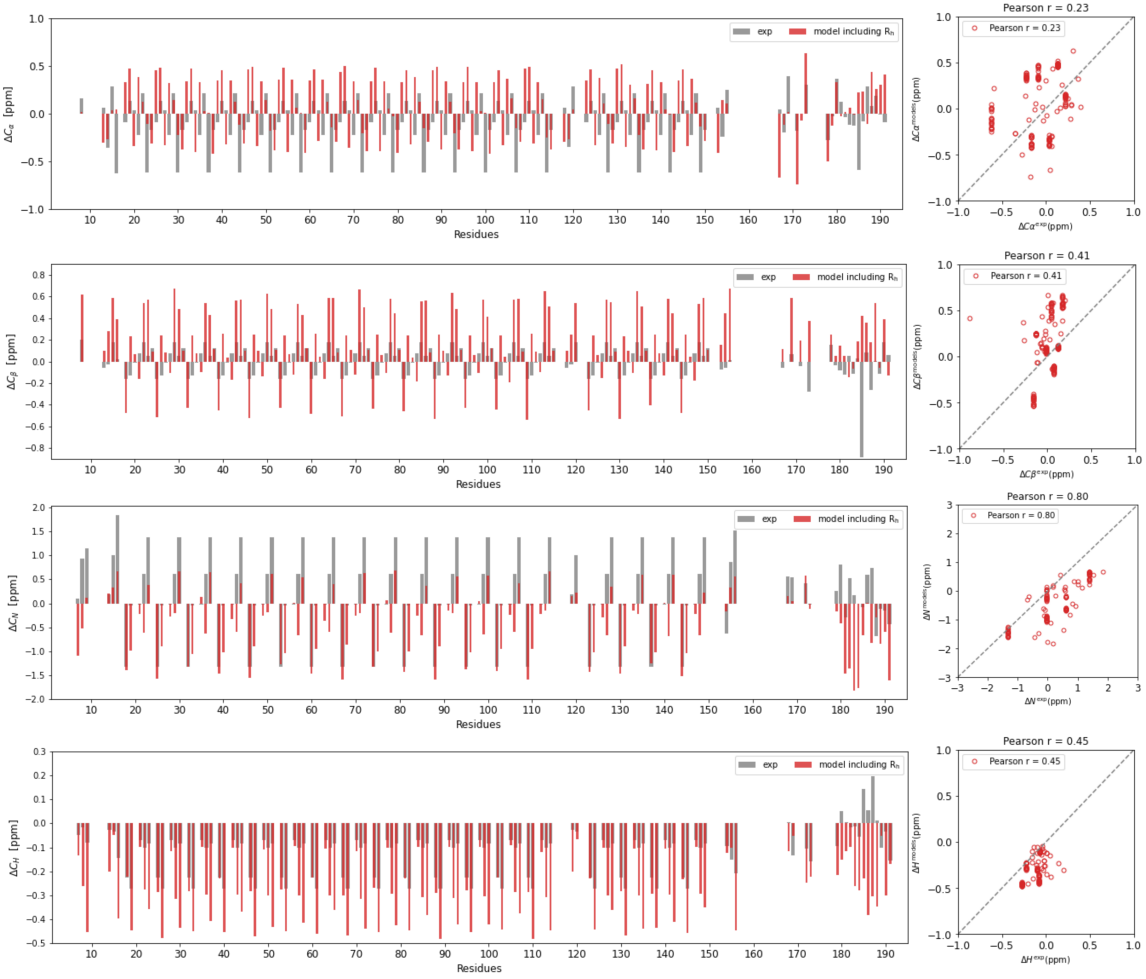

Fig. 3.7: Comparison of NMR Chemical Shifts for yCTD Fragment Simulations and Full-Length Models from Hierarchical Chain Growth: The analysis of NMR chemical shifts for yCTD fragment simulations and full-length models obtained from hierarchical chain growth comparing with secondary shift values for $C\alpha$,$C\beta$,N, and HN atoms. These values are derived from different sources: fragment all-atom molecular dynamics (MD) simulations, assembled full-length models, and experimental data.

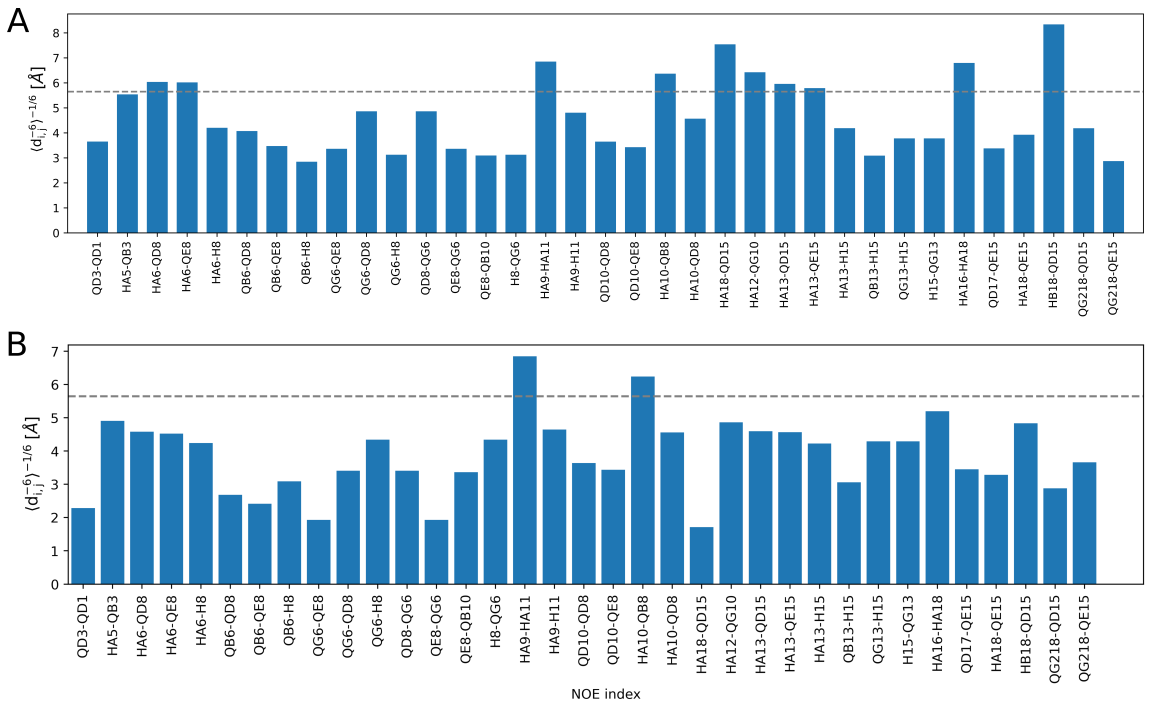

Fig. 3.8: NOE analysis: A) Without any refinement, 29 out of 36 experimental NOE contacts within 5.65 Åwere matched, with two additional contacts at the cutoff. B) NOE after the final reweighting of the 21mer fragment extracted from the full-length chains. After minimal refinement, the 3R structures from yCTD HCG matched 33 out of 36 NOE contacts within the threshold, with one contact slightly exceeding it.

Extraction and Comparison of Fragments

To further validate our yCTD ensemble against experimental data, we extracted 10^5 fragments, each 21 amino acids long, with the sequence YSPTSPS from the yCTD ensemble. We then calculated NOE contacts for these fragment ensembles to compare with the NMR data of 3R-CTD (Fig 3.8). NOE contacts provide information about spatial proximity between atoms, which is crucial for structure determination. Initially, without any refinement, we successfully matched 29 out of 36 experimental NOE contacts that were within 5.65 Å, with two additional contacts at the cutoff, indicating a promising level of agreement between our computational predictions and experimental observations fig. 3.8A. With minimal additional refinement, the 3R structures from yCTD HCG matched 33 out of 36 NOE contacts within the threshold, with one contact slightly exceeding it fig. 3.8B. The $S_{\mathrm{KL}}^{\mathrm{bias}}$ was below 1, suggesting that the importance sampling generated relevant structures. Our analysis demonstrated a good fit to the experimental data, with minimal changes in weights, reflecting strong agreement with the experimental chemical shifts while remaining close to the initial weights.

Residual Dipolar Coupling Analysis

To further validate the structural ensemble of the 21-mer RHCG domain of the yCTD, residual dipolar coupling (RDC) analysis was employed. RDCs provide long-range information about

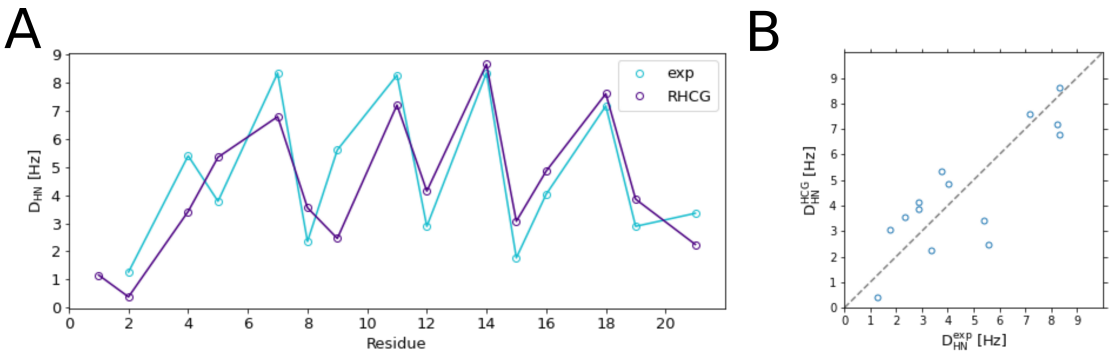

Fig. 3.9: Comparison of experimental RDCs (cyan) along the yCTD sequence with RDCs calculated from the RHCG ensemble (violet). (A) Experimental RDCs are compared with the RHCG ensemble structures. (B) Scatter plot showing the correlation between RDCs calculated from the RHCG ensemble and experimental RDCs. RDCs were calculated using the PALES method.

the molecule's shape. RDCs for the ensemble were calculated using PALES [198]. Importantly, RDCs were not used in our ensemble refinement procedure (Fig. 3.4). The results were averaged over the ensemble, taking into account the weight of each chain. Figure 3.9A shows a direct comparison between calculated RDC values from the ensemble and experimental RDC measurements.

The calculated RDCs closely matched the experimental values, indicating that the ensemble effectively captures both the overall shape and local structural details. This agreement is visually observed in Figure 3.9A. To quantify this agreement, the Pearson correlation coefficient (R-value) was calculated, yielding a value of 0.8 (Figure 3.9B). This strong positive correlation indicates that the ensemble accurately represents the structure and dynamics of the 21-mer RHCG domain in solution. The close match between computed and experimental RDC data supports the model and enables further studies of the yCTD RHCG domain.

Additionally, the yCTD ensembles were refined to match the hydrodynamic radius (RH) measured by NMR, employing the BioEn library for this ensemble refinement. For NOE contact matching, we extracted 10^5 fragments of 21 amino acids (YSPTSPS) from the yCTD ensemble and calculated their chemical shifts and NOE contacts for comparison with the NMR data of 3R-CTD.

The NMR biased ensembles matched the experimentally determined hydrodynamic radii (Fig. 3.11). Notably, the Y1 position was preferentially situated in turn regions within both the 3R-CTD and the yCTD ensemble. Turn conformations of $S_{-2}P_{-1}S_0Y_1$, characterized by O-N distances below 5 Å, were observed in approximately $\sim 16\%$ of the conformers (Fig 3.10).

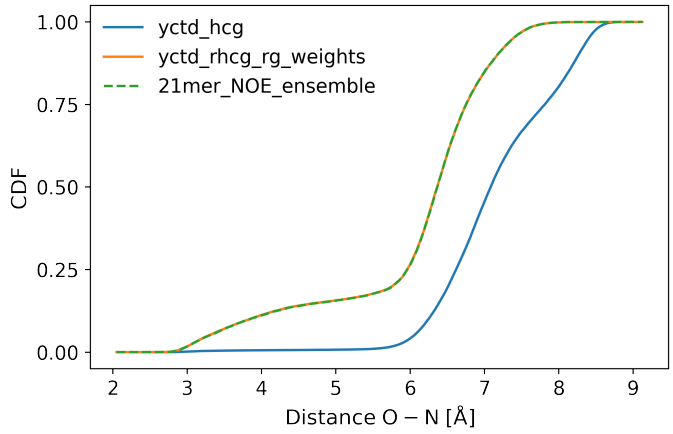

Fig. 3.10: Turn conformations of $S_{-2}P_{-1}S_0Y_1$„ with O-N distances for both the 3R-CTD and the yCTD ensemble. O-N distances below 5 Å, were observed in approximately $\sim 16\%$ of the conformers.

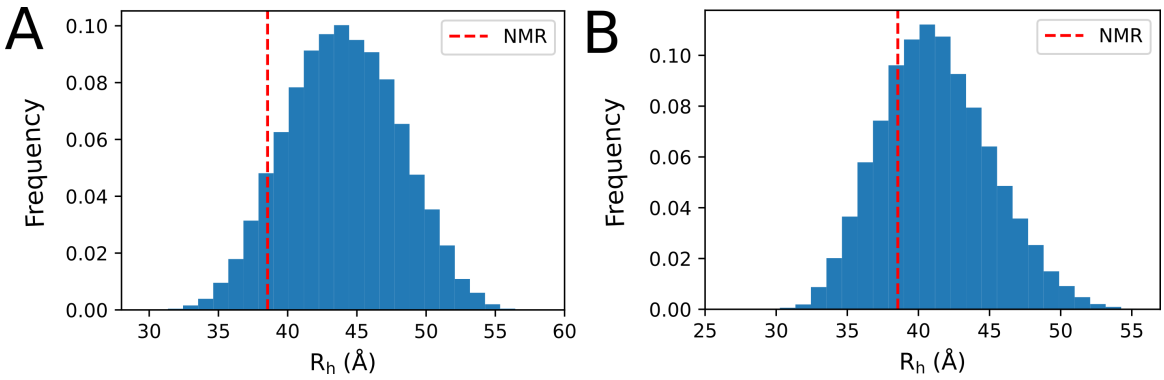

Fig. 3.11: A) Histogram distribution of R_h values for the HCG structures of yCTD compared with the experimental value at 5°C (red dashed line). A) Without reweighing, the experimental value is to the left of our distribution.

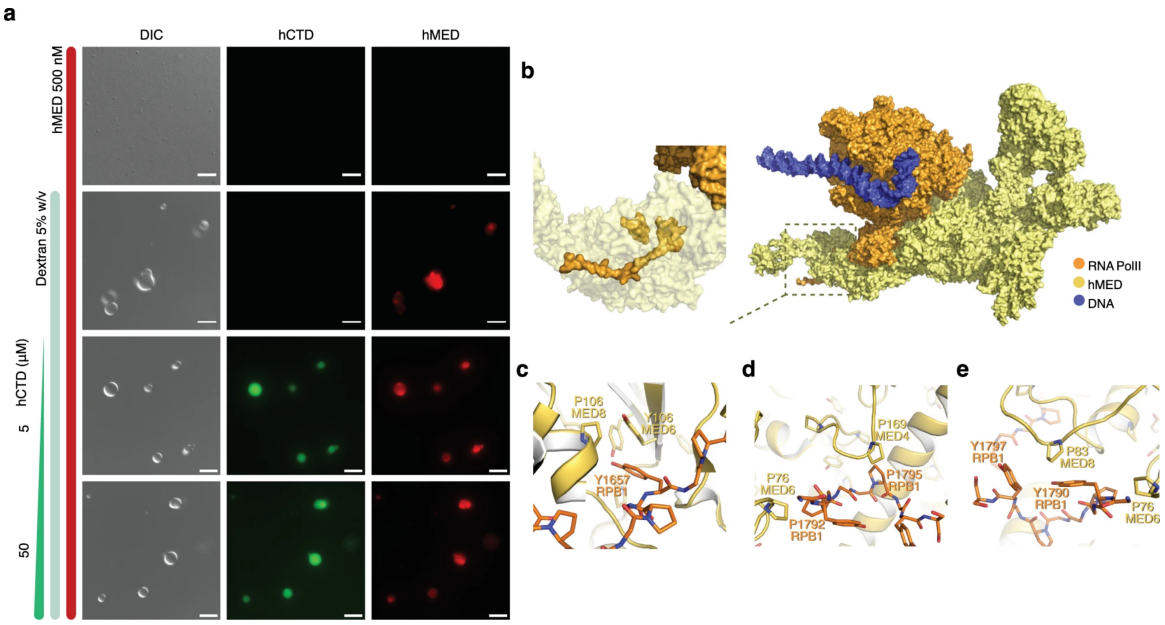

Fig. 3.12: Phase separation and structural interactions of human CTD and Mediator complex. A) Microscopy images showing phase separation of hCTD (green) and human Mediator complex (red). B)Structure of the Mediator-bound preinitiation complex. The inset shows the structure of hMED-bound CTD. c-e. Close-up views of specific interactions within the Mediator-CTD complex: c. Tyrosine-tyrosine (Tyr-Tyr) contacts d. Proline-proline (Pro-Pro) contacts e. Tyrosine-proline (Tyr-Pro) contacts. Adapted from [190]

3.3.1 Tyr-Pro interactions are important for the contacts of CTD with the Mediator complex

Intriguingly, the Pro-Tyr contacts, which are highlighted by the RHCG and NMR studies of CTD on its own, are also found in interactions of CTD with the Mediator complex (Fig 3.12) [190]. Specifically, the study demonstrates that the CTD phase separates together with the Mediator complex, suggesting a functional link between these interactions and mediator recruitment. These interactions observed via NMR, indicate a direct physical association between Pro-Tyr, Pro-Pro, and Tyr-Tyr contacts between CTD and Mediator complex(Fig 3.12 e-f). This co-recruitment suggests that Tyr-Pro interactions facilitate the CTD and Mediator complex interaction. These interactions are a key aspect of CTD's role in transcription.

3.4 Conclusions

This study offers a novel approach to address the longstanding structural ambiguity of the yeast RNA polymerase II C-terminal domain (yCTD), a critical regulator of gene expression. By integrating a hierarchical Monte Carlo chain-growth algorithm (HCG), replica-exchange molecular dynamics (REMD), and NMR data, we have mapped the dynamic structural ensemble of yCTD with high resolution.

Previous research has presented conflicting views on the structure of the CTD, with some studies proposing compact conformations such as a β-spiral, which helps to incorporate into transcription initiation complexes. However, the low chemical shift dispersion observed in NMR experiments contradicts the existence of a stable 3D fold, consistent with its intrinsic disorder. This persistent structural ambiguity has hindered precise mechanistic understanding of the CTD's role in transcriptional regulation.

Our simulations reveal that yCTD exhibits a dynamic, predominantly random coil structure with a slight tendency to form turns, particularly at the Y1 heptad position. These findings align with existing experimental observations. The hydrodynamic radii (Rh) fall within a range intermediate to those typically observed for IDPs and denatured proteins in the yeast CTD. This indicates that the CTD exhibits unique structural properties falling between those of fully unstructured protein or compact IDPs. Importantly, our REMD simulations captured and characterized the conformational heterogeneity of yCTD. With experimental data to reweight them, the diverse structures that arise from the HCG simulation are consistent with observations that suggest the absence of stable interactions. This successful reconciliation of NMR and structural behaviour.

A key advancement of our approach is the explicit treatment of proline isomerization. Proline residues are abundant in yCTD and are known to exhibit cis- and trans- conformations that can significantly impact local and global protein structure. REMD simulations allowed us to overcome energy barriers and efficiently sample both cis- and trans- proline states, revealing the dynamic interplay between these conformers and their influence on yCTD's overall conformational ensemble.

By generating structural ensembles that are not only consistent with experimental NMR data but also explicitly account for proline isomerization dynamics, our approach offers a more accurate and complete representation of yCTD structure. This level of detail is crucial for understanding how specific sequence elements and cis/trans proline equilibria cooperate to modulate yCTD function and facilitate its interactions with other proteins involved in transcription.

4

Stochastic Models for phosphorylation and their application in molecular dynamics simulations of protein condensates

This chapter explores how cells use chemical modifications, specifically phosphorylation, to control the formation and behaviour of biomolecular condensates. The study focuses on the C-terminal domain (CTD) of RNA Polymerase II, a protein critical for transcribing DNA into RNA, which we discussed in detail in the previous chapters. The CTD undergoes frequent phosphorylation, where phosphate groups are added or removed, altering its properties and interactions. The research explores how phosphorylation mechanisms influence condensate formation, stability, and organization. By examining varying phosphorylation and dephosphorylation rates, this work aims to provide insights into the non-equilibrium control of cellular structures and their adaptive responses to changing conditions.

4.1 Introduction

Biomolecular condensates are typically driven by weak and multivalent interactions between biomolecules, often involving intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) of multi-domain proteins with both folded and disordered regions. Interestingly, in comparison to folded protein domains, IDRs are often enriched in phosphorylation sites, making them key targets for post-translational modifications (PTMs) that regulate condensate behavior.[208].

Phosphorylation— the reversible addition of phosphate groups to proteins— is a key post-translational modification (PTM) that responds to cellular signals. In IDRs, phosphorylation can alter the charge, structure, and interaction properties of proteins, influencing the stability, assembly, and disassembly of condensates. [209, 23, 210, 211]. This dynamic regulation is essential for processes like gene expression, signal transduction, and stress responses[212]. Enzymes called kinases phosphorylate proteins in the process of consuming adenosine triphosphate (ATP) and yield adenosine diphosphate (ADP). Conversely, phosphatases remove phosphate groups from these proteins, increasing the concentration of inorganic

phosphate (Pi) in the solution. Cycles of phosphorylation and dephosphorylation represent non-equilibrium steady states when ATP, ADP, and Pi concentrations are maintained by external processes, as occurs in cells. Chemically driven dynamics can determine the size of protein droplets [213], and such systems can be considered active emulsions.

Dynamic phosphorylation of disordered proteins is essential for regulating gene transcription[214, 215, 216]. Hnisz et al. proposed that dynamic phosphorylation would determine whether condensates underpinning the transcription of genes form. A specific example of this interplay between phosphorylation dynamics and condensation is seen with the C-terminal domain (CTD) of RNA Polymerase II [217, 215]. During transcription initiation, the CTD forms condensates or is recruited to condensates formed by transcription factors [218, 219, 215, 216, 220]. These condensates could create local concentrations of factors necessary for gene activation. As transcription initiation progresses, the CTD of RNA Polymerase II is progressively phosphorylated. Phosphorylation of CTD [220] is essential for switching from the initiation phase of transcription to the elongation of RNA transcripts[133].

However, phosphorylation events can dissolve these condensates [217]. Simulations show that CTD, phosphorylated CTD (pCTD) can, in principle, partially demix and form distinct condensate phases[221], explaining why condensates of initiating and elongating RNA Polymerase II are partially separated in *C. elegans* embryos. Simulations further showed that the resulting morphologies can be understood by considering the interfacial tensions of the different phases[181, 221]. Co-localization of pCTD and protein required for the splicing of nascent RNA transcripts has also been reported in cells[183].

Molecular dynamics simulations can directly reveal how chemical modifications of individual residues change the properties of proteins[211, 210], but chemically driven nonequilibrium dynamics such as phosphorylation and dephosphorylation cycles have not been extensively studied through molecular simulations. Relaxation from a state far from thermodynamic equilibrium has been simulated by Grazaizar et al. [222]. Chemical reactions in condensates have been simulated with a reactive force field close to atomic resolution[223]. The enzymatic phosphorylation of proteins in condensates can be simulated in residue-level coarse-grained molecular dynamics[224] by combining molecular dynamics with Monte Carlo steps for chemical reactions[225, 226]. However, even at a residue-level resolution, it is difficult to simulate how condensates are shaped by enzymatic reactions, as large simulation systems are required and simulation trajectories have to be simulated for long durations to adequately sample how condensates are shaped by these reactions.

To better understand these dynamics we have developed models to simulate the stochastic switching between phosphorylated and unphosphorylated states in molecular dynamics simulations. We studied the stochastic switching of serine (Ser) to phospho-serine in a system composed of two components: pCTD (phosphorylated C-terminal domain) and CTD (unphosphorylated C-terminal domain). The dynamics of this system are examined under different rates for phosphorylation and dephosphorylation reactions, considering two distinct models:

the random-uniform switching model and the Processive chain switching model. These reflect two different biological mechanisms based on the way kinases can interact with their substrates. In the random-uniform switching model, phosphorylation and dephosphorylation events are assumed to occur independently across all serine residues within a CTD chain. Each serine residue has an equal probability of undergoing modification, regardless of the state of neighbouring residues. This would correspond to kinases or phosphatases binding to a protein chain, modifying one residue and then immediately dissociating. By contrast, in the Processive chain switching model, the entire protein chain is either phosphorylated or dephosphorylated at once. In this model, phosphorylation and dephosphorylation events occur processively, with the modification of one serine residue coupled to the phosphorylation of neighbouring residues along the CTD chain. This processive behaviour is inspired by the processive nature of many kinases and phosphatases, which can move along protein chains, catalyzing multiple modification events in sequence before dissociating from their substrate. These models highlight how different phosphorylation mechanisms can impact the behavior of biomolecular condensates. The random-uniform switching model may better represent situations where modifications are carried out by a large number of freely diffusing enzymes, leading to a more random distribution of phosphorylation sites. In contrast, the Processive chain switching model captures scenarios where processive enzymes create localized regions of high or low phosphorylation, potentially leading to more dramatic changes in protein interactions and condensate properties. Understanding how the CTD of RNA Polymerase II is phosphorylated by kinases such as CDK9 is important as it might severely alter the critical concentrations required for this rate-limiting step[227].

Our modeling approaches demonstrate that phosphorylation dynamics profoundly affect the formation, stability, and internal organization of biomolecular condensates. For example, simulations have shown that changes in phosphorylation rates can lead to transitions between different condensate states, such as from a single large droplet to multiple smaller droplets or even to the complete dissolution of condensates. These findings provide mechanistic insights into how cells might use phosphorylation to rapidly reorganize subcellular structures in response to various stimuli.

4.2 Methods

In this chapter, we explore stochastic switching processes as a model for biochemical reactions driven by chemical fuel. Specifically, we focus on phosphorylation and dephosphorylation cycles, which are mediated by kinases and phosphatases.

To simplify the modeling, we assume that the concentrations of ATP, ADP, and inorganic phosphate Pi are maintained at steady-state levels through external cellular processes. This assumption allows us to treat reaction rates as fixed parameters rather than dynamically adjusting them based on local energy changes or environmental factors. Consequently, our

model represents a non-equilibrium system actively driven by chemical fuel input. The energy change upon switching and its influence on the reaction rate based on the local environment are not explicitly incorporated. The probability of reactions occurring is determined by molecular populations and fixed rate constants.

It is important to note that detailed balance is not enforced in this framework because the system is inherently out of equilibrium. While energy dissipation occurs during ATP hydrolysis, it is not explicitly tracked in our simulations. Instead, we focus on the stochastic dynamics of switching events and their implications for biochemical networks. It enables us to analyze key properties such as steady-state probabilities and reaction dynamics without considering the microscopic details of switching.

Gillespie's First-Reaction Method

To simulate stochastic processes in systems with multiple reactions, we employ Gillespie's First-Reaction Method. The algorithm treats the reaction system as a Markov process and simulates it using two key steps for each event: determining the time until the next reaction and selecting which reaction occurs next. One of the methods for simulating stochastic processes in systems with multiple reactions is the First-Reaction Method. The steps in the First-Reaction Method are as follows:

1. Initialization: Set the initial time $t = 0$ and define initial molecule counts X_i for each species S_i (where $i = 1, \ldots, N$). Compute the propensity a_μ for each reaction R_μ (where $\mu = 1, \ldots, M$): $a_\mu = c_\mu \cdot h_\mu$, where c_μ is the reaction rate constant and h_μ is the number of distinct combinations of reactants.

2. Generate Tentative Reaction Times: For each reaction R_v (where $v = 1, 2, \ldots, M$), generate a 'tentative' time τ_v at which it would occur. This time is sampled from an exponential distribution with rate $a_v = h_v c_v$, where: $\tau_v = -\frac{1}{a_v} \ln(r_v)$, and r_v is sampled from a uniform distribution$[0, 1]$.

3. Select the First Reaction: Identify the reaction with the smallest τ_v, as this is the first to occur: $\tau = \min(\tau_1, \tau_2, \ldots, \tau_M)$. The index μ of the reaction corresponding to τ is selected as the next reaction.

4. Update the State: Advance the system's time by τ: $t \to t + \tau$. Update the populations of the species involved in reaction R_μ.

The First Reaction Method is based on the fact that the time to the next reaction follows an exponential distribution. The joint probability of a reaction R_μ occurring at time τ is given by:

$$P(\tau, \mu) = a_\mu e^{-\sum_{v=1}^{M} a_v \tau}.$$

The algorithm achieves this distribution by treating each reaction as if it operates independently and selecting the first event that occurs.

Stochastic Switching of Bits in Molecular Dynamics Simulations

Building on Gillespie's First-Reaction Method, we developed a tailored approach to simulate stochastic switching between neutral and phosphorylated states within molecular dynamics (MD) simulations. This framework models phosphorylation-dephosphorylation cycles as stochastic events driven by fixed rates corresponding to enzymatic activity.

1. The state variable for phosphorylated and unphosphorylated bits are defined as: $\widetilde{N}_n$ and $\widetilde{N}_p$ respectively, with the total number of bits conserved according to the equation $\widetilde{N}_n + \widetilde{N}_p = Nm$.

2. Transition rates are r_{np} for neutral to phosphorylated and r_{pn} for phosphorylated to neutral. The algorithm models two events: a neutral bit switches to phosphorylated, and a phosphorylated bit switches to neutral.

3. In the molecular dynamics (MD) simulation, the next time step is determined by the smaller of two sampled exponential times, represented as: $t_{\text{next}} = \min(\widetilde{E}_{np}, \widetilde{E}_{pn})$, where $\widetilde{E}_{np} \sim \text{Exp}(\widetilde{N}_n r_{np})$ and $\widetilde{E}_{pn} \sim \text{Exp}(\widetilde{N}_p r_{pn})$.

4. After an event occurs, update the state by modifying $\dot{\widetilde{N}}_n$ and $\dot{\widetilde{N}}_p$. Repeat the process as necessary.

The rate corresponds to the activity of kinases and phosphatases, enzymes that add and remove phosphate groups to proteins. Higher rates indicate more frequent phosphorylation/dephosphorylation events, reflecting higher enzyme activity or concentration.

Building on this framework, Prof. Lisa Hartung suggest the following two algorithms to model the stochastic effects of switching between different states in our system. The base model consists of N chains, each composed of m bits. Each bit can exist in one of two states: neutral (n) or phosphorylated (p).

The uniform random switching model

The algorithm

In this model, each bit switches states independently, reflecting the random nature of phosphorylation and dephosphorylation events in cellular environments. Let $\widetilde{N}_n$ represent the number of neutral bits and $\widetilde{N}_p$ represent the number of phosphorylated bits. The total number of bits is given by:

$$\widetilde{N}_n + \widetilde{N}_p = Nm. \tag{4.1}$$

Each bit can switch its state independently, simulating the random nature of phosphorylation and dephosphorylation events in cellular environments.

1. **Generate Random Variables:** Create an exponentially distributed random variable

$\widetilde{E}_{np}$ with parameter $\widetilde{N}_n r_{np}$. This variable represents the time until the next neutral-to-phosphorylated transition occurs. Similarly, create another random variable $\widetilde{E}_{pn}$ with parameter $\widetilde{N}_p r_{pn}$, representing the time until the next phosphorylated-to-neutral transition.

2. **Run Simulation:** Execute the MD simulation for a duration equal to the smaller of the two random variables, $\min(\widetilde{E}_{np}, \widetilde{E}_{pn})$.

3. **Switch Bit State:**

   - If $\min(\widetilde{E}_{np}, \widetilde{E}_{pn}) = \widetilde{E}_{np}$, randomly select a neutral bit and switch it to phosphorylated. This simulates an event where a kinase enzyme adds a phosphate group to a protein.

   - If $\min(\widetilde{E}_{np}, \widetilde{E}_{pn}) = \widetilde{E}_{pn}$, randomly select a phosphorylated bit and switch it to neutral. This represents an event where a phosphatase enzyme removes a phosphate group.

4. **Update Counts:** Adjust $\widetilde{N}_n$ and $\widetilde{N}_p$ based on the changes made, and repeat the process from step 1. This iterative process allows for continuous monitoring and adjustment of bit states as they fluctuate over time.

Equilibrium Considerations

Over time, the numbers of neutral and phosphorylated bits, $\widetilde{N}_n$ and $\widetilde{N}_p$, will fluctuate around an equilibrium value. This equilibrium satisfies the condition:

$$\widetilde{N}_n r_{np} = \widetilde{N}_p r_{pn} \tag{4.2}$$

At equilibrium, phosphorylated bits are uniformly distributed across all chains. This uniform distribution means that most chains will appear similar, with only minor fluctuations in their composition. Occasionally, rare events may lead to larger deviations from this uniformity, but these are infrequent. This model helps us understand how phosphorylation dynamics can influence molecular behaviour in systems where stochastic switching plays a significant role.

The Processive chain switching model

The Processive chain switching model focuses on phosphorylation dynamics at the level of entire protein chains rather than individual bits. This approach reflects biological processes where enzymes like kinases and phosphatases act processively on proteins, often modifying multiple sites before dissociating from the substrate protein.

Model Description

We denote by N_n the number of chains without any phosphorylated bits (fully neutral chains) and by N_p the number of chains with at least one phosphorylated bit. The total number of chains remains constant:

$$N_n + N_p = N \tag{4.3}$$

We introduce two key rates to describe the transitions between these states:

- c_{np}: the rate at which a neutral chain is phosphorylated by a motor. This rate reflects how often kinase enzymes target and modify these chains.
- c_{pn}: the rate at which a phosphorylated chain is dephosphorylated by a motor, representing phosphatase activity.

When a neutral chain undergoes phosphorylation, each of its m bits becomes phosphorylated. Similarly, when a phosphorylated chain is targeted for dephosphorylation, all its phosphorylated bits revert to neutral simultaneously. Here, phosphorylation and dephosphorylation are rapid and coordinated.

Algorithm

The algorithm for implementing the Processive chain switching model in molecular dynamics simulations involves the following steps:

1. **Generate Random Variables:** Create an exponentially distributed random variable E_{np} with parameter $N_n c_{np}$. This variable represents the time until the next phosphorylation event occurs. Similarly, generate another random variable E_{pn} with parameter $N_p c_{pn}$, representing the time until the next dephosphorylation event.
2. **Run Simulation:** Conduct the MD simulation for a duration equal to $\min(E_{np}, E_{pn})$.
3. **Switch Chain State:**
    a) If $\min(E_{np}, E_{pn}) = E_{np}$: Select a neutral chain uniformly at random. For each bit in this chain, flip it to the phosphorylated state with probability $p_p = 1$. This simulates independent Bernoulli trials for each bit.
    b) If $\min(E_{np}, E_{pn}) = E_{pn}$: Select a phosphorylated chain uniformly at random. Convert all phosphorylated bits in this chain back to neutral.
    c) **Update Counts:** Adjust N_n and N_p based on the changes made and repeat from step 1.

Unlike the uniform random switching model, this approach leads to distinct subpopulations of chains: fully neutral chains and chains with multiple phosphorylated bits. This results in more pronounced fluctuations and highlights how collective phosphorylation events can influence system dynamics. After some time, the number of phosphorylated chains will fluctuate around the equilibrium values

$$N_n c_{np} = N_p c_{pn}. \tag{4.4}$$

If a chain gets phosphorylated by a Processive chain $p_p m$ many bits get phosphorylated. Note that in this model the phosphorylated bits will be less uniform: There will always be a percentage of chains that are neutral. The phosphorylated chains have many bits that are phosphorylated. This model highlights significant differences in interaction behaviours between neutral and phosphorylated chains. Unlike the random-uniform switching model where phosphorylation occurs bitwise across all chains ($mN \gg N$), here we observe more pronounced fluctuations due to whole-chain phosphorylation events. As a result, random fluctuations in the number of phosphorylated chains are more visible. The Processive chain switching model provides insights into how collective phosphorylation events can lead to distinct subpopulations within a system, influencing overall dynamics and interaction patterns.

Implementation

We used HOOMD-blue version 3.8.1 to simulate phosphorylation events in our protein systems. HOOMD-blue [150] allows us to combine molecular dynamics and stochastic switching of particle identities[224, 226]. In this way, we simulate the random transitions between phosphorylated and neutral states of protein chains within the molecular dynamics simulation. We set up the initial protein configurations, defined the rates of phosphorylation and dephosphorylation, and ran simulations to observe how protein phosphorylation states changed over time. We used our random-uniform and processive chain switching model to generate these random switching events. We used the HPS residue-level coarse-grained model for our simulations[139]. Parameters for phosphorylated Ser were taken from Ref [148].

Simulation setup

Three sets of systems were simulated at different levels of phosphorylation over dephosphorylation r_{np}/r_{pn}. The first set comprised slab simulations with an initial configuration of 200 CTD chains and 2 pCTD chains. To create this initial system, the chains were arranged linearly adjacent to one another and subjected to a 100 ns simulation. This preliminary system was then utilized as the starting point for conducting stochastic phosphorylation studies. The additional 2 pCTD chains were included to introduce the identity of phosphorylated serine.

The second system consisted of 300 CTD chains and 300 fully phosphorylated CTD chains in a cubical box. The initial system for this was a condensate formed by simulating linearly arranged chains for 0.5 μs. This setup, containing an equal number of CTD and pCTD chains, was then used to simulate stochastic phosphorylation and dephosphorylation events.

The third system consisted of 600 CTD, 600 pCTD, and 1200 HRD chains. The initial configuration was formed by simulating linearly arranged chains for 5 μs. During this simulation, the peptide phase separated, forming a condensate with a core-shell morphology. The core was composed of CTD chains, while the shell was formed by pCTD-HRD interactions. This initial configuration was then used to study the effects of stochastic phosphorylation and dephosphorylation under different rates, allowing for an examination of how these processes influence the complex core-shell structure of the condensate.

4.3 Results and Discussions

Comparison of the two models for the phosphorylation dynamics

Coarse-grained simulations of C-terminal domain (CTD) chains reveal that increasing phosphorylation levels ($r_{phosp}/r_{dephosp}$) drive a transition from a single large condensate to multiple smaller ones, eventually leading to complete dissolution at high phosphorylation levels. These findings underscore the sensitivity of condensates to phosphorylation dynamics and their ability to reorganize rapidly in response to changes. The simulation system initially comprised 200 CTD chains and 2 fully phosphorylated CTD (pCTD) chains, each modelled as a coarse-grained polymer with 140 amino acid residues. Phosphorylation and dephosphorylation events were introduced dynamically to explore the effects of post-translational modifications. Two distinct phosphorylation models were employed: the Processive Chain Switching Model and the Random-Uniform Switching Model (Fig. 4.1). At low phosphorylation levels, with a phosphorylation rate of $r_{np} = 0.1$ and $c_{np} = 0.1$ for the cases of Random-Uniform Switching and Processive Chain Switching Model respectively (Fig 4.1C), CTD chains predominantly remain in their neutral state, promoting strong inter-chain interactions that favour the formation of a single large condensate. This behaviour is consistent across both the Processive Chain Switching and Random-Uniform Switching models. However, as phosphorylation levels increase, the system undergoes fragmentation into multiple smaller condensates, as observed at $r_{np} = 0.5$ and $c_{np} = 0.5$ Fig. 4.1E. This fragmentation arises from phosphorylated regions altering interaction propensities, disrupting favourable contacts between protein chains.

Interestingly, at an intermediate phosphorylation level, with a phosphorylation rate $r_{np} = 0.3$ and $c_{np} = 0.3$, clear differences are observed between the two models. In the random-uniform switching model, the system remains as a single large condensate (Fig. 4.1D). However, in the Processive chain switching model, the system splits into two distinct condensates. The

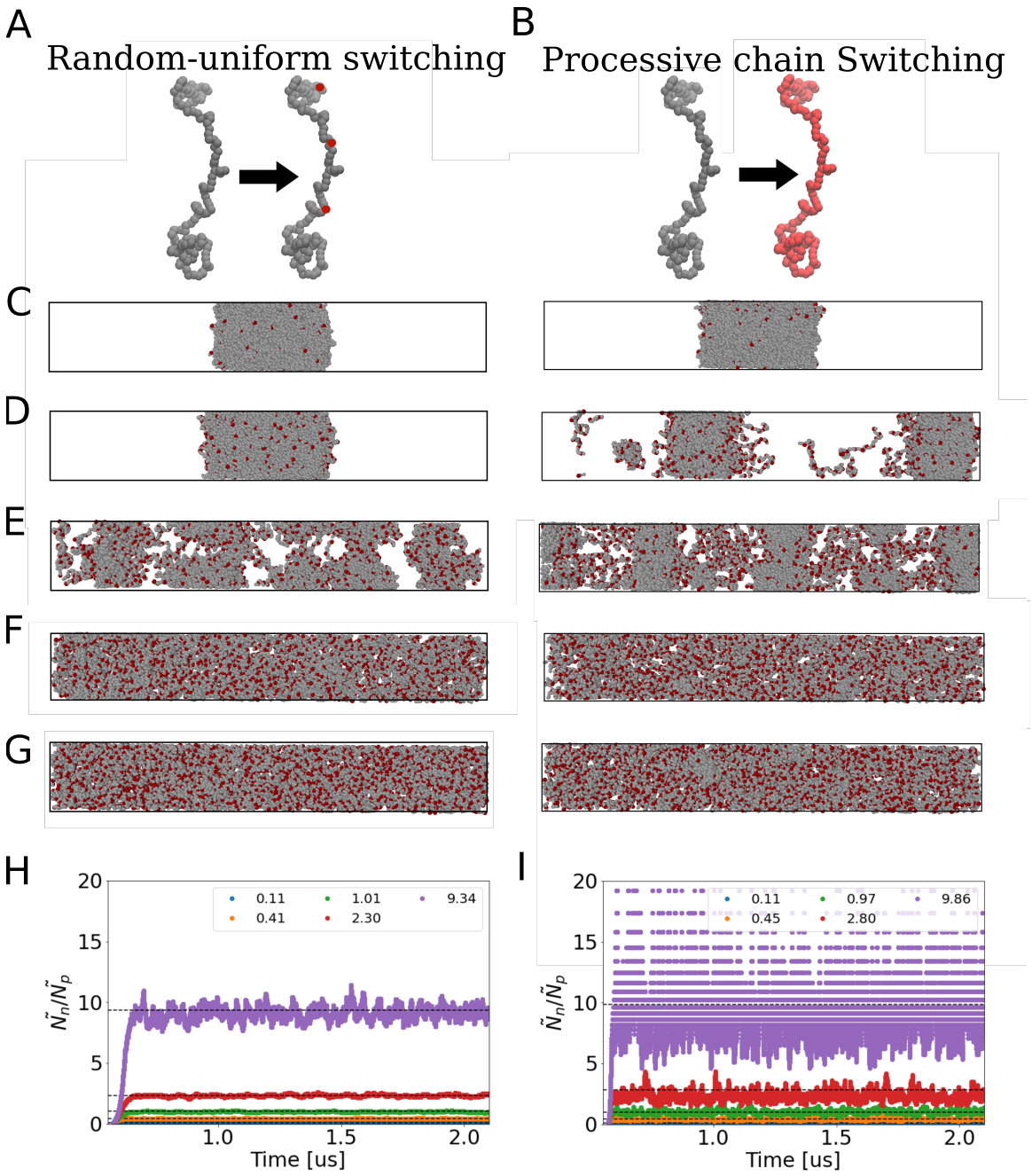

Fig. 4.1: A) Schematic of random-uniform switching Model and B) Processive chain Switching Model. C-G) Coarse-grained simulations of C-terminal domain (CTD) chains varying the relative phosphorylation and dephosphorylation rates. Simulations demonstrating the effect of phosphorylation level on condensate formation. The system's behaviour transitions from a single large condensate to multiple smaller condensates as the phosphorylation rate increases. At a higher phosphorylation level, the condensates dissolve. Simulations on the left illustrate the random uniform switching model, where the residues get phosphorylated randomly. Processive chain Switching Model simulations are shown on the right side. H-I) Time-series plot showing the fluctuations in the number of phosphorylated and dephosphorylated residues/chains. The system oscillates around an equilibrium value.

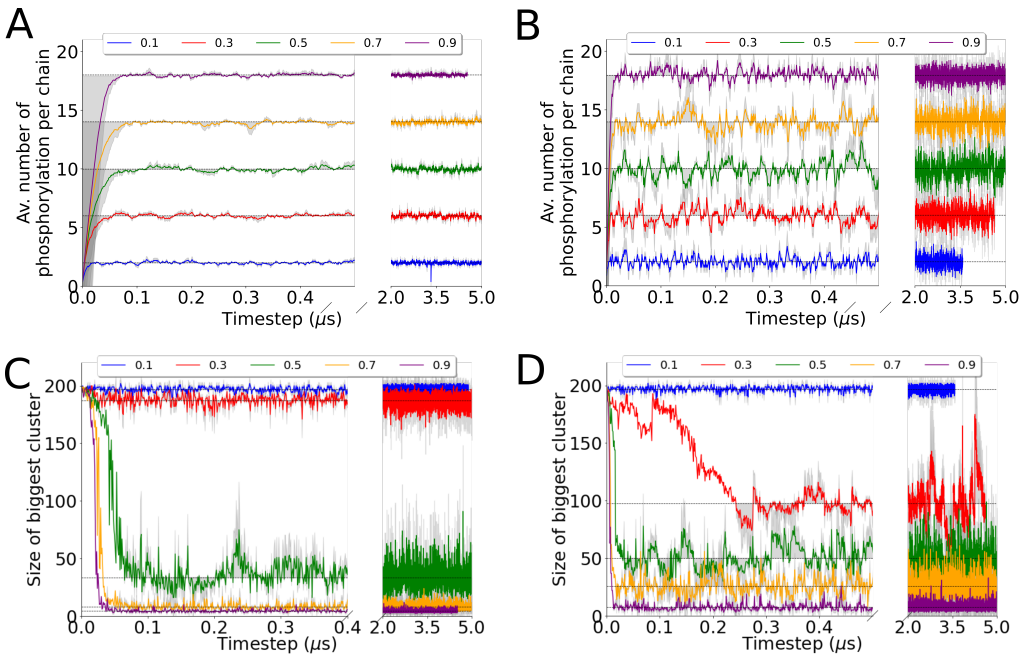

Fig. 4.2: Stochastic switching of phosphorylation and dynamics of the size of the largest cluster. Average number of phosphorylations per chain over time. A) Processive chain switching. B) random-uniform switching. Phosphorylation occurs more rapidly per chain in Processive chain switching, reaching equilibrium faster compared to random-uniform switching C) Size of the largest cluster over time with phosphorylation and dephosphorylation dynamics as the time step increases. D) Processive chain switching: consistent with Figure 1A, the condensate splits into multiple clusters at $c_{np} > 0.1$. D) random-uniform switching: consistent with Figure 1B, multiple clusters form beyond $r_{np} = 0.3$.

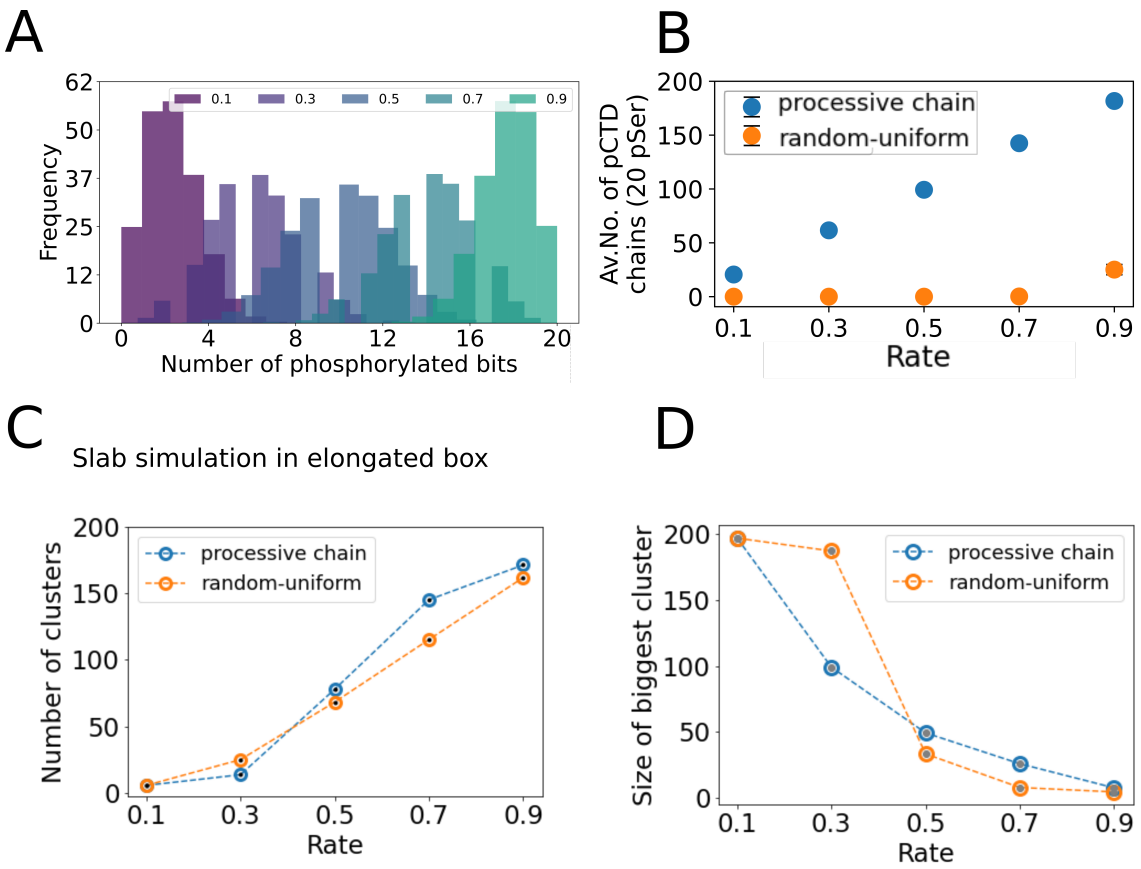

Fig. 4.3: Analysis of phosphorylation states and cluster formation in CTD condensates. A) Simulations were conducted in an elongated box with 200 chains. The plot shows the distribution of phosphorylation along the CTD in the random-uniform switching Model. Histogram showing the number of phosphorylated bits per chain as the phosphorylation rates increase. The distribution shifts towards higher phosphorylation levels with increasing r_{np}/r_{pn}. B) The plot demonstrates the increase in the average number of fully phosphorylated chains (20 pSer per chain) as the rates increase in both the random-uniform and processive chain switching. In the processive chain switching model, the average number increases linearly. In the random-uniform switching model, fully phosphorylated chains are observed at higher phosphorylation rates (0.7 and 0.9), with a significant increase at 0.9. C) Cluster analysis from simulations in elongated boxes shows that the number of clusters formed in both Processive chain and random-uniform switching models increases with phosphorylation rates. D) The size of the biggest cluster formed decreases as the phosphorylation levels increase in both cases.

differences in behaviour between the two models at an intermediate phosphorylation rate can be attributed to several factors, including the mechanisms of phosphorylation and the role of charge asymmetry in protein interactions. In the Processive chain switching model, where phosphorylation is treated as a coordinated event affecting entire chains, the introduction of multiple negative charges can significantly alter the interactions between these chains and the unphosphorylated CTD chains. The increased electrostatic repulsion among negatively charged phosphorylated bits can lead to fragmentation of the large condensate. Some highly phosphorylated chains are not part of any protein cluster but are in an aqueous solvent. Conversely, in a random-uniform switching model, where phosphorylation occurs independently at each bit, the overall charge distribution remains more balanced, allowing strong interchain interactions to persist, enabling the system to maintain a single condensate.

At a higher phosphorylation level, with a phosphorylation rate $c_{np} = 0.5$ and $r_{np} = 0.5$ and above, both the processive chain and random-uniform switching models result in the system splitting into multiple smaller condensates. The increased rates lead to a higher concentration of negatively charged phosphate groups on the protein chains. This charge asymmetry enhances electrostatic repulsion between phosphorylated regions, making it energetically unfavourable for the chains to remain aggregated in a single large condensate. As a result, the system tends to fragment into smaller, more stable condensates to minimize repulsive interactions. Both models exhibit a saturation effect at high phosphorylation levels. In the random-uniform switching model, independent phosphorylation events still occur, but as more bits become phosphorylated, the overall charge distribution becomes increasingly imbalanced. This imbalance contributes to the destabilization of larger condensates. At higher phosphorylation rates ($r_{np} > 0.5$ and $c_{np} > 0.5$), both the processive chain and random-uniform switching models lead to the disruption of the entire structure of biomolecular condensates. In both models, this repulsive force becomes significant enough to overcome the attractive interactions that stabilize larger condensates, leading to the dissolution of the entire structure.

The interplay between phosphorylation and dephosphorylation establishes a dynamic equilibrium, as shown in the time-series plot of phosphorylated and unphosphorylated chains (Fig. 4.1H-I). This plot reveals a dynamic equilibrium where the ratio of neutral and phosphorylated entities oscillates around a mean value. The equilibrium mean value corresponds to the ratio of the rate of phosphorylation and dephosphorylation. These oscillations reflect the continuous competition between phosphorylation and dephosphorylation processes, capturing the dynamic nature.

As we observe the dynamic equilibrium between phosphorylation and dephosphorylation, it becomes clear that this balance significantly affects the structure of biomolecular condensates over time. To further understand the impact of phosphorylation dynamics on condensate structure, we analyzed the size of the largest cluster as a function of phosphorylation rate. In random-uniform switching, the condensate remains as a single large cluster until the

phosphorylation rate r_{np} reaches 0.3. Beyond this point, multiple smaller clusters begin to form due to the independent phosphorylation of individual bits on the chains. Initially, this independent phosphorylation maintains a more balanced charge distribution, reducing overall electrostatic repulsion. However, as the phosphorylation rate increases, the charge imbalance grows, leading to increased electrostatic repulsion (Fig. 4.2D-G). In contrast, the Processive Chain Switching model exhibits fragmentation into smaller clusters at a lower phosphorylation rate of $c_{np} = 0.1$ due to the rapid introduction of negative charges across entire chains.

The mechanism of phosphorylation is a key aspect that distinguishes the processive chain Switching and random-uniform switching models, even though the overall phosphorylation level can be the same in both cases. In the processive chain switching model, phosphorylation occurs quickly because it affects entire chains simultaneously. This coordinated approach rapidly introduces negative charges, leading to a faster equilibrium (Fig. 4.2B). The swift introduction of these negative charges accelerates the system's transition to a stable state, where the number of phosphorylated entities stabilizes quickly. In contrast, the Random-Uniform Switching model involves the independent phosphorylation of individual sites, resulting in a slower accumulation of negative charges. Consequently, the system takes longer to reach equilibrium, with the number of phosphorylated residues increasing more gradually over time, as illustrated in Fig. 4.2A. It is important to note that eventually, at equilibrium, the average number of phosphorylated residues per chain is the same in both models (Fig. 4.2A, B). However, fluctuations around the mean are more pronounced in the Processive Chain Switching model, reflecting its coordinated phosphorylation mechanism.

Droplet size and shapes in simulation of cubic boxes

To characterize how phosphorylation kinetics influence the size and shape of protein droplets, we extended our analysis to simulations conducted in a cubical box setup. The simulations were conducted with CTD chains, with snapshots showing the equilibrated system at different phosphorylation levels Fig. 4.4A, with the random-uniform switching model. As the phosphorylation rates increase, the system transitions from a single large condensate to multiple smaller condensates.

In the random-uniform switching model (Fig. 4.4B), the condensate remains largely intact at lower phosphorylation rates (e.g., 0.1 and 0.3). However, once the phosphorylation rate exceeds 0.3, the cluster size decreases significantly, leading to the formation of smaller condensates. This behaviour aligns with our earlier slab simulations, where independent phosphorylation events initially maintain a balanced charge distribution. In contrast, the Processive Chain Switching model exhibits a more immediate response. At a moderate phosphorylation rate $c_{np} = 0.3$, the cluster size decreased significantly due to the coordinated phosphorylation of entire chains, similar to slab simulation.

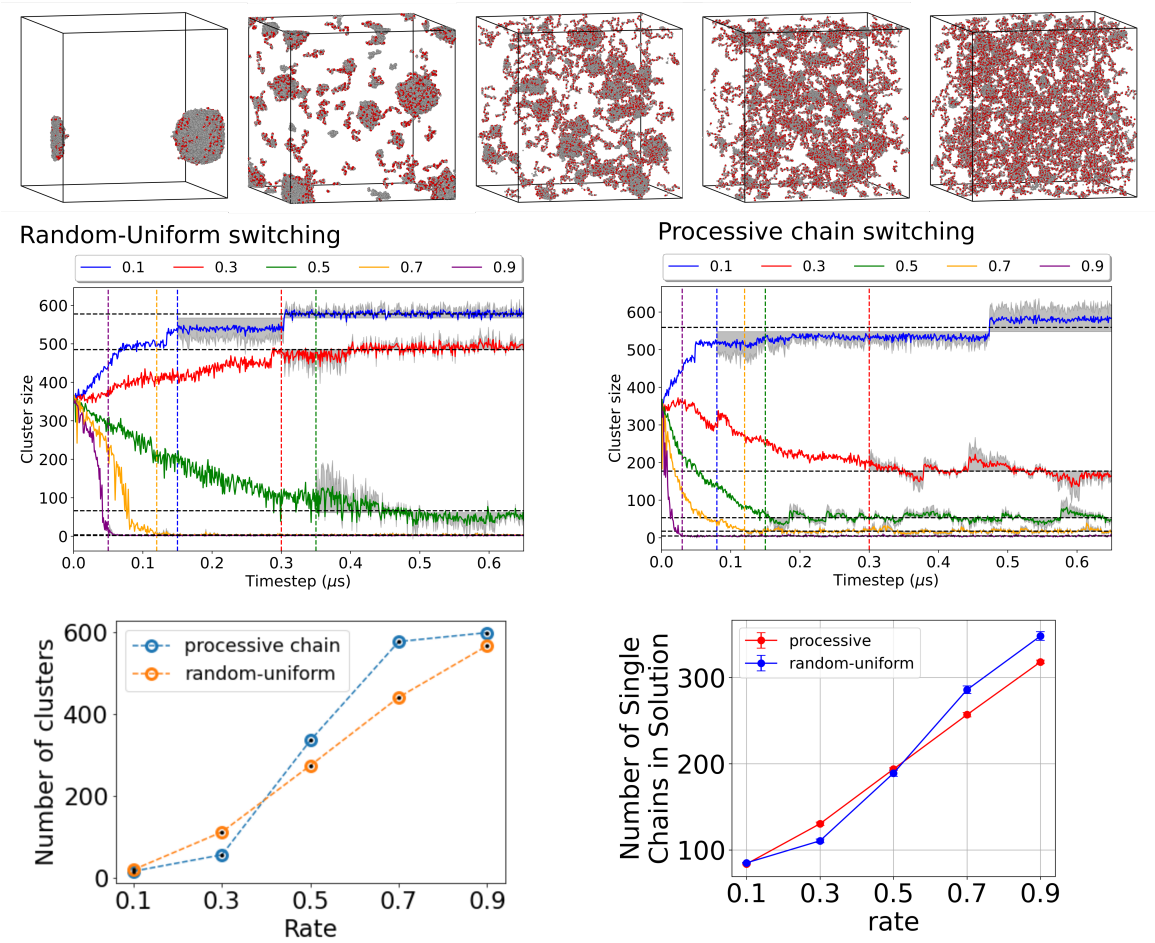

Fig. 4.4: Characterisation of droplets of CTD and pCTD in cubic box simulations. A) Simulation Setup: Simulations are conducted in a cubical box containing 300 CTD chains (grey) and 300 fully phosphorylated CTD (pCTD) chains (red). The snapshots show the equilibrated system at different levels of phosphorylation. As the rates increase, the system transitions from one large condensate to multiple smaller condensates. B) Cluster Size at random-uniform switching: This plot shows the size of the largest cluster at different phosphorylation rates in the random-uniform switching model. C) Cluster Size at processive chain Switching: This plot shows the size of the largest cluster at various phosphorylation rates in the processive chain Switching model. In contrast to the random-uniform switching model, the cluster size in the processive chain D) The plot shows the increase in the number of clusters as the rates increase. E) Number of free chains in the dilute phase.

The total number of clusters also increased in both models as the phosphorylation rate increased(Fig. 4.4D). Additionally, Fig 4.4E shows that the number of CTD chains in the dilute solution rises as the phosphorylation rate increases. These findings again support that higher phosphorylation levels disrupt cluster formation, leading to more chains outside any cluster in the dilute phase.

Phosphorylation dynamics of a multiphasic condensate

To further understand the complex interactions and phase behaviour of biomolecular condensates, we simulated a system consisting of CTD and histidine-rich domains (HRD), with the CTD being phosphorylated stochastically. The charge alterations caused by phosphorylation play a critical role in cellular processes. However, during transcription, phosphorylated CTD also functions as an interaction interface with other proteins, such as the HRD of Cycline T1. In these simulations, we started with CTD and HRD components and applied the processive chain Switching model to phosphorylate the CTD chains. We took simulation snapshots to observe the equilibrium steady state at various phosphorylation rates (Fig 4.5). The shell thickness around the CTD condensate increased with a higher phosphorylation rate. This indicates that phosphorylating CTD chains leads to a more extensive and dense shell formation, as pCTD and HRD interact via electrostatic potential. At a high rate of 0.9, the system changed from a multiphasic state to a single-phase state. It is also important to note that the CTD condensate, which forms the core of the system, has an irregular shape when phosphorylation kinetics are active. In contrast, when the kinetics are turned off, the CTD condensate has a more regular shape. This irregularity under active phosphorylation indicates that the dynamic introduction of negative charges disrupts the cohesive forces within the CTD condensate. Fig. 4.6, shows the asphericity of CTD condensates as a function of rate, comparing scenarios with and without kinetics. At lower ratios c_{np}/c_{pn} (0.1, 0.3, and 0.5), the asphericity values are similar for both conditions. However, at higher rates (0.7 and 0.9), the asphericity increases for both cases, with condensates formed with kinetics exhibiting greater asphericity than those formed without kinetics. This indicates that including kinetic effects leads to more irregular or non-spherical shapes in the condensates at higher rates.

We analyzed the cluster size of CTD for different phosphorylation rates to quantify the effects of phosphorylation on the structure of the CTD condensate. At lower phosphorylation rates, CTD chains form larger clusters. However, as the phosphorylation level increases, the introduction of negative charges leads to enhanced electrostatic repulsion, causing the clusters to break apart into smaller clusters.

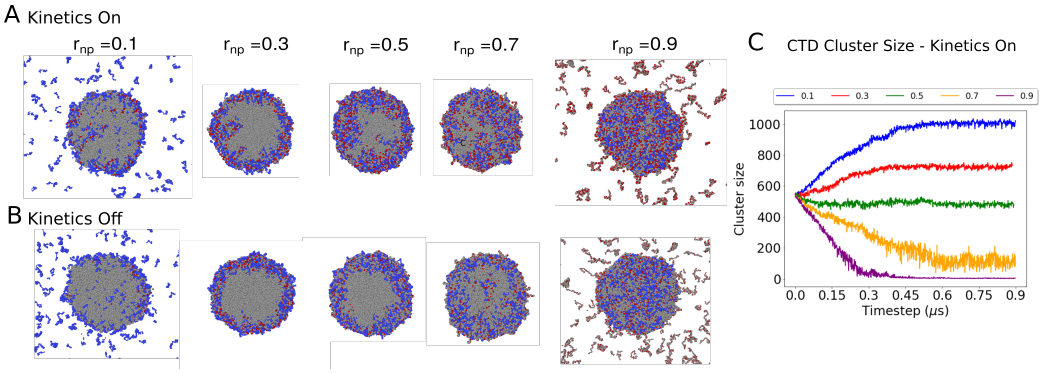

Fig. 4.5: A) Simulation Setup: Simulations of a system consisting of CTD and HRD domains, with CTD phosphorylated stochastically following the processive chain Switching model. Snapshots show the equilibrium steady state at various phosphorylation rates. As the rates increase, the thickness of the shell around the CTD condensate increases. C) Simulations with the same system at equilibrium and with kinetics off. The CTD condensate exhibits a more regular shape when kinetics are off. C) Cluster size of CTD analyzed for different rates of phosphorylation. Higher rates lead to smaller cluster sizes.

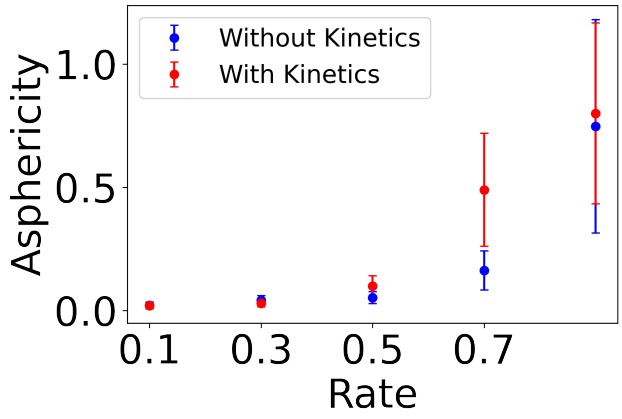

Fig. 4.6: Asphericity of CTD condensate with and without kinetics as a function of phosphorylation rates. Higher asphericity values for condensate formed with kinetics compared to those without kinetics, particularly at rates of 0.7 and 0.9. This indicates that kinetic effects lead to more non-spherical condensate at higher rates.

4.4 Conclusion

We introduced a simple and efficient approach to simulations of phosphorylation dynamics of protein condensates. Phosphorylation and Phosphorylation can be simulated by explicitly representing enzymes in coarse-grained molecular dynamics simulations. However, simulations of long duration are required to observe how a condensate responds to a given phosphorylation level. By modelling the phosphorylation and de-phosphorylation reactions as kinetic rate processes without 3D representations of the enzymes, we can efficiently capture how indices respond to changes in protein phosphorylation levels.

The two chemical reaction models we considered represent extreme cases: one assumes no correlation between successive reactions, while the other models the behavior of a maximally processive enzyme. The processive model leads to larger fluctuations in the number of phosphorylated Ser residues and greater variation in the size of clusters, especially at intermediate phosphorylation-to-dephosphorylation ratios. To make the processive chain switching model less extreme, we could set the probability of an individual residue reacting to a value smaller than 1. This would allow for partial modification during a single processive event. The switching between events where droplets are present and where they dissolve has been proposed to regulate transcription initiation [214]. Droplets would recruit the transcription machinery, supporting transcription initiation. These switches between droplet formation and dissolution could also help explain the experimentally observed bursts in gene transcription.

Including chemical-driven phosphorylation in our simulations, we observe similar morphologies as in equilibrium simulations of condensates of phosphorylated and unphosphorylated CTD as well as HRD of the elongation factor Cyclin T1. Our simulations show that high levels of phosphorylation make droplets less spherical, which can be clearly seen for droplets of CTD, pCTD and HRD and also for CTD and pCTD droplets, but these droplets split and dissolve at larger ratios of phosphorylation to de-phosphorylation. Interestingly, droplets become more spherical in *C. elegans* as Cyclin T1 expression is inhibited by RNA interference (RNAi), which leads to a loss of HRD and reduction of CTD phosphorylation [221]. There are likely additional proteins with positively charged disordered regions which support transcription elongation by interacting with pCTD [174]. While there is not an exact correspondence between simulations and in vivo experiments, it is still interesting to note that lower phosphorylation levels are associated with more spherical droplets in simulations and in vivo. It is important to note that the changes in CTD droplet shapes in *C. elegans* could also be due to changes in interactions with many other proteins as well as direct and indirect interactions with RNA and DNA, which are beyond the scope of this study.

5

Sequence determinants of protein phase separation through molecular dynamics and active learning

Title: "Sequence determinants of protein phase separation and recognition by protein phase-separated condensates through molecular dynamics and active learning"

Contributions:

- **Simulation setup and execution**: I was responsible for setting up and executing simulations for this study. This involved writing scripts for the three kinds of sequence optimisation analysis.
- **Data analysis**: I analyzed the simulation trajectories to calculate the order parameters for optimisation.
- **active learning workflow**: I helped implement the active learning workflow with the MD simulations. I was responsible for model training and optimisation methodology.
- **Figure preparation**: I created the plots shown in the paper.
- **Manuscript contribution**: I contributed to writing and revising the manuscript.

Reprint permissions:

Reprinted from " Sequence determinants of protein phase separation and recognition by protein phase-separated condensates through molecular dynamics and active learning A. Changiarath, A. Arya, V. A. Xenidis, J. Padeken and L. S. Stelzl, Faraday Discuss., 2025, 256, 235 DOI: 10.1039/D4FD00099D," with the permissions from the RSC Publishing.

Faraday Discussions

Cite this: *Faraday Discuss.*, 2025, **256**, 235

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Sequence determinants of protein phase separation and recognition by protein phase-separated condensates through molecular dynamics and active learning†

Arya Changiarath, [a] Aayush Arya,[a] Vasileios A. Xenidis, [b] Jan Padeken[c] and Lukas S. Stelzl *[cde]

Received 10th May 2024, Accepted 26th July 2024
DOI: 10.1039/d4fd00099d

Elucidating how protein sequence determines the properties of disordered proteins and their phase-separated condensates is a great challenge in computational chemistry, biology, and biophysics. Quantitative molecular dynamics simulations and derived free energy values can in principle capture how a sequence encodes the chemical and biological properties of a protein. These calculations are, however, computationally demanding, even after reducing the representation by coarse-graining; exploring the large spaces of potentially relevant sequences remains a formidable task. We employ an "active learning" scheme introduced by Yang *et al.* (*bioRxiv*, 2022, https://doi.org/10.1101/2022.08.05.502972) to reduce the number of labelled examples needed from simulations, where a neural network-based model suggests the most useful examples for the next training cycle. Applying this Bayesian optimisation framework, we determine properties of protein sequences with coarse-grained molecular dynamics, which enables the network to establish sequence–property relationships for disordered proteins and their self-interactions and their interactions in phase-separated condensates. We show how iterative training with second virial coefficients derived from the simulations of disordered protein sequences leads to a rapid improvement in predicting peptide self-interactions. We employ this Bayesian approach to efficiently search for new sequences that bind to condensates of the disordered C-terminal domain (CTD) of RNA Polymerase II, by simulating molecular recognition of peptides to phase-separated condensates in coarse-grained molecular dynamics. By searching for protein sequences which prefer to self-interact rather than interact with another protein

[a]*Institute of Physics, Johannes Gutenberg University (JGU) Mainz, Germany*
[b]*Department of Biology, Aristotle University of Thessaloniki, Greece*
[c]*Institute of Molecular Biology (IMB) Mainz, Germany*
[d]*Institute of Molecular Physiology, Johannes Gutenberg University (JGU) Mainz, Germany. E-mail: lstelzl@uni-mainz.de*
[e]*KOMET1, Institute of Physics, Johannes Gutenberg University (JGU) Mainz, Germany*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4fd00099d

sequence we are able to shape the morphology of protein condensates and design multiphasic protein condensates.

1 Introduction

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Proteins and other biomolecules within cells can phase separate to form liquid-like droplets.[1] Such phase-separated droplet regions provide a unique microenvironment essential for several important cellular processes and thus are an important way of regulating biological function. The human proteome has a very rich diversity in terms of composition.[2] A predictive insight into which protein sequences could undergo phase-separation would improve understanding of the pathogenesis of several neurodegenerative diseases including Alzheimer's disease,[3,4] where protein phase separation and the formation of toxic aggregates of these proteins have been linked to pathogenesis. Different condensates or condensates with multiple distinct phases provide a way for biological processes to be regulated in time and space.[5–7] Spatially separating different biochemical reactions is also a promising blueprint for biotechnological applications.[8] This necessitates efficient strategies to design novel phase separating condensates and to develop molecules, such as artificial proteins, to bind selectively to phase-separated condensates. However, in other contexts such as designing protein interaction modules for synthetic biology,[9] one might specifically want to reduce the propensity of proteins to interact and form condensates.

Molecular dynamics (MD) simulations have played a major role in our understanding of protein–protein interactions and protein phase separation.[1,10–12] The development of state-of-the-art molecular force fields has enabled accurate representation of molecular interactions, allowing these simulation models to successfully predict phase separation propensities of diverse protein sequences.[11,13] Atomistic molecular dynamics simulations can provide essential insights into molecular driving forces by simplifying the description. However, building an exact quantitative understanding of a protein sequence–property relationship, to be able to predict the behaviour of a sequence without the necessity of a simulation or an experiment, has remained difficult to achieve. This is in part due to the large space of possible biological sequences, and also how sequence context will change the interaction characteristics of the amino acids.

One way to model such complicated relationships is *via* machine learning, and recently, important progress has been made in learning such relationships and also applying this progress to the interactions and phase behaviour of flexible and disordered proteins.[14–17] A very successful strategy has been to combine physics-based simulations with machine learning and artificial intelligence.[16,18,19] Ginell *et al.* have trained a neural network with a physics-based simulation model to predict how different disordered proteins will interact.[14] With a physics-based simulation model von Bülow *et al.* have successfully trained a neural network to predict the phase separation propensities of proteins.[15] In recent years, protein language models have been shown to organise protein sequences in biologically meaningful ways.[20–23] Language models also seem to encode useful physical and chemical information, which makes it imperative to see how we can leverage the information encoded in such models to predict protein behaviour. Language models have been used in iterative design[24] and for proposing entirely novel proteins.[23] However, a key difficulty

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

in mapping sequence–property relationships is the limited availability of "labelled" data needed for training. The property of interest may require experimental input or computationally expensive simulations. One must also note that not all labelled data is equally useful, and some specific examples may be more beneficial for training than others.

Active learning[27] is a machine learning paradigm meant to address exactly this issue, where the model interacts with the human to suggest which labelled examples should be provided for training. Within such a framework, Bayesian Optimisation (BO) can be invoked to rank the most useful sequences in terms of their potential utility to the model and thus can be used to judiciously direct one's resources. These approaches can be utilised to understand how chemical structure and biological sequences give rise to chemical[28] and biological properties. Recently, An *et al.*[29] used active learning to discover strongly interacting proteins that tend to phase-separate, while optimising the trade-off between thermodynamic stability and transport properties. In another context, it was demonstrated that active learning helps to design molecules selectively interacting with one phase of a lipid membrane over another phase of this membrane[30] and to find small molecules binding to a protein.[31] Chemical structure and biological sequences have to be inputted in a way that is amenable to machine learning. One possibility is to combine various metrics. Deep learning provides a way of achieving this and promising results based on autoencoder have been obtained.[30] For biological sequences, language models may be a natural way of representing

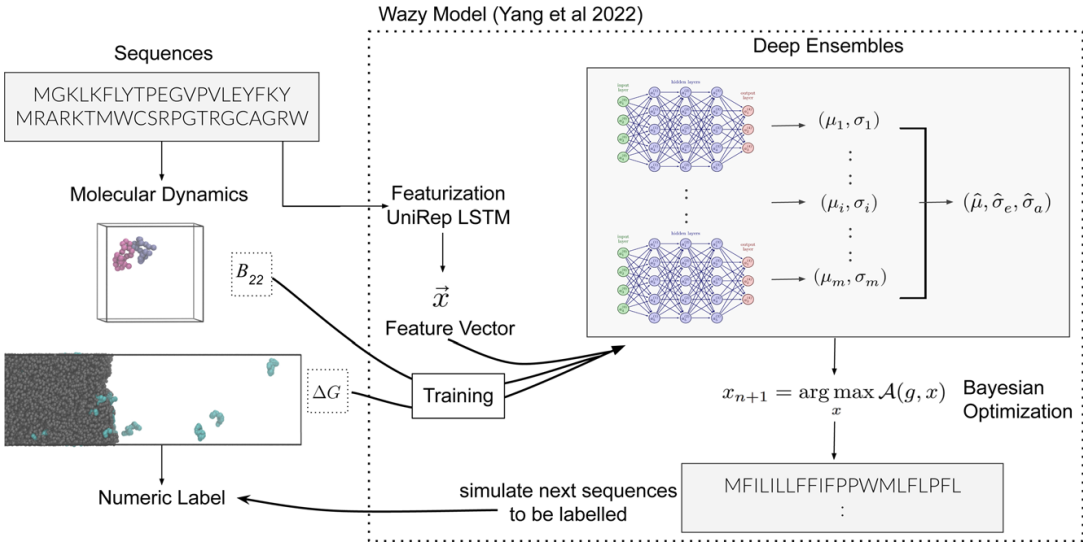

Fig. 1 Schematic overview of the active learning framework for predicting peptides using Bayesian Optimisation (BO) and Molecular Dynamics (MD) simulations: (A) MD simulations of protein chains to compute second virial coefficient (B_{22}) and partition free energy (ΔG) for model training. This initial sample is fed to the `wazy` BO package,[25] which utilises UniRep LSTM,[24,26] to parse the protein sequence into a feature vector $\vec{x}$. Deep ensemble training on feature vectors and computed properties to establish sequence–property relationships. To choose which sequences the model would benefit the most from, an acquisition function $\mathcal{A}$ is used to rank and suggest the next set of sequences. The suggested sequences are simulated again, generating new data for subsequent training cycles. This iterative process helps the model learn the sequence–property relationship with fewer examples and simultaneously explore sequence space regions that contain proteins with the most extreme values of the property we studied.

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

sequences and recent advances make it possible to use pre-trained language models for featurisation.[24,25] Language models show a lot of promise for BO, as exemplified by a recent study by Hoffbauer and Strodel,[17] where the authors showed that the biochemical knowledge encoded in an advanced pre-trained language model enables efficient active learning even when only limited data is available. Yang *et al.* have developed a very promising workflow based on a pre-trained language model[25] and deep ensembles to express the learned relationships. Here we investigated three related questions: (1) Can we identify the sequence determinants which lead to protein self interactions and phase separation? (2) Can we understand and design new sequences which selectively bind to phase-separated protein condensates? (3) Can we design multiphasic condensates? To address these important questions in biophysics and physical chemistry we utilize the BO approach by Yang *et al.*[25] (Fig. 1).

2 Methods

We study the mapping of a protein sequence to one of its properties, as represented by a numeric quantity $f(\vec{x})$. As we shall describe later, $f(x)$ could be the second virial coefficient B_{22}, or another quantity that provides an estimate of the affinity of a sequence to bind to a condensate. Our scientific problem thus reduces to training a model to learn the mapping between $\vec{x}$ and $f(\vec{x})$, while simultaneously exploring the protein sequence space to maximise this label. We utilize the `wazy` package[25] which provides the framework for performing such optimisation tasks implemented here.

Fig. 1 illustrates our workflow: We begin with an initial set of protein sequences and conduct coarse-grained molecular dynamics simulations to extract for instance, B_{22} or ΔG values. We initialize a surrogate model within the protein sequence Bayesian Optimisation (BO) package `wazy`[25] and provide it with an initial set of sequences and labels to calibrate the initial sequence–property mapping. During that, `wazy` uses UniRep—a multiplicative Long-Short Term Memory (mLSTM) model[24,26] for parsing the amino acid sequence into a feature vector x that contains the essential protein features in a fixed-length numeric vector.

UniRep, trained on 24 million primary amino acid sequences in the UniRef50 database, has shown the ability to accurately partition structurally and evolutionarily similar proteins, even when they share little sequence identity.[24] The resulting feature vectors, combined with the computed labels, serve as input for training the surrogate models, which in the context of `wazy` are ensembles of multi-layer perceptrons (MLPs). Such *deep ensembles* are appropriate for uncertainty quantification and offer greater expressive power compared to the Gaussian processes (GPs) that are otherwise typically chosen as surrogates.

Following this initial calibration of sequence–property mapping, the Bayesian optimiser uses an *acquisition function* to rank which proteins would serve as the most useful examples for further training. This process simultaneously explores the sequence space, in search of regions with proteins that possess the highest values of the property we are studying. A batch of sequences is drawn from this for the next cycle of training. Coarse-grained simulations of these sequences are again performed, and the entire process is repeated for several iterations. In the subsequent sections, we provide a detailed account of the training and optimisation methodology.

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

2.1 Model description and optimisation

As mentioned before, the surrogate model in `wazy` is an ensemble of feedforward neural networks (called a *deep ensemble*[32]). In such a model, the quantity of interest $f(x)$ is independently predicted by not one but an ensemble of neural network models, which in this case is a Multi-Layer Perceptron (MLP). In the configuration used here, the ensemble comprises of 5 MLPs. UniRep[24] parses an input sequence into a fixed-length feature vector of dimension $N = 1900$. Therefore, each individual MLP takes an input of dimension 1900, followed by three layers having 128, 32, and 2 neurons each. The last layer outputs two numbers: a mean $\hat{\mu}$ and a standard deviation $\hat{\sigma}$ that characterise a normal distribution $\mathcal{N}(\hat{\mu}, \hat{\sigma})$. The dispersion in the mean values $\hat{\mu}$ predicted by the individual models, and the sum of the variances $\hat{\sigma}^2$ of this $\mathcal{N}$ by the different models, together give a total uncertainty estimate. All networks layer but the last one is followed with the SiLU (also known as Swish) activation function. In addition, standard techniques such as dropout and weight decay are utilised to reduce overfitting. For further technical description about `wazy`, we refer the reader to Yang *et al.*, who designed it.[25]

The advantage of deep ensembles is that neural networks possess more expressive power than Gaussian processes, while providing a way of uncertainty quantification. Bayesian optimisation works *via* ranking sequences using an acquisition function $\mathscr{A}$. Unless stated otherwise, for most of the work presented here, we chose the "upper confidence bound" (UCB) acquisition function[33] which balances both exploitation of the known maximas and exploration of the parameter space:

$$\mathscr{A}(x;\lambda) = \mu(x) + \lambda\sigma(x) \tag{1}$$

here, the exploitation term $\mu(x)$, represents the current estimate of the value known to the surrogate, while the exploration term $\sigma(x)$ quantifies the uncertainty. The parameter λ acts as a control to adjust the balance between exploitation and exploration. λ is set to a value 2.

Another commonly used acquisition function is Expected Improvement (EI). The fundamental principle behind EI is to select the next point to evaluate based on the current best result. The choice of the acquisition function sensitively affects which regions of sequence space get explored for, and thus the arrival at, an optimal solution. Later, in Section 3.1, we will show the impact of this choice as observed in our study.

2.2 Training and validation

Optimisation of the surrogate model—by asking it to suggest sequences to be measured in a lab or simulated—requires an initial calibration. For most of the tasks studied here, an initial set of sequences was drawn from ProtGPT2,[23] a protein language model constructed for *de novo* protein design. Quantities $f(x)$, either B_{22} or ΔG values, were computed from coarse-grained molecular dynamics. This computed value was treated as the *true* or *expected* value of the property of interest. The length of all sequences used for training was chosen to be 21 amino acids.

While training, to track the learning of the model, we set aside two sets of sequences for *validation* of the model's predictive ability, which were never used

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

for training. The first set was drawn the same way as the training data itself, with 42 21-amino acid length sequences drawn from the ProtGPT2 language model,[23] spanning a wide range of B_{22} values. To track how well the approach captures strongly interacting peptides which are likely only a small fraction of natural protein sequences, we generated a second set of validation data. We took 22 21-amino acid sequences with very negative B_{22} values from the study of An *et al.* (2024)[29] (see their ESI Tables). The true values of B_{22} for these validation sequences were computed in exactly the same way, from coarse-grained simulations. The coarse-grained simulations for calculating B_{22} were conducted for 10 μs.

For designing peptides that bind to condensates, an initial calibration set of 70 21-mer peptides was selected from An *et al.* (2024). For validation, 30 peptides were chosen from ProtGPT2. The simulations involving the condensate and peptides were performed for 0.6 μs.

Similar to Yang *et al.*,[25] we will refer to an "optimisation step" as being the usage of one sequence-label pair after which the model is updated. Subsequent to each optimisation step, we predicted the numeric label for the validation set sequences and tracked the mean squared error from the residual between true (computed from simulations) and values predicted by the model. The exact size of the calibration and validation sets were chosen separately for the different tasks studied here (described in their respective subsections). In all that follows, we shall refer to the initial calibration as *iteration* 0.

2.3 Specifications of the coarse-grained molecular dynamics simulations

Residue-level coarse-grained simulations with the HPS model[34] were run in HOOMD-blue v. 2.9.6 [35] extended with azplugins (v. 2.9.2). Simulations were also conducted with the more recent CALVADOS2 simulation model.[13] The second virial coefficient (B_{22}) was calculated from the simulations of two 21-amino acid peptide sequences in a cubic box of dimension 8 nm × 8 nm × 8 nm. To determine the partition free energy, condensate simulations were conducted in a slab geometry of dimension 15 × 15 × 250 nm. The condensate was simulated with 300 chains of 140-amino acid CTD of RNA Polymerase II, and 30 chains of 21-amino acid peptides were added to the preformed condensate.

2.4 Computing the second virial coefficient from coarse-grained simulations

The second virial coefficient B_{22} is used as a measure of protein self interactions. This is well-reasoned, as B_{22} is an estimate of the intermolecular forces between two molecules in dilute solutions.[34] It is well established that a negative B_{22} correlates strongly with the tendency to phase separate, an idea that led An *et al.*[29] to similarly use B_{22} as a proxy label. We compute it using equilibrium molecular dynamics simulations with durations of 10 μs. The second virial coefficient is defined as:

$$B_{22} = \int_0^\infty 2\pi r^2 [\exp(-U(r)/k_B T) - 1] \mathrm{d}r, \tag{2}$$

where $U(r)$ is the potential energy at radius r, k_B is the Boltzmann's constant and T the temperature.

We compute B_{22} from the radial distribution function $g(r)$

$$B_{22} = \int_0^\infty 2\pi r^2 [g(r) - 1] \mathrm{d}r. \quad (3)$$

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

2.5 Designing binders to a specific condensate

In our simulations, we investigate sequence-specific binding to condensate formed by the intrinsically disordered C-terminal domain (CTD) of RNA Polymerase II. Here we consider two types of CTD condensates: one composed of ideal 140-mer CTD sequences (each containing 20 ideal YSPTSPS heptad repeats), and another representing the CTD of human RNA Polymerase II.

To quantify the binding affinity of the designed peptides for the condensed phase, we calculated the partition free energy (ΔG) from the equilibrium densities in the coexisting dilute and dense phases. We calculated the partition free energy of the peptides to the condensate using the expression,[36] $\Delta G = -RT \log(\rho_{\mathrm{dense}}/(\rho_{\mathrm{dilute}} + \varepsilon))$, which became the numeric label $f(x)$ we wanted to train for. Here ΔG is the binding free energy, R is the gas constant, T is the temperature, ρ_{dense} is the density of the peptide in the dense condensate phase, and ρ_{dilute} is the density of the peptide in the dilute phase. ε is a small constant ($=10^{-5}$) added to the denominator to prevent the divergence of ΔG when peptides strongly bind to the condensate.

2.6 Protein sequence descriptors

In addition to predictability of B_{22} or ΔG by the model, we kept track of the amino acid diversity, sequence properties such as mean hydropathy, fraction of residues that contribute to disorder or chain expansion, the number of charged residues, *etc.* For this, the amino acid compositions and relevant descriptors were estimated with the `localCIDER` Python library from Holehouse *et al.*[37]

3 Results

As mentioned previously, we trained the model for several different tasks. In the following, we demonstrate the performance on each of these. Note that for each task, the model was trained independently from scratch. In Section 3.1, we show how the model successfully learnt to predict strongly self-interacting proteins, and suggest sequences that are likely to phase separate. In Section 3.3, utilising a different property—namely ΔG, which gives the binding affinity—we trained to find peptides that like to bind with pre-formed condensates, using CTD of the RNA Polymerase II as a test case. The method can be extended to multi-phasic condensates by defining a suitable fitness function to shape the morphologies of multi-phasic condensates. The success of the model on its ability to perform on these very different tasks with limited training data shows the generalised applicability of the method.

3.1 Sequence-determinants of protein self-interaction from active learning

As the training progresses, the deep ensemble (Fig. 1) quickly learns the relationship between protein sequences and the second virial coefficient B_{22}, quantifying a protein's tendency to interact with another chain with identical

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

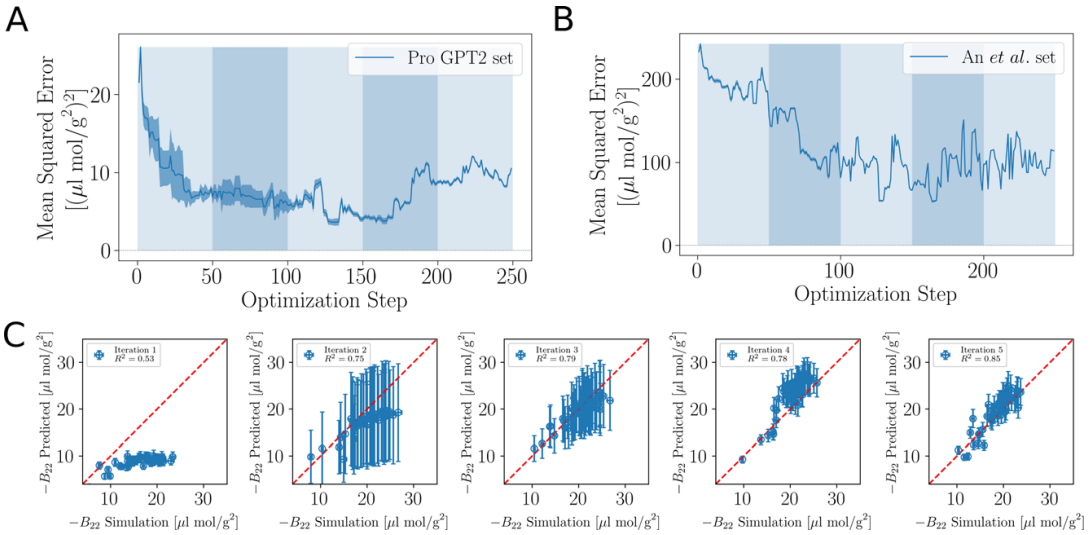

Fig. 2 Accuracy of the learned relationship between protein sequence and B_{22}. (A) Feedback between ML model and simulations. The shaded areas highlight the transitions between the sequences 0–50, 50–100, 100–150 and 150–200 in the first, second, and third iterations of the active learn procedure. The validation set is generated with ProtGPT2. (B) The same but with a validation set which includes more strongly interacting sequences from An *et al.* (C) Predicted B_{22} from the active learning and computed from coarse-grained molecular dynamics. After the network was trained with $N = 50$ sequences. At the second iteration with a total of $N = 100$. Third iteration with $N = 150$ sequences (Fig. S1†).

sequence. In Fig. 2A, we see a rapid decline in the mean squared error (MSE) in its predictions for B_{22}, on validation sequences the model has never seen during training, with the MSE dropping from more than 25 to just above 5. The alternating shaded regions mark the different *iterations* in the active learning process. Note that in Fig. 2, each optimisation step represents the training and update of the deep ensemble using one sequence and its corresponding label. The validation set used in Fig. 2A features sequences which follow the rules of natural protein sequences as encapsulated by ProteinGPT2.[23] Note that while improvement against this set begins to stagnate at around 50 steps, we find a continued improvement with the second validation set of 21 mer sequences designed by An *et al.*[29] (Fig. 2B). This is because, after the first 50 steps when the model is interrogated for suggestion, it already picks up chains with the most negative B_{22} values (Fig. 3A). Continued training on those sequences therefore does not improve the predictability of the average sequence, but improvement is seen against the An *et al.*[29] sequences, which are enriched with more strongly interacting proteins. The MSE decreases from values above 200 to values below 100. After seeing more than 150 sequences, the MSE rises again, which could hint at overfitting. In addition, the decrease in the uncertainty estimate after 150 sequences in Fig. 2A may also be due to overfitting, as this suggests all models within the ensemble are predicting values closer to each other's, while not improving in accuracy. Note that mean-squared error alone is not a complete indicator of progress during training.

When the model is interrogated for suggesting training examples, we also ask it to predict B_{22} values for these sequences. After coarse-grained molecular dynamics simulations of these suggested sequences were performed, we

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

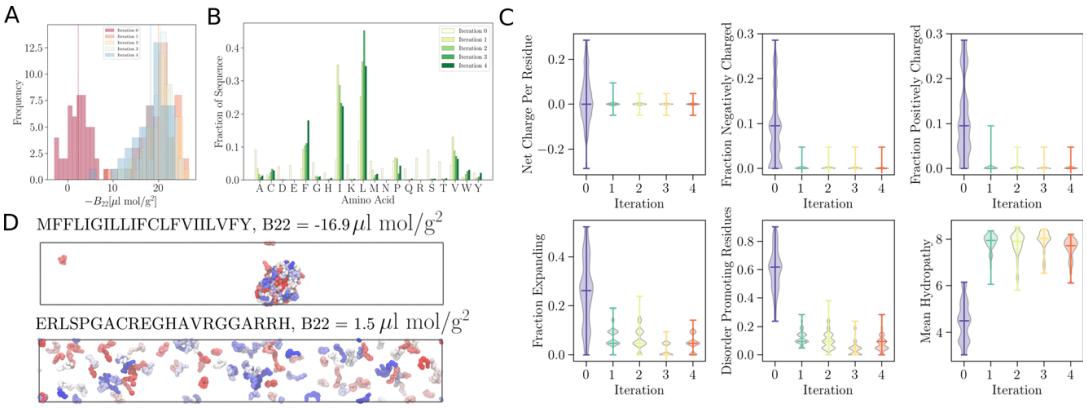

Fig. 3 `wazy` trained to maximise the $-B_{22}$ values, thereby suggesting interacting sequences. (A) Distribution of $-B_{22}$ after each iteration shifts towards more negative values. (B) The plot depicts the amino acid fractions in each iteration, showing a higher fraction of isoleucine (I), leucine (L), valine (V), and phenylalanine (F). (C) Analysis of the amino acid composition of the sequences suggested by the model after each iteration provides insight into the evolving composition of the optimised sequences. (D) Representative simulation snapshot of the protein sequences suggested by `wazy` in the third iteration.

compared how well those forecasts matched with the true values (Fig. 2C). This improvement is quantitatively reflected in the coefficient of determination (R^2), which indicates how well the model predictions match the observed data. Initially, there is poor agreement between predicted and true estimates, indicated by $R^2 = 0.53$. The correlation between predicted and expected values continued to improve in the subsequent iterations as evidenced by the increasing R^2 values. By the fourth iteration, the R^2 value had risen to 0.85.

Note that a significantly longer calibration (*e.g.* 200 sequences instead of 50) in the initial training does not improve the overall outcome (see Fig. S1 in the ESI†). We tentatively conclude that results from the suggested sequences in active learning are more important, or that there is not much new information in these additional initial sequences.

As remarked previously, after just 50 such steps, when the model was interrogated to suggest the next batch of sequences to train with, it already picked up sequences with very negative B_{22} values. When coarse-grained molecular dynamics simulations of these sequences were run, we found that the distribution of B_{22} values had clearly shifted from a mean value of -2.3 μL mol^{-1} g^{-2} to -18.7 μL mol^{-1} g^{-2} (see iteration 1 in Fig. 3A). Subsequent iterations of training with examples requested by the model led to a further shift to even further negative values.

In addition to these quantitative metrics around B_{22} values, in Fig. 3B we show that even the amino acid composition of the suggested sequences changed in biologically meaningful ways. The sequences get enriched with hydrophobic amino acids such as Ile, Leu, and Val. Charged residues are already lost in the first stage of learning. Impressively, all of these biologically and chemically sensible changes emerged without explicit input regarding information about the charge or hydrophobicity of the residues during training. We attribute this to the featurisation with the UniRep LSTM, and it illustrates the remarkable ability of the model to learn and capture relevant physicochemical properties from the sequence data alone. Many of the sequences we provided initially feature charged residues (Fig. 3C). While these sequences are not highly charged, with a net

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

charge per residue close to 0, some sequences are enriched in positively or negatively charged residues. The model seems to learn from the input sequences that a simple way to enhance protein–protein chain interactions is to eliminate charged residues. It is important to note that complementary patches of positively and negatively charged residues can give rise to strong interactions and protein condensation *via* complex coacervation.[7,38] The input sequence set includes sequences which feature a third or half of the residues favouring expanded conformations. As the training progresses, the number of disorder promoting residues decreases strongly with iteration. At the same time the mean hydrophobicity of the suggested sequences increases.

We expect that sequences with large negative values of B_{22} interact strongly and phase separate, which we verified by simulating multiple chains for

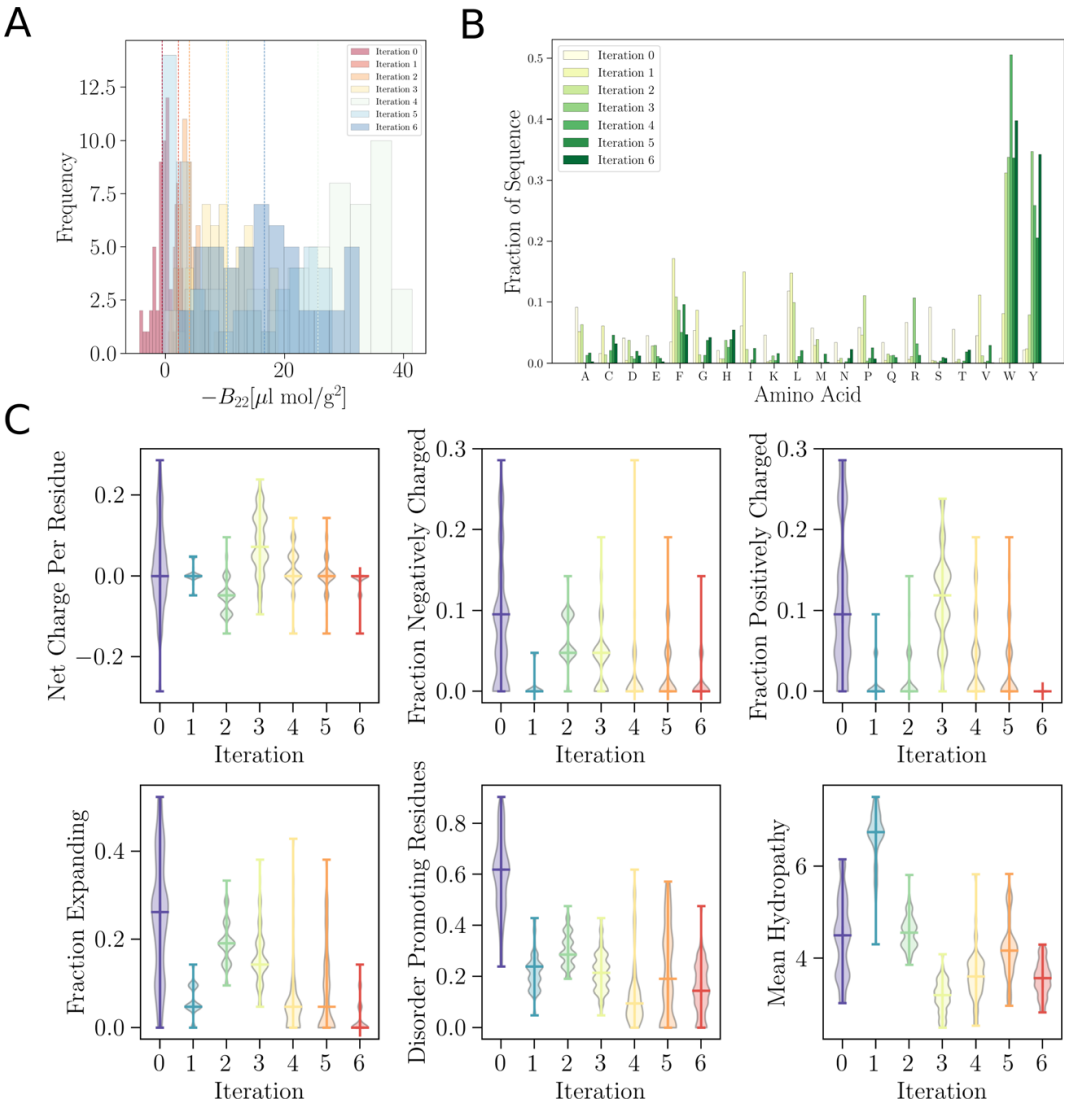

Fig. 4 Optimisation by maximising $-B_{22}$ values, where the simulations are conducted using the CALVADOS force field.[13] (A) The distribution of B_{22} values for 5 iterations of optimisation. The observed distribution shifts towards increasingly negative values compared with the HPS model, indicating a trend towards sequences with stronger interactions. (B) The plot illustrates the fraction of amino acids present in the sequences suggested by `Wazy` during each iteration of optimisation. Notably, there is a noticeable increase in the fraction of isoleucine (I), leucine (L), valine (V), and phenylalanine (F) over successive iterations. (C) Analysis of the amino acid compositions of the sequences suggested by the model after each iteration.

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

a sequence from iteration 3 (Fig. 3D). Multiple chains of this peptide form a condensate as expected given its favourable inter-chain interactions. Conversely, for sequences with a positive B_{22} we expect that the chains are not interacting with each other and that the chains prefer to stay separate from each other and do not phase separate as shown for a sequence from the validation data set (Fig. 3D). We note that the HPS model which underestimates interactions of Arg residues likely underestimates the self interactions of this peptide.

3.1.1 Influence of the force field in the coarse-grained simulations. To examine how the coarse-grained simulation model we use affects the suggested sequences, we conducted coarse-grained simulations using the CALVADOS2 force field,[13,39] which is a reparameterisation of HPS and more accurately describes the interactions of disordered proteins. We used the same initial set of sequences to calibrate. The distribution of $-B_{22}$ from the sequences suggested by training shifted towards larger values (Fig. 4A) of $-B_{22}$ of close to 0 μL mol^{-1} g^{-2} (iteration 0) to approximately 10 μL mol^{-1} g^{-2} (iteration 3). The sequences the model suggested are enriched with Tyr and Trp residues (Fig. 4B). Consistent with BO employing the HPS model, the fraction of disorder-promoting residues decreases across the iterations. However, in contrast to the HPS model, the mean hydropathy value of the suggested sequences decreases when using the CALVADOS force field (Fig. 4C).[13,39] This can be attributed to the increase in the fraction of tyrosine and tryptophan residues, which have a lower hydropathy index compared to isoleucine, leucine, valine, and phenylalanine, as reported in the Kyte–Doolittle hydropathy scale.[40]

3.1.2 Sensitivity to acquisition function: test with expected improvement. From the previous discussion, we showed that active learning enabled us to very efficiently find strongly interacting sequences. However, the peptide sequences we suggested are likely to be hydrophobic for the actual realisation of peptide-based condensate material in a laboratory experiment. For the practical application of designing novel peptide-based condensate materials, one likely needs to perform multi-parameter optimisation to achieve a trade-off between protein interactions and fluidity[29] of the resulting materials, and change the acquisition function accordingly.

Therefore, to potentially identify alternative solutions while exploring the sequence space, we independently trained the model using sequences ranked best by the Expected Improvement (EI) acquisition function. The mean squared error (MSE) between the predicted and expected $-B_{22}$ of the ProteinGPT2

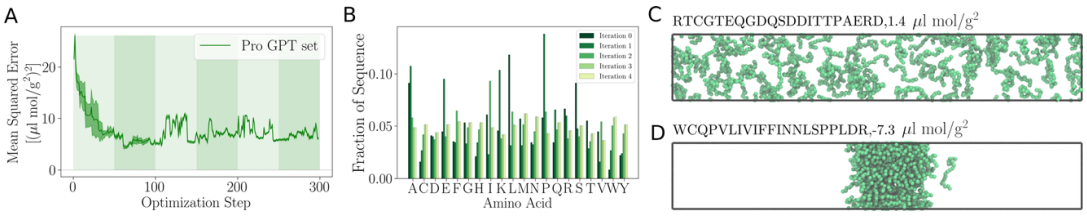

Fig. 5 Optimisation using Expected Improvement (EI) acquisition function to find the optimal value of $-B_{22}$. (A) Evolution of mean squared error (MSE) between the predicted and expected $-B_{22}$ of the ProteinGPT2 validation set. (B) Fraction of amino acids in the sequences suggested by `Wazy` after each iteration. (C) Snapshot of the simulations of sequences suggested by `wazy` at the 4th iteration. (D) The sequences with more negative B_{22} showed clear phase separation (Fig. S2B†).

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

validation set, decreased with each training step (Fig. 5A), similar to what we observed in the UCB-based model. The model suggests sequences with diverse amino acid compositions (Fig. 5B). Examining the suggested sequences, we found that the EI-based optimisation found less strong interactors (smaller $-B_{22}$ values) compared to the model trained using the Upper Confidence Bound (UCB) acquisition function (Fig. S2†). The UCB-based optimisation predominantly predicted sequences with high negative B_{22} values, while the EI-based model predicted a wider range of sequences, mostly with negative B_{22} values and a small fraction with positive B_{22} values.

When we simulated the sequences suggested by the model in the 4th iteration, those with negative B_{22} values in particularly exhibited clear phase separation—forming dense, condensed phases (see Fig. 5C and D). The phase-separating sequence not only features aromatic residues such as Trp and Phe and hydrophobic residues such as Val, but also negatively charged Asp and positively charged Arg at its C-terminal end. Interactions of positively and negatively charged amino acids can also contribute to phase separation,[38,41] provided the net charge of the sequence is low to moderate.[38] Overall the optimisation using UCB led to the identification of sequences with stronger interactions than EI. However, we note that, in the search for an optimal solution, UCB had led to less diverse sequence structures. On the other hand, the sequences ranked by EI were more diverse in amino acid composition but this came at the cost of not touching the sequence space of the most negative B_{22} chains. These differences are evident in the distributions shown in Fig. S2.†

3.2 Designing peptides that do not interact

Employing the same approach, we can design peptides which do not interact with copies of themselves. For many biotechnological applications, one would want to minimise the interactions between proteins, *e.g.*, to reduce the aggregation of therapeutic antibodies.[42] Non-interacting sequences can be useful building blocks in protein design,[9] for example when combining protein domains with different functionalities. To decrease the propensity for protein self-interaction, we maximise B_{22} rather than $-B_{22}$, as previously done for promoting interactions.

As the BO proceeds, suggested sequences tend to have more positive B_{22} values (Fig. 6A) than the initial set of sequences we have used to start the BO (referred to as "Iteration 0"). Interestingly, at the first iteration, the proposed sequences feature a combination of negatively charged amino acids, such as glutamic acid (Glu) and some aspartic acid (Asp) residues, as well as positively charged amino acids, including lysine (Lys) and arginine (Arg) (Fig. 6B and C). In the subsequent second, third and fourth iterations, Arg residues become highly enriched (Fig. 6B). Note that the enrichment in Arg residues may be partially due to the specific choices made in the parameterisation of the coarse-grained simulation model we employ. While Arg is positively charged and increasing the net positive or negative charge of a peptide can prevent phase separation,[38] Arg residues have a very low tendency to interact with other protein residues in the version of residue-level coarse-grained HPS model[34] we employ here. Tesei *et al.*, have shown that a higher interaction strength of Arg for other amino acids leads to a better overall description of disorder proteins, their phase behaviour and properties.[13,39] The net charge per residue increases as the training progresses (Fig. 6C). Residues

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

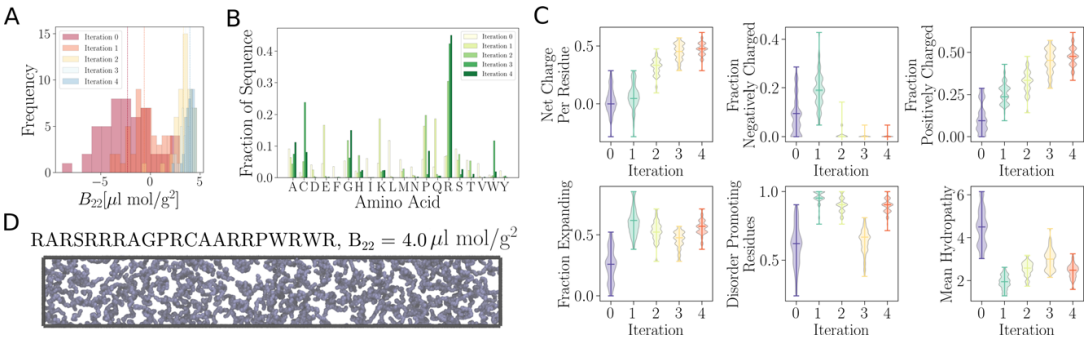

Fig. 6 Design of non-interacting sequences using `wazy`, by training to maximise $+B_{22}$ values, thereby suggesting sequences with minimal intermolecular interactions. (A) Distribution of B_{22} values for the sequences suggested by `wazy` after each iteration, showing a shift towards higher B_{22} values, indicating the evolution of non-interacting sequences. (B) Amino acid composition of the suggested sequences, exhibiting an increasing fraction of the positively charged amino acid arginine (Arg) after each iteration. (C) Analysis of amino acid composition in the suggested sequences. The fraction of positively charged residues increases with each iteration, while the mean hydropathy decreases, reflecting the enrichment of charged residues. (D) Simulation conducted with the sequence suggested by `wazy` after the 4th iteration. The simulation reveals that the suggested sequence does not undergo phase separation, remaining uniformly distributed throughout the simulation box.

tend to favour extended rather than collapsed conformations. The fraction of negatively charged amino acids in the selected sequences becomes negligible after the first iteration (Fig. 6C). Concurrently, the fraction of amino acids associated with favouring extended rather than collapsed conformations increases compared to the initially input sequences. Analogously, the fraction of disorder promoting residues increases with each iteration. However, in the third iteration, there is a temporary drop in the fraction of disorder promoting residues, but this trend is reversed in the fourth iteration, where a high level of disorder-promoting residues is observed once again. Conversely, the mean hydrophobicity decreases as there are almost no aromatic residues (Phe, Tyr, and Trp) present in the sequences suggested by the active learning process.

To test whether these sequences actually interact with each other or not, we observed their behaviour in our CG simulations. In Fig. 6D, we show a snapshot of such a simulation of protein with $B_{22} = 4.0$ μL mol g^{-2}. Clearly, the suggested sequence remains uniformly distributed throughout the simulation box, preferring to not interact with its identical chains and therefore not showing any sign of phase separation.

3.3 Designing peptides binding to phase-separated condensates

To investigate the ability of peptides to interact with pre-formed condensates, we initiated 70 simulations with peptides and a pre-formed condensate of the C-terminal domain of RNA Polymerase II (CTD).[7]

By optimising the peptide sequences to maximise the magnitude of the negative partition free energy ($-\Delta G$), our active learning approach effectively identifies peptides with an enhanced propensity to interact with the condensed phase. Just like for the B_{22} task, the performance of the model was evaluated using the mean squared error (MSE) calculated for a validation set of peptides. Remarkably, the MSE showed a significant decrease after just 50 training steps—

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

similar to other tasks studied here—with MSE dropping from above 200 to less than 50 indicating that the model could effectively capture the underlying patterns and characteristics of the peptide-CTD condensate interactions. In the second iteration hydrophobic residues such as Cys, Phe, Leu, and Trp residues are very prominently enriched (Fig. 7C, S4A and B†). After four iterations of the Bayesian optimisation cycle, peptides are highly enriched in Trp residues. At this stage, peptides are uniformly distributed in the dense condensate phase. At the same time sampling the dilute concentration of the peptides becomes very challenging and long simulations are required for proper equilibration of the peptides between the dilute and condensate phases. Once the peptides are interacting strongly with the condensates and the dilute concentration becomes

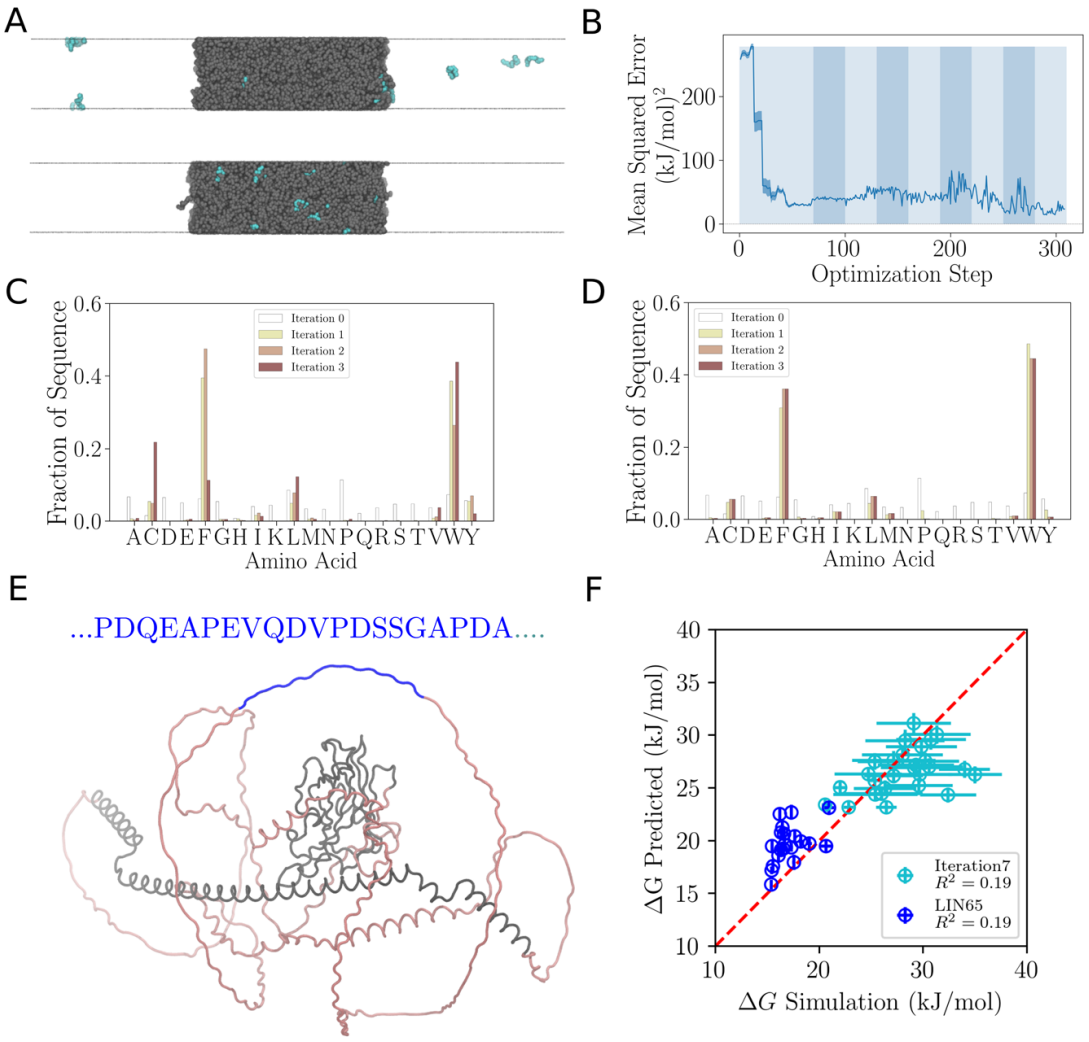

Fig. 7 Designing peptides which bind to a phase-separated condensate of CTD. (A) Grey CTD chains. Light blue peptide chains. (Top) Peptide weakly interacting with condensate. (Bottom) Peptide from BO, which interacts strongly. (B) Progress of the training by comparison to the validation data set. Each training step corresponds to an additional sequence and corresponding binding propensity being processed by `wazy`. (C) Changes in the amino acid composition of sequences binding with ideal CTD condensate after each iteration. (D) For human CTD condensate. (E) AlphaFold (Q95XN0) structure of LIN65. The intrinsically disordered region is shown in pink, and the ordered regions, excluded from the analysis, are shown in grey. A 20-mer sequence from the disordered region is shown in blue. (F) Correlation plot of predicted and computed ΔG values for LIN-65 fragments and peptides suggested by `Wazy` in iteration 7.

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.



difficult to sample further, so enhancing the interactions of the peptides with the condensates becomes difficult. However, at this stage, the design task has in some sense been successfully completed as strongly interacting protein sequences have been found. We note that the lack of sequence diversity means that these peptides are likely too hydrophobic to be useful for biochemical and biotechnological applications, but our investigation shows that it is possible to quickly identify sequences which bind to condensates given a coarse-grained simulation model.

We also performed iterative optimisation to identify sequences that bind effectively to human CTD condensates.[43] The human CTD (hCTD) sequence deviates from the ideal CTD sequence, which consists of a consensus repeat of heptad motifs. During the training process with `wazy`,[25] the algorithm suggested peptide sequences with a similar composition to the sequences found in the previous case, but featuring a higher fraction of the following amino acid residues: Trp, Phe, Leu, and Cys (Fig. 7D, S4C and D).†

Once trained, such a model can quickly evaluate many novel sequences for their potential to interact with specific condensates without requiring time-consuming simulations for each sequence. It can generalize from learned patterns to predict the interaction propensity of sequences different from those in the training set. For instance, we used the trained model to investigate interactions between LIN-65 and RNA Polymerase II (Pol II). In *C. elegans*, LIN-65 (Fig. 7E) is an important regulator of chromatin organisation by supporting the formation of heterochromatin which represses the transcription of genes. Given that the CTD of RNA Pol II is critical for transcription, but LIN-65 is associated with the repression of transcription, we surmised that disordered sequences in the LIN-65 disordered regions may be disfavoured from binding to CTD condensates and this preference encoded in the sequence of LIN-65 could contribute to its specificity in separating transcribed genes from heterochromatin. Since the model is trained on 20-mer sequences, we divided the intrinsically disordered region of LIN-65 into 20-mer fragments (Fig. 7E, pin and blue regions). We then predicted the propensity of LIN-65 fragments to be recruited to CTD condensates using the trained model. The correlation plots shown in Fig. 7F, indicate that the recruitment propensity of LIN-65 fragments is lower than that of sequences suggested by `wazy` after the seventh iteration, as the ΔG is lower for LIN-65 fragments compared to the sequences `wazy` suggested (Fig. S4E and F†). This predictive ability without new simulations showcases the powerful application of machine learning in studying protein–condensate interactions.

3.4 Design of multiphasic condensates

To explore the nature of sequences that can be recruited to CTD condensates while forming multiphasic structures, we employed a strategy based on the second virial coefficient (B_{22}). We calculated B_{22} for a given sequence interacting with itself and with a 21-mer CTD sequence. The residue-level coarse-grained HPS force field we employ for these simulations captures the interactions of 21-mer CTD sequences well, as established through comparisons to all-atom simulations with explicit solvent.[7] The fitness function, $\Delta B_{22} = -B_{22}[\text{seq–seq}] + B_{22}[\text{seq–ctd}]$, was optimised using `wazy`, aiming to identify sequences that interact more strongly with sequences of the same type than with CTD. Through this optimisation process, `wazy` suggested highly hydrophobic sequences that exhibited

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

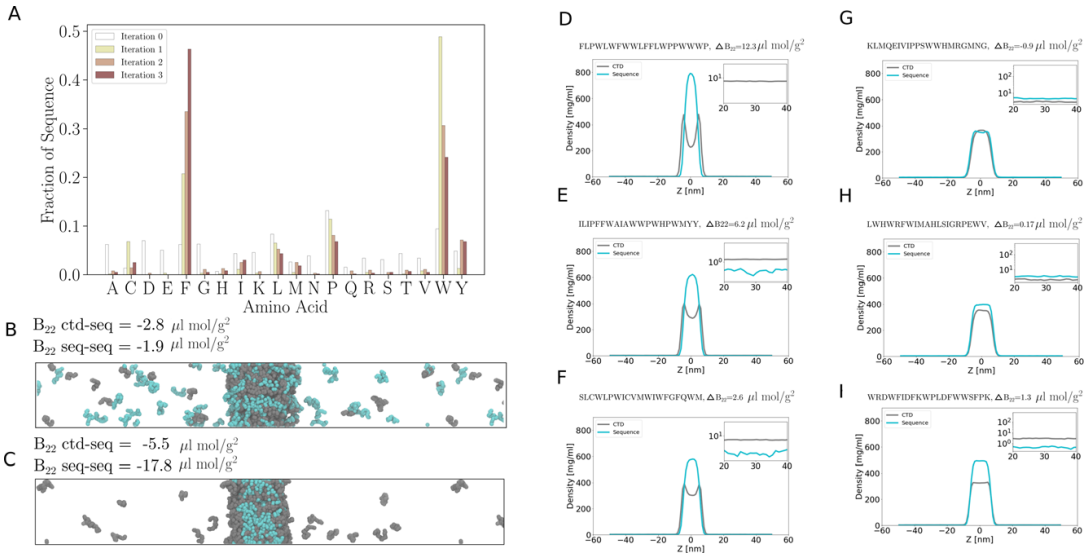

Fig. 8 Designing multiphasic condensates using `wazy`. (A) Amino acid composition of the sequences suggested by `wazy` after each iteration of optimising the fitness function. (B) Simulation with 300 chains each of the optimised sequence and CTD, illustrating the formation of a mixed condensate when the B_{22} values of sequence–sequence and sequence–CTD interactions are comparable. The corresponding density profile is shown in (G). (C) Formation of a remixed phase when B_{22} of the sequence–sequence is notably lower than sequence–CTD. The corresponding density profile is shown in (D). (D–F) Density profile showing the formation of multiphasic condensate when the ΔB_{22} is higher, indicating stronger sequence–sequence interactions compared to sequence–CTD interactions. (G–I) Density profile in the case of mixed phase condensate formed when the ΔB_{22} is lower, suggesting weaker sequence–sequence interactions relative to sequence–CTD interactions.

stronger interactions with themselves compared to their interactions with the CTD sequence (Fig. 8 and S3).† To validate the formation of multiphasic condensates, we conducted simulations with a system comprising 300 chains of the 21-mer CTD sequence and 300 chains of the highly hydrophobic sequences suggested by `wazy` (Fig. 8B and C). Analysis of the density profiles from these simulations revealed a distinct phase separation, where the highly hydrophobic sequences formed the core of the condensate, while the comparatively less hydrophobic CTD sequences formed a surrounding shell (Fig. 8C–F). For lower ΔB_{22} values, the density profile shows a mixed condensate(Fig. 8G–I). This observation suggests that the highly hydrophobic sequences identified through the optimisation process have the propensity to be recruited to CTD condensates while maintaining a distinct phase separation, leading to the formation of a multiphasic condensate structure with a hydrophobic core and a CTD-rich shell.

4 Conclusions

While modelling sequence–property relationships remains a difficult challenge, here we showed it is possible to train neural network models with molecular dynamics simulations to design disordered proteins to: (1) self-interact and phase separate and (2) to bind to phase separated condensates and (3) even shape the morphologies of phase-separated condensates. For this we employed the approach developed by Yang *et al.*[25] that utilises appropriate featurisation from pre-trained sequence models,[24,25,44] and Bayesian Optimisation (BO) for sampling

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

the candidates most useful for training and protein design (Fig. 1). For these design tasks a limited number of sequences, with numeric labels from molecular dynamics, are sufficient, *e.g.*, 50–100 sequences with labels are enough to predict B_{22} with reasonable accuracy (Fig. 2C). Even when no explicit information about the hydrophobicity and charge of individual amino acids is given,[20] approaches based on pre-trained language models are able to learn biologically and chemically relevant information from the sequence data alone.

Active learning provides a significant advantage when individual computations are expensive and an efficient way of traversing a large parameter space is needed. While training and optimising for certain fitness criteria, one must appropriately tune the balance between exploration and exploitation. The improvement using CALVADOS force-field[13,39] highlighted the importance of the availability of state-of-the-art molecular force fields. The sequences identified from the coarse-grained simulations could be further explored with atomistic resolution along with explicit solvent, to include the role of water, ions and secondary structure formation and provide further insight into the molecular driving forces.[30]

A key limitation of the present work is that many of the peptides we suggest here may be too hydrophobic to be realised in a laboratory experiment. Recent work shows that one could explicitly consider the synthesizability of peptides in a neural network.[45] As pointed out by An *et al.*,[29] optimising interaction strength comes with a trade-off in terms of transport properties of the material and considering this in multi-parameter optimisation will yield protein sequences that are more balanced with regards to hydrophobic and non-hydrophobic residues. While we have not attempted to perform multi-parameter optimisation, for future practical applications to inform the design of novel materials underpinned by protein phase separation, multi-parameter optimisation is likely beneficial.

Data availability

All codes and datasets related to this study can be accessed at the following GitHub repository: **https://github.com/cartilage-ftw/active-learning**.

Author contributions

A. C. S. designed and ran most of the HOOMD simulations. A. A. implemented the active learning workflow with the MD simulations. V. X. provided the protein sequences used for training. A. C. S., A. A., J. P. and L. S. conceived the study and interpreted the data. L. S. oversaw the study. A. C. S., A. A. and L. S. wrote the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

A. C. S. and L. S. S. thank M^{3}ODEL for support. L. S. S. acknowledges support by ReALity (Resilience, Adaptation and Longevity) and Forschungsinitiative des Landes Rheinland-Pfalz. This project was funded by SFB 1551 Project No.

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

464588647 of the DFG (Deutsche Forschungsgemeinschaft). We acknowledge a mini grant from the Emergent AI (EAI) Center of Carl Zeiss Foundation. L. S. S. and A. A. thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 233630050 – TRR 146 for support. We gratefully acknowledge the advisory services offered and the computing time granted on the supercomputers Mogon II at Johannes Gutenberg University Mainz, which is a member of the AHRP (Alliance for High Performance Computing in Rhineland Palatinate) and the Gauss Alliance e.V. We thank Dr Noelia Ferruz, Prof. Dr Susanne Gerber, Yehor Tuchkov, Prof. Dr Heinz Koeppl, Dr Sören von Bülow and Prof. Dr Shikha Dhiman for inspiring discussions.

Notes and references

1 E. W. Martin, A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu and T. Mittag, *Science*, 2020, **367**, 694–699.
2 B. Tsang, I. Pritišanac, S. W. Scherer, A. M. Moses and J. D. Forman-Kay, *Cell*, 2020, **183**, 1742–1756.
3 A. Zbinden, M. Pérez-Berlanga, P. De Rossi and M. Polymenidou, *Dev. Cell*, 2020, **55**, 45–68.
4 L. A. Gruijs da Silva, F. Simonetti, S. Hutten, H. Riemenschneider, E. L. Sternburg, L. M. Pietrek, J. Gebel, V. Dötsch, D. Edbauer, G. Hummer, L. S. Stelzl and D. Dormann, *EMBO J.*, 2022, **41**, e108443.
5 R. M. Welles, K. A. Sojitra, M. V. Garabedian, B. Xia, W. Wang, M. Guan, R. M. Regy, E. R. Gallagher, D. A. Hammer, J. Mittal and M. C. Good, *Nat. Chem.*, 2024, **16**, 1062–1072.
6 M. Farag, W. M. Borcherds, A. Bremer, T. Mittag and R. V. Pappu, *Nat. Commun.*, 2023, **14**, 5527.
7 A. Changiarath, J. J. Michels, R. H. Rodriguez, S. M. Hanson, F. Schmid, J. Padeken and L. S. Stelzl, *bioRxiv*, 2024, preprint, DOI: **10.1101/2024.03.16.585180**.
8 M. Dzuricky, B. A. Rogers, A. Shahid, P. S. Cremer and A. Chilkoti, *Nat. Chem.*, 2020, **12**, 814–825.
9 A. Gräwe, M. Merkx and V. Stein, *Bioconjugate Chem.*, 2022, **33**, 1415–1421.
10 S. Rekhi, C. G. Garcia, M. Barai, A. Rizuan, B. S. Schuster, K. L. Kiick and J. Mittal, *Nat. Chem.*, 2024, **16**, 1113–1124.
11 J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar and R. Collepardo-Guevara, *Nat. Comput. Sci.*, 2021, **1**, 732–743.
12 W. M. Jacobs, *J. Chem. Theory Comput.*, 2023, **19**, 3429–3445.
13 G. Tesei and K. Lindorff-Larsen, *Open Research Europe*, 2023, **2**, 94.
14 G. M. Ginell, R. J. Emenecker, J. M. Lotthammer, E. T. Usher and A. S. Holehouse, *bioRxiv*, 2024, preprint, DOI: **10.1101/2024.06.03.597104**.
15 S. von Bülow, G. Tesei and K. Lindorff-Larsen, *bioRxiv*, 2024, preprint, DOI: **10.1101/2024.06.03.597109**.
16 F. Pesce, A. Bremer, G. Tesei, J. B. Hopkins, C. R. Grace, T. Mittag and K. Lindorff-Larsen, *Sci. Adv.*, 2024, **10**(35), eadm9926, DOI: **10.1126/sciadv.adm9926**.
17 T. Hoffbauer and B. Strodel, *bioRxiv*, 2024, preprint, DOI: **10.1101/2024.01.12.575432**.

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

18 X. Zeng, C. Liu, M. J. Fossat, P. Ren, A. Chilkoti and R. V. Pappu, *APL Mater.*, 2021, **9**, 021119.
19 P. Y. Chew, J. A. Joseph, R. Collepardo-Guevara and A. Reinhardt, *Chem. Sci.*, 2023, **14**, 1820–1836.
20 A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik and B. Rost, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022, **44**, 7112–7127.
21 N. Brandes, D. Ofer, Y. Peleg, N. Rappoport and M. Linial, *Bioinformatics*, 2022, **38**, 2102–2110.
22 R. Verkuil, O. Kabeli, Y. Du, B. I. M. Wicky, L. F. Milles, J. Dauparas, D. Baker, S. Ovchinnikov, T. Sercu and A. Rives, *bioRxiv*, 2022, preprint, DOI: **10.1101/2022.12.21.521521**.
23 N. Ferruz, S. Schmidt and B. Höcker, *Nat. Commun.*, 2022, **13**, 4348.
24 E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi and G. M. Church, *Nat. Methods*, 2019, **16**, 1315–1322.
25 Z. Yang, K. A. Milas and A. D. White, *bioRxiv*, 2022, preprint, **https://doi.org/10.1101/2022.08.05.502972**.
26 E. J. Ma and A. Kummer, *bioRxiv*, 2020, preprint, DOI: **10.1101/2020.05.11.088344**.
27 M. Aldeghi and C. W. Coley, *Chem. Sci.*, 2022, **13**, 8221–8223.
28 K. Shmilovich, S. S. Panda, A. Stouffer, J. D. Tovar and A. L. Ferguson, *Digital Discovery*, 2022, **1**, 448–462.
29 Y. An, M. A. Webb and W. M. Jacobs, *Sci. Adv.*, 2024, **10**, eadj2448.
30 B. Mohr, K. Shmilovich, I. S. Kleinwächter, D. Schneider, A. L. Ferguson and T. Bereau, *Chem. Sci.*, 2022, **13**, 4498–4511.
31 Y. Khalak, G. Tresadern, D. F. Hahn, B. L. de Groot and V. Gapsys, *J. Chem. Theory Comput.*, 2022, **18**, 6259–6270.
32 B. Lakshminarayanan, A. Pritzel and C. Blundell, in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6405–6416.
33 P. I. Frazier, *arXiv*, preprint, arXiv:1807.02811, 2018, DOI: **10.48550/arXiv.1807.02811**.
34 G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best and J. Mittal, *PLoS Comput. Biol.*, 2018, **14**, e1005941–23.
35 J. A. Anderson, J. Glaser and S. C. Glotzer, *Comput. Mater. Sci.*, 2020, **173**, 109363.
36 J. A. Riback, L. Zhu, M. C. Ferrolino, M. Tolbert, D. M. Mitrea, D. W. Sanders, M.-T. Wei, R. W. Kriwacki and C. P. Brangwynne, *Nature*, 2020, **581**, 209–214.
37 A. S. Holehouse, R. K. Das, J. N. Ahad, M. O. G. Richardson and R. V. Pappu, *Biophys. J.*, 2017, **112**, 16–21.
38 A. Bremer, M. Farag, W. M. Borcherds, I. Peran, E. W. Martin, R. V. Pappu and T. Mittag, *Nat. Chem.*, 2022, **14**, 196–207.
39 G. Tesei, T. K. Schulze, R. Crehuet and K. Lindorff-Larsen, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**(44), e2111696118.
40 J. Kyte and R. F. Doolittle, *J. Mol. Biol.*, 1982, **157**, 105–132.
41 D. J. Bauer, L. S. Stelzl and A. Nikoubashman, *J. Chem. Phys.*, 2022, **157**, 154903.
42 T. M. Prass, P. Garidel, M. Blech and L. V. Schäfer, *J. Chem. Inf. Model.*, 2023, **63**, 6129–6140.

43 D. Flores-Solis, I. P. Lushpinskaia, A. A. Polyansky, A. Changiarath, M. Boehning, M. Mirkovic, J. Walshe, L. M. Pietrek, P. Cramer, L. S. Stelzl, B. Zagrovic and M. Zweckstetter, *Nat. Commun.*, 2023, **14**, 5979.
44 Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, *Science*, 2023, **379**, 1123–1130.
45 R. Zhang, H. Wu, Y. Xiu, K. Li, N. Chen, Y. Wang, Y. Wang, X. Gao and F. Zhou, *arXiv*, preprint, arXiv:1807.02811, 2023, DOI: **10.48550/arXiv.1807.02811**.

Open Access Article. Published on 03 August 2024. Downloaded on 1/30/2025 7:52:31 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

6 Summary and outlook

Biomolecular condensates, formed via liquid-liquid phase separation (LLPS), are critical for organizing the intracellular environment by concentrating macromolecules both spatially and temporally. These dynamic assemblies facilitate biochemical reactions and provide a flexible organizational mechanism. Condensates enable cells to modulate molecular processes in response to dynamic cues. The C-terminal domain (CTD) of RNA Polymerase II, an intrinsically disordered region (IDR) that regulates transcription, exemplifies such a system. The CTD consists of tandem heptapeptide repeats that undergo extensive post-translational modifications, primarily phosphorylation. These modifications dictate the recruitment of transcription factors and RNA-processing machinery, thereby orchestrating the transitions between different stages of transcription, such as initiation, elongation, and termination. Despite its functional importance, the precise structural and dynamic mechanisms underlying CTD phase separation and transcriptional regulation remain incompletely understood due to the highly flexible and disordered nature of the domain.

This thesis addresses these challenges by integrating advanced simulation techniques, machine learning, and experimental collaborations to systematically investigate the conformational landscape, phase behaviour, and functional consequences of CTD phase separation. The research is structured into four key chapters: Chapter 1 examines the role of phosphorylation in phase separation and transcription regulation, Chapter 2 explores the conformational equilibrium of the RNA Polymerase II CTD by integrating atomically-detailled models of proteins with experimental data, Chapter 3 introduces a novel approach to simulate phosphorylation-driven condensate dynamics, and Chapter 4 leverages machine learning to design phase-separating proteins and shows how to binding of peptides to CTD condensates and the morphology of CTD condensates can be designed and shaped in silico. Collectively, these studies provide critical insights into the molecular mechanisms governing transcriptional condensates. The research not only advances the fundamental understanding of IDPs in the transcription process but also offers computational tools and predictive frameworks that can be applied to other biomolecular condensates. These findings have broad implications for synthetic biology, drug discovery, and molecular medicine, enabling the design of novel biomaterials and therapeutic interventions targeting condensate-mediated processes.

Key Findings

1. **Phase Separation and Transcription Regulation:** Phosphorylation of CTD plays a crucial role in the formation of distinct phase-separated condensates that regulate transcription initiation and elongation. These condensates facilitate interactions with IDRs and other transcription-related proteins. Molecular dynamics simulations revealed the effects of phosphorylation on CTD interactions, highlighting the role of hydrophobic and electrostatic forces in phase separation. The research shows phosphorylation as a molecular switch that alters protein behaviour and condensate morphology. Experimental validation through super-resolution microscopy confirmed these findings, linking condensate morphology with transcriptional activity.

2. **Structural Dynamics of Intrinsically Disordered CTD:** Through RHCG and REMD simulations, all-atom structural ensembles of yCTD were generated that align well with experimental NMR data. These simulations successfully captured rare but biologically significant events, such as proline isomerization, demonstrating the power of computational approaches in modelling intrinsically disordered proteins (IDPs). The agreement between simulated and experimental structures validates the robustness of this methodology and paves the way for further computational studies of IDPs.

3. **Phosphorylation-Driven Dynamics of Protein Condensates:** An efficient computational approach was introduced to model phosphorylation dynamics in protein condensates, capturing how phosphorylation and dephosphorylation modulate condensate morphology and stability. By simulating phosphorylation as a kinetic rate process rather than explicitly modelling enzymes, condensates that influence transcriptional activity were observed. The findings suggest that phosphorylation-dependent condensate dissolution could be a mechanism for regulating transcriptional bursts.

4. **Machine Learning-Guided Design of Disordered Proteins:** One important contribution of this work is the application of neural network models trained on molecular dynamics simulations to design novel disordered protein sequences. Using active learning and Bayesian Optimization, sequences were successfully identified that phase separate, bind to condensates, and even shape condensate morphology. These results highlight the potential for computational approaches to inform the rational design of IDPs for synthetic biology and biomaterials applications.

Implications and Future Directions

This research addresses the behavior of IDPs, focusing on the CTD of RNA Polymerase II in regulating gene transcription. It combines advanced molecular dynamics simulations, machine learning, and experimental validation to reveal how protein phase separation modulates transcription.

A key contribution is the mechanistic insight into transcriptional regulation, demonstrating how CTD phosphorylation leads to the formation of phase-separated condensates, which recruit and organize the transcription machinery. The ability to predict and design protein interactions using machine learning and simulations enables the understanding of cellular processes and development of targeted therapies.

This work establishes a computational framework applicable beyond RNA Polymerase II. Integrating computational modeling, experimental validation, and machine learning offers a method for studying molecular mechanisms of protein phase separation. The framework allows for modeling disordered proteins, simulating protein interactions, and predicting molecular behaviors in biological systems.

The research presented in this thesis opens up numerous directions for future investigation into intrinsically disordered proteins, transcriptional condensates, and computational protein design. The computational methods developed in this thesis provide a strong foundation for investigating complex biological systems using advanced modelling techniques. Future studies could focus on developing more sophisticated computer simulations that combine different approaches like machine learning, atomic-level modelling, and experimentation. By expanding the current methodological frameworks, it becomes possible to design new protein sequences with specific functions, understand how proteins form condensates, and explore their interactions with other cellular components. The computational strategies can help unravel how different cellular conditions affect phase separation behaviour and molecular interactions. Potential directions include creating more nuanced models of phosphorylation dynamics, investigating broader ranges of protein variants, and developing multi-parameter optimization techniques that capture the intricate mechanisms of cellular processes.

Bibliography

[1]Salman F Banani, Hyun O Lee, Anthony A Hyman, and Michael K Rosen. “Biomolecular condensates: organizers of cellular biochemistry”. In: *Nature reviews Molecular cell biology* 18.5 (2017), pp. 285–298 (cit. on pp. 1, 6, 7).

[2]Diana M Mitrea and Richard W Kriwacki. “Phase separation in biology; functional organization of a higher order”. In: *Cell Communication and Signaling* 14 (2016), pp. 1–20 (cit. on p. 1).

[3]Clifford P Brangwynne, Christian R Eckmann, David S Courson, et al. “Germline P granules are liquid droplets that localize by controlled dissolution/condensation”. In: *Science* 324.5935 (2009), pp. 1729–1732 (cit. on p. 1).

[4]Dan Bracha, Mackenzie T Walls, and Clifford P Brangwynne. “Probing and engineering liquid-phase organelles”. In: *Nature biotechnology* 37.12 (2019), pp. 1435–1445 (cit. on p. 2).

[5]Bradley W Davis, William M Aumiller, Negar Hashemian, et al. “Colocalization and sequential enzyme activity in aqueous biphasic systems: experiments and modeling”. In: *Biophysical journal* 109.10 (2015), pp. 2182–2194 (cit. on p. 2).

[6]Brian G O’Flynn and Tanja Mittag. “The role of liquid–liquid phase separation in regulating enzyme activity”. In: *Current opinion in cell biology* 69 (2021), pp. 70–79 (cit. on p. 2).

[7]Denes Hnisz, Krishna Shrinivas, Richard A Young, Arup K Chakraborty, and Phillip A Sharp. “A phase separation model for transcriptional control”. In: *Cell* 169.1 (2017), pp. 13–23 (cit. on p. 2).

[8]Ibraheem Alshareedah, Mahdi Muhammad Moosa, Muralikrishna Raju, Davit A Potoyan, and Priya R Banerjee. “Phase transition of RNA- protein complexes into ordered hollow condensates”. In: *Proceedings of the National Academy of Sciences* 117.27 (2020), pp. 15650–15658 (cit. on p. 2).

[9]Jingyi Fei, Mahdieh Jadaliha, Tyler S Harmon, et al. “Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution”. In: *Journal of cell science* 130.24 (2017), pp. 4180–4192 (cit. on pp. 2, 41).

[10]Marina Feric, Nilesh Vaidya, Tyler S Harmon, et al. “Coexisting liquid phases underlie nucleolar subcompartments”. In: *Cell* 165.7 (2016), pp. 1686–1697 (cit. on pp. 2, 41).

[11]Simon Alberti and Anthony A Hyman. “Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing”. In: *Nature reviews Molecular cell biology* 22.3 (2021), pp. 196–213 (cit. on pp. 2, 8).

[12]Tanja Mittag and Rohit V Pappu. “A conceptual framework for understanding phase separation and addressing open questions and challenges”. In: *Molecular cell* 82.12 (2022), pp. 2201–2214 (cit. on p. 2).

[13]Simon Alberti, Amy Gladfelter, and Tanja Mittag. “Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates”. In: *Cell* 176.3 (2019), pp. 419–434 (cit. on pp. 2, 35).

[14]David Whitford. *Proteins: structure and function*. John Wiley & Sons, 2013 (cit. on p. 3).

[15]M Madan Babu. “The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease”. In: *Biochemical Society Transactions* 44.5 (2016), pp. 1185–1200 (cit. on p. 3).

[16]Natasha S Latysheva, Tilman Flock, Robert J Weatheritt, Sreenivas Chavali, and M Madan Babu. “How do disordered regions achieve comparable functions to structured domains?” In: *Protein Science* 24.6 (2015), pp. 909–922 (cit. on p. 3).

[17]Christine A Orengo, Annabel E Todd, and Janet M Thornton. “From protein structure to function”. In: *Current opinion in structural biology* 9.3 (1999), pp. 374–382 (cit. on p. 3).

[18]Christine Vogel, Matthew Bashton, Nicola D Kerrison, Cyrus Chothia, and Sarah A Teichmann. “Structure, function and evolution of multidomain proteins”. In: *Current opinion in structural biology* 14.2 (2004), pp. 208–216 (cit. on p. 3).

[19]Robin Van Der Lee, Marija Buljan, Benjamin Lang, et al. “Classification of intrinsically disordered regions and proteins”. In: *Chemical reviews* 114.13 (2014), pp. 6589–6631 (cit. on p. 3).

[20]Tobias Sikosek and Hue Sun Chan. “Biophysics of protein evolution and evolutionary protein biophysics”. In: *Journal of The Royal Society Interface* 11.100 (2014), p. 20140419 (cit. on p. 4).

[21]Celeste J Brown, Audra K Johnson, A Keith Dunker, and Gary W Daughdrill. “Evolution and disorder”. In: *Current opinion in structural biology* 21.3 (2011), pp. 441–446 (cit. on p. 4).

[22]Shahin Ramazi and Javad Zahiri. “Post-translational modifications in proteins: resources, tools and prediction methods”. In: *Database* 2021 (2021), baab012 (cit. on p. 4).

[23]Alaji Bah and Julie D Forman-Kay. “Modulation of intrinsically disordered protein function by post-translational modifications”. In: *Journal of Biological Chemistry* 291.13 (2016), pp. 6696–6705 (cit. on pp. 4, 5, 89).

[24]Anton Abyzov, Martin Blackledge, and Markus Zweckstetter. “Conformational dynamics of intrinsically disordered proteins regulate biomolecular condensate chemistry”. In: *Chemical Reviews* 122.6 (2022), pp. 6719–6748 (cit. on p. 4).

[25]Alex S Holehouse and Birthe B Kragelund. “The molecular basis for cellular function of intrinsically disordered protein regions”. In: *Nature Reviews Molecular Cell Biology* 25.3 (2024), pp. 187–211 (cit. on pp. 4, 5).

[26]Jeong-Mo Choi, Alex S Holehouse, and Rohit V Pappu. “Physical principles underlying the complex biology of intracellular phase transitions”. In: *Annual review of biophysics* 49.1 (2020), pp. 107–133 (cit. on p. 5).

[27]Rahul K Das, Kiersten M Ruff, and Rohit V Pappu. “Relating sequence encoded information to form and function of intrinsically disordered proteins”. In: *Current opinion in structural biology* 32 (2015), pp. 102–112 (cit. on pp. 5, 35).

[28]Sara Calabretta and Stéphane Richard. “Emerging roles of disordered sequences in RNA-binding proteins”. In: *Trends in biochemical sciences* 40.11 (2015), pp. 662–672 (cit. on p. 5).

[29]Kim Van Roey, Bora Uyar, Robert J Weatheritt, et al. “Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation”. In: *Chemical reviews* 114.13 (2014), pp. 6733–6778 (cit. on p. 5).

[30]Liliana Malinovska, Sonja Kroschwald, and Simon Alberti. “Protein disorder, prion propensities, and self-organizing macromolecular collectives”. In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1834.5 (2013), pp. 918–931 (cit. on p. 6).

[31]Andrew S Lyon, William B Peeples, and Michael K Rosen. “A framework for understanding the functions of biomolecular condensates across scales”. In: *Nature reviews Molecular cell biology* 22.3 (2021), pp. 215–235 (cit. on p. 7).

[32]Benjamin R Sabari, Alessandra Dall’Agnese, and Richard A Young. “Biomolecular condensates in the nucleus”. In: *Trends in biochemical sciences* 45.11 (2020), pp. 961–977 (cit. on pp. 7, 10).

[33]Chih-Yung S Lee, Andrea Putnam, Tu Lu, et al. “Recruitment of mRNAs to P granules by condensation with intrinsically-disordered proteins”. In: *Elife* 9 (2020), e52896 (cit. on p. 7).

[34]Benjamin S Schuster, Roshan Mammen Regy, Elliott M Dolan, et al. “Biomolecular condensates: Sequence determinants of phase separation, microstructural organization, enzymatic activity, and material properties”. In: *The journal of physical chemistry B* 125.14 (2021), pp. 3441–3451 (cit. on p. 7).

[35]Ivan Peran and Tanja Mittag. “Molecular structure in biomolecular condensates”. In: *Current opinion in structural biology* 60 (2020), pp. 17–26 (cit. on p. 7).

[36]Clifford P Brangwynne, Timothy J Mitchison, and Anthony A Hyman. “Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes”. In: *Proceedings of the National Academy of Sciences* 108.11 (2011), pp. 4334–4339 (cit. on p. 7).

[37]Yifan Dai, Lingchong You, and Ashutosh Chilkoti. “Engineering synthetic biomolecular condensates”. In: *Nature Reviews Bioengineering* 1.7 (2023), pp. 466–480 (cit. on p. 7).

[38]Ibraheem Alshareedah, Taranpreet Kaur, and Priya R Banerjee. “Methods for characterizing the material properties of biomolecular condensates”. In: *Methods in enzymology*. Vol. 646. Elsevier, 2021, pp. 143–183 (cit. on p. 7).

[39]Anthony A Hyman, Christoph A Weber, and Frank Jülicher. “Liquid-liquid phase separation in biology”. In: *Annual review of cell and developmental biology* 30.1 (2014), pp. 39–58 (cit. on p. 8).

[40]Ibraheem Alshareedah, Wade M Borcherds, Samuel R Cohen, et al. “Sequence-specific interactions determine viscoelasticity and ageing dynamics of protein condensates”. In: *Nature Physics* 20.9 (2024), pp. 1482–1491 (cit. on p. 8).

[41]Diana M Mitrea, Matthäus Mittasch, Beatriz Ferreira Gomes, Isaac A Klein, and Mark A Murcko. “Modulating biomolecular condensates: a novel approach to drug discovery”. In: *Nature Reviews Drug Discovery* 21.11 (2022), pp. 841–862 (cit. on p. 8).

[42]Michele Vendruscolo and Monika Fuxreiter. “Protein condensation diseases: therapeutic opportunities”. In: *Nature communications* 13.1 (2022), p. 5550 (cit. on p. 8).

[43]Brent S Visser, Wojciech P Lipiński, and Evan Spruijt. “The role of biomolecular condensates in protein aggregation”. In: *Nature Reviews Chemistry* (2024), pp. 1–15 (cit. on p. 8).

[44]Simon Alberti and Anthony A Hyman. “Are aberrant phase transitions a driver of cellular aging?” In: *BioEssays* 38.10 (2016), pp. 959–968 (cit. on p. 8).

[45]Lara A Gruijs da Silva, Francesca Simonetti, Saskia Hutten, et al. “Disease-linked TDP-43 hyperphosphorylation suppresses TDP-43 condensation and aggregation”. In: *The EMBO journal* 41.8 (2022), e108443 (cit. on p. 8).

[46]Simon Alberti and Dorothee Dormann. “Liquid–liquid phase separation in disease”. In: *Annual review of genetics* 53.1 (2019), pp. 171–194 (cit. on p. 8).

[47]Mario Hofweber, Saskia Hutten, Benjamin Bourgeois, et al. “Phase separation of FUS is suppressed by its nuclear import receptor and arginine methylation”. In: *Cell* 173.3 (2018), pp. 706–719 (cit. on p. 8).

[48]Charles H Li, Eliot L Coffey, Alessandra Dall’Agnese, et al. “MeCP2 links heterochromatin condensates and neurodevelopmental disease”. In: *Nature* 586.7829 (2020), pp. 440–444 (cit. on p. 8).

[49]Patrick Cramer. “Organization and regulation of gene transcription”. In: *Nature* 573.7772 (2019), pp. 45–54 (cit. on p. 9).

[50]Kevin M Harlen and L Stirling Churchman. “The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain”. In: *Nature reviews Molecular cell biology* 18.4 (2017), pp. 263–273 (cit. on p. 9).

[51]Ibrahim I Cisse, Ignacio Izeddin, Sebastien Z Causse, et al. “Real-time dynamics of RNA polymerase II clustering in live human cells”. In: *Science* 341.6146 (2013), pp. 664–667 (cit. on pp. 9, 69).

[52]Huasong Lu, Dan Yu, Anders S Hansen, et al. “Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II”. In: *Nature* 558.7709 (2018), pp. 318–323 (cit. on p. 9).

[53]Gaofeng Pei, Heankel Lyons, Pilong Li, and Benjamin R Sabari. “Transcription regulation by biomolecular condensates”. In: *Nature Reviews Molecular Cell Biology* (2024), pp. 1–24 (cit. on p. 10).

[54]Johannes Söding, David Zwicker, Salma Sohrabi-Jahromi, Marc Boehning, and Jan Kirschbaum. “Mechanisms for active regulation of biomolecular condensates”. In: *Trends in cell biology* 30.1 (2020), pp. 4–14 (cit. on p. 10).

[55]Martin Stortz, Diego M Presman, and Valeria Levi. “Transcriptional condensates: a blessing or a curse for gene regulation?” In: *Communications Biology* 7.1 (2024), p. 187 (cit. on p. 10).

[56]Justin Demmerle, Siyuan Hao, and Danfeng Cai. “Transcriptional condensates and phase separation: condensing information across scales and mechanisms”. In: *Nucleus* 14.1 (2023), p. 2213551 (cit. on p. 10).

[57]Boyuan Deng and Gang Wan. “Technologies for studying phase-separated biomolecular condensates”. In: *Advanced Biotechnology* 2.1 (2024), p. 10 (cit. on p. 11).

[58]Christian F. Pantoja, Alain Ibáñez de Opakua, Maria-Sol Cima-Omori, and Markus Zweckstetter. “Determining the Physico-Chemical Composition of Biomolecular Condensates from Spatially-Resolved NMR”. In: *Angewandte Chemie International Edition* 62.17 (2023), e202218078 (cit. on p. 11).

[59]Simone Kosol, Sara Contreras-Martos, Cesyen Cedeño, and Peter Tompa. “Structural characterization of intrinsically disordered proteins by NMR spectroscopy”. In: *Molecules* 18.9 (2013), pp. 10802–10828 (cit. on p. 12).

[60]Eric B Gibbs, Erik C Cook, and Scott A Showalter. “Application of NMR to studies of intrinsically disordered proteins”. In: *Archives of biochemistry and biophysics* 628 (2017), pp. 57–70 (cit. on p. 12).

[61]H Jane Dyson and Peter E Wright. “NMR illuminates intrinsic disorder”. In: *Current opinion in structural biology* 70 (2021), pp. 44–52 (cit. on p. 12).

[62]Maxwell CK Leung, Phillip L Williams, Alexandre Benedetto, et al. “Caenorhabditis elegans: an emerging model in biomedical and environmental toxicology”. In: *Toxicological sciences* 106.1 (2008), pp. 5–28 (cit. on p. 13).

[63]Philip M Meneely, Caroline L Dahlberg, and Jacqueline K Rose. “Working with worms: Caenorhabditis elegans as a model organism”. In: *Current Protocols Essential Laboratory Techniques* 19.1 (2019), e35 (cit. on p. 13).

[64]Andrea Putnam, Madeline Cassani, Jarrett Smith, and Geraldine Seydoux. “A gel phase promotes condensation of liquid P granules in Caenorhabditis elegans embryos”. In: *Nature structural & molecular biology* 26.3 (2019), pp. 220–226 (cit. on p. 13).

[65]Titus Kaletta and Michael O Hengartner. “Finding function in novel targets: C. elegans as a model organism”. In: *Nature reviews Drug discovery* 5.5 (2006), pp. 387–399 (cit. on p. 13).

[66]Pantira Singkum, San Suwanmanee, Potjaman Pumeesat, and Natthanej Luplertlop. “A powerful in vivo alternative model in scientific research: Galleria mellonella”. In: *Acta microbiologica et immunologica Hungarica* 66.1 (2019), pp. 31–55 (cit. on p. 13).

[67]Friederike Schmid. “Theory and simulation of multiphase polymer systems”. In: *arXiv preprint arXiv:1001.1265* (2010) (cit. on p. 14).

[68]Rohit V Pappu, Xiaoling Wang, Andreas Vitalis, and Scott L Crick. “A polymer physics perspective on driving forces and mechanisms for protein aggregation”. In: *Archives of biochemistry and biophysics* 469.1 (2008), pp. 132–141 (cit. on p. 14).

[69]Clifford P Brangwynne, Peter Tompa, and Rohit V Pappu. “Polymer physics of intracellular phase transitions”. In: *Nature Physics* 11.11 (2015), pp. 899–904 (cit. on p. 14).

[70]Ian W Hamley. *Introduction to soft matter: synthetic and biological self-assembling materials*. John Wiley & Sons, 2007 (cit. on p. 15).

[71]Tomohiro Yamazaki and Tetsuya Yamamoto. “Statistical Thermodynamics Approach for Intracellular Phase Separation”. In: *piRNA: Methods and Protocols*. Springer, 2022, pp. 361–393 (cit. on p. 15).

[72]Gregor Munro Anderson. *Thermodynamics of natural systems*. Cambridge University Press, 2005 (cit. on p. 15).

[73]Maurice L Huggins. “Solutions of long chain compounds”. In: *The Journal of chemical physics* 9.5 (1941), pp. 440–440 (cit. on p. 15).

[74]Paul J Flory. “Thermodynamics of high polymer solutions”. In: *The Journal of chemical physics* 9.8 (1941), pp. 660–660 (cit. on p. 15).

[75]Robert J Young and Peter A Lovell. *Introduction to polymers*. CRC press, 2011 (cit. on p. 16).

[76]Ian W Hamley. “Introduction to soft matter: polymers, colloids, amphiphiles and liquid crystals”. In: *(No Title)* (2000) (cit. on p. 18).

[77]Nafisa Begam, Olga Matsarskaia, Michael Sztucki, Fajun Zhang, and Frank Schreiber. “Unification of lower and upper critical solution temperature phase behavior of globular protein solutions in the presence of multivalent cations”. In: *Soft Matter* 16.8 (2020), pp. 2128–2134 (cit. on p. 19).

[78]Joel Berry, Clifford P Brangwynne, and Mikko Haataja. “Physical principles of intracellular organization via active and passive phase transitions”. In: *Reports on Progress in Physics* 81.4 (2018), p. 046601 (cit. on p. 19).

[79]Masao Doi. *Introduction to polymer physics*. Oxford university press, 1996 (cit. on p. 20).

[80]Somendra M Bhattacharjee, Achille Giacometti, and Amos Maritan. “Flory theory for polymers”. In: *Journal of Physics: Condensed Matter* 25.50 (2013), p. 503101 (cit. on p. 20).

[81]Zhen-Gang Wang. “50th Anniversary Perspective: Polymer Conformation A Pedagogical Review”. In: *Macromolecules* 50.23 (2017), pp. 9073–9114 (cit. on p. 20).

[82]Michael Rubinstein and Ralph H Colby. *Polymer physics*. Oxford university press, 2003 (cit. on pp. 20, 21).

[83]Alex S Holehouse and Rohit V Pappu. “Collapse transitions of proteins and the interplay among backbone, sidechain, and solvent interactions”. In: *Annual Review of biophysics* 47.1 (2018), pp. 19–39 (cit. on p. 20).

[84]Pierre-Gilles De Gennes. *Scaling concepts in polymer physics*. Cornell university press, 1979 (cit. on p. 21).

[85]IWAO Teraoka and Polymer Solutions. “An introduction to physical properties”. In: *Polymer Solutions* (2002) (cit. on pp. 21, 22).

[86]Wenbing Hu. *Polymer physics: a molecular approach*. Springer Science & Business Media, 2012 (cit. on p. 22).

[87]Mustapha Carab Ahmed, Ramon Crehuet, and Kresten Lindorff-Larsen. “Computing, analyzing, and comparing the radius of gyration and hydrodynamic radius in conformational ensembles of intrinsically disordered proteins”. In: *Intrinsically disordered proteins: methods and protocols* (2020), pp. 429–445 (cit. on pp. 22, 77, 82).

[88]CI Addison, AA Louis, and JP Hansen. “Influence of solvent quality on polymer solutions: A Monte Carlo study of bulk and interfacial properties”. In: *The Journal of chemical physics* 121.1 (2004), pp. 612–620 (cit. on p. 22).

[89]Edward Yu Bormashenko. *Physics of wetting: Phenomena and applications of fluids on surfaces*. Walter de Gruyter GmbH & Co KG, 2017 (cit. on pp. 23, 24).

[90]Benjamin Widom. “Capillarity and wetting phenomena: drops, bubbles, pearls, waves”. In: *Physics Today* 57.12 (2004), pp. 66–67 (cit. on p. 24).

[91]Tharwat F Tadros. “Emulsion formation, stability, and rheology”. In: *Emulsion formation and stability* (2013), pp. 1–75 (cit. on p. 26).

[92]Jan Guzowski, Piotr M Korczyk, Slawomir Jakiela, and Piotr Garstecki. “The structure and stability of multiple micro-droplets”. In: *Soft Matter* 8.27 (2012), pp. 7269–7278 (cit. on p. 26).

[93]S. Torza and S. G. Mason. “Coalescence of two immiscible liquid drops”. In: *Science* 163.3869 (1969), pp. 813–814 (cit. on p. 26).

[94]A Werner, F Schmid, M Müller, and K Binder. ““Intrinsic” profiles and capillary waves at homopolymer interfaces: A Monte Carlo study”. In: *Physical Review E* 59.1 (1999), p. 728 (cit. on pp. 28, 44).

[95]Martin-D Lacasse, Gary S Grest, and Alex J Levine. “Capillary-wave and chain-length effects at polymer/polymer interfaces”. In: *Physical review letters* 80.2 (1998), p. 309 (cit. on p. 28).

[96]Ting Ge, Mark O Robbins, Dvora Perahia, and Gary S Grest. “Healing of polymer interfaces: Interfacial dynamics, entanglements, and strength”. In: *Physical Review E* 90.1 (2014), p. 012602 (cit. on p. 29).

[97]Scott A Hollingsworth and Ron O Dror. “Molecular dynamics simulation for all”. In: *Neuron* 99.6 (2018), pp. 1129–1143 (cit. on pp. 29, 33).

[98]Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2023 (cit. on pp. 29, 31).

[99]Michael P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford university press, 2017 (cit. on pp. 29, 31).

[100]Yi Isaac Yang, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Qin Gao. “Enhanced sampling in molecular dynamics”. In: *The Journal of chemical physics* 151.7 (2019) (cit. on p. 32).

[101]Jérôme Hénin, Tony Lelièvre, Michael R Shirts, Omar Valsson, and Lucie Delemotte. “Enhanced sampling methods for molecular dynamics simulations”. In: *arXiv preprint arXiv:2202.04164* (2022) (cit. on p. 32).

[102]Yuji Sugita and Yuko Okamoto. “Replica-exchange molecular dynamics method for protein folding”. In: *Chemical physics letters* 314.1-2 (1999), pp. 141–151 (cit. on p. 32).

[103]Alessandro Laio and Michele Parrinello. “Escaping free-energy minima”. In: *Proceedings of the national academy of sciences* 99.20 (2002), pp. 12562–12566 (cit. on p. 32).

[104]Thomas E Gartner III and Arthi Jayaraman. “Modeling and simulations of polymers: a roadmap”. In: *Macromolecules* 52.3 (2019), pp. 755–786 (cit. on p. 33).

[105]Thorben Fröhlking, Mattia Bernetti, Nicola Calonaci, and Giovanni Bussi. “Toward empirical force fields that match experimental observables”. In: *The Journal of chemical physics* 152.23 (2020) (cit. on p. 33).

[106]David A Case, Thomas E Cheatham III, Tom Darden, et al. “The Amber biomolecular simulation programs”. In: *Journal of computational chemistry* 26.16 (2005), pp. 1668–1688 (cit. on p. 33).

[107]Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, et al. “CHARMM: the biomolecular simulation program”. In: *Journal of computational chemistry* 30.10 (2009), pp. 1545–1614 (cit. on p. 33).

[108]William George Noid. “Perspective: Coarse-grained models for biomolecular systems”. In: *The Journal of chemical physics* 139.9 (2013) (cit. on p. 33).

[109]Marissa G Saunders and Gregory A Voth. “Coarse-graining methods for computational biology”. In: *Annual review of biophysics* 42.1 (2013), pp. 73–93 (cit. on p. 33).

[110]Florian Müller-Plathe. “Coarse-graining in polymer simulation: from the atomistic to the mesoscopic scale and back”. In: *ChemPhysChem* 3.9 (2002), pp. 754–769 (cit. on p. 33).

[111]Satyen Dhamankar and Michael A Webb. “Chemically specific coarse-graining of polymers: Methods and prospects”. In: *Journal of Polymer Science* 59.22 (2021), pp. 2613–2643 (cit. on p. 33).

[112]Christine Peter and Kurt Kremer. “Multiscale simulation of soft matter systems–from the atomistic to the coarse-grained level and back”. In: *Soft Matter* 5.22 (2009), pp. 4357–4366 (cit. on p. 33).

[113]K Michael Salerno, Anupriya Agrawal, Dvora Perahia, and Gary S Grest. “Resolving dynamic properties of polymers through coarse-grained computational studies”. In: *Physical review letters* 116.5 (2016), p. 058302 (cit. on p. 33).

[114]Daniel T Gillespie. “Simulation methods in systems biology”. In: *Formal Methods for Computational Systems Biology: 8th International School on Formal Methods for the Design of Computer, Communication, and Software Systems, SFM 2008 Bertinoro, Italy, June 2-7, 2008 Advanced Lectures 8* (2008), pp. 125–167 (cit. on p. 34).

[115]Daniel T Gillespie. “Stochastic chemical kinetics”. In: *Handbook of Materials Modeling: Methods*. Springer, 2005, pp. 1735–1752 (cit. on p. 34).

[116]Daniel T Gillespie. “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. In: *Journal of computational physics* 22.4 (1976), pp. 403–434 (cit. on p. 35).

[117]Daniel T Gillespie. “Exact stochastic simulation of coupled chemical reactions”. In: *The journal of physical chemistry* 81.25 (1977), pp. 2340–2361 (cit. on p. 35).

[118]WM Young and EW Elcock. “Monte Carlo studies of vacancy migration in binary ordered alloys: I”. In: *Proceedings of the Physical Society* 89.3 (1966), p. 735 (cit. on p. 35).

[119]Soyoung Jeon, Yeram Jeon, Ji-Youn Lim, et al. “Emerging regulatory mechanisms and functions of biomolecular condensates: implications for therapeutic targets”. In: *Signal Transduction and Targeted Therapy* 10.1 (2025), p. 4 (cit. on p. 35).

[120]Danny M Hatters. “Grand challenges in biomolecular condensates: structure, function, and formation”. In: *Frontiers in Biophysics* 1 (2023), p. 1208763 (cit. on p. 35).

[121]Yongdae Shin and Clifford P Brangwynne. “Liquid phase condensation in cell physiology and disease”. In: *Science* 357.6357 (2017), eaaf4382 (cit. on p. 35).

[122]Sebastian Coupe and Nikta Fakhri. “Nonequilibrium phases of a biomolecular condensate facilitated by enzyme activity”. In: *bioRxiv* (2024) (cit. on p. 35).

[123]Kadi L Saar, Daoyuan Qian, Lydia L Good, et al. “Theoretical and data-driven approaches for biomolecular condensates”. In: *Chemical Reviews* 123.14 (2023), pp. 8988–9009 (cit. on p. 35).

[124]Giacomo Janson, Gilberto Valdes-Garcia, Lim Heo, and Michael Feig. “Direct generation of protein conformational ensembles via machine learning”. In: *Nature Communications* 14.1 (2023), p. 774 (cit. on p. 36).

[125]Matteo Aldeghi and Connor W Coley. “A focus on simulation and machine learning as complementary tools for chemical space navigation”. In: *Chemical Science* 13.28 (2022), pp. 8221–8223 (cit. on p. 37).

[126]Garrett M Ginell, Ryan J Emenecker, Jeffrey M Lotthammer, Emery T Usher, and Alex S Holehouse. “Direct prediction of intermolecular interactions driven by disordered regions”. In: *bioRxiv* (2024) (cit. on p. 37).

[127]Sören von Bülow, Giulio Tesei, Fatima Kamal Zaidi, Tanja Mittag, and Kresten Lindorff-Larsen. “Prediction of phase-separation propensities of disordered proteins from sequence”. In: *Proceedings of the National Academy of Sciences* 122.13 (2025), e2417920122 (cit. on p. 37).

[128]Yaxin An, Michael A Webb, and William M Jacobs. “Active learning of the thermodynamics-dynamics trade-off in protein condensates”. In: *Science Advances* 10.1 (2024), eadj2448 (cit. on p. 37).

[129]William M. Jacobs and Daan Frenkel. “Phase Transitions in Biological Systems with Many Components”. In: *Biophys. J.* 112.4 (2023/01/03 2017), pp. 683–691 (cit. on p. 40).

[130]David Zwicker and Liedewij Laan. “Evolved interactions stabilize many coexisting phases in multicomponent liquids”. In: *Proc. Natl. Acad. Sci. U. S. A.* 119.28 (2022), e2201250119 (cit. on p. 40).

[131]Hemali P. Phatnani and Arno L. Greenleaf. “Phosphorylation and functions of the RNA polymerase II CTD”. In: *Genes Dev.* 20.21 (2006), pp. 2922–2936. eprint: `http://genesdev.cshlp.org/content/20/21/2922.full.pdf+html` (cit. on pp. 40, 50).

[132]Patrick Cramer. “Organization and regulation of gene transcription”. In: *Nature* 573 (7772 2019), pp. 45–54 (cit. on p. 40).

[133]Kevin M. Harlen and L. Stirling Churchman. “The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain”. In: *Nat. Rev. Mol. Cell Biol.* 18.4 (2017), pp. 263–273 (cit. on pp. 40, 50, 90).

[134]Porfirio Quintero-Cadena, Tineke L. Lenstra, and Paul W. Sternberg. “RNA Pol II Length and Disorder Enable Cooperative Scaling of Transcriptional Bursting”. In: *Molecular Cell* 79.2 (2020), 207–220.e8 (cit. on pp. 40, 67).

[135]Yick Hin Ling, Ziyang Ye, Chloe Liang, et al. “Disordered C-terminal domain drives spatiotemporal confinement of RNAPII to enhance search for chromatin targets”. In: *Nature Cell Biology* 26.4 (2024), pp. 581–592 (cit. on p. 40).

[136]Denes Hnisz, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp. “A Phase Separation Model for Transcriptional Control”. In: *Cell* 169 (1 2017), pp. 13–23 (cit. on p. 40).

[137]Ann Boija, Isaac A. Klein, Benjamin R. Sabari, et al. “Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains”. In: *Cell* 175.7 (2018), 1842–1855.e16 (cit. on p. 40).

[138]Ron O. Dror, Robert M. Dirks, J.P. Grossman, Huafeng Xu, and David E. Shaw. “Biomolecular Simulation: A Computational Microscope for Molecular Biology”. In: *Annu. Rev. Biophys.* 41.1 (2012). PMID: 22577825, pp. 429–452. eprint: `https://doi.org/10.1146/annurev-biophys-042910-155245` (cit. on p. 41).

[139]Gregory L. Dignon, Wenwei Zheng, Young C. Kim, Robert B. Best, and Jeetain Mittal. “Sequence determinants of protein phase behavior from a coarse-grained model”. In: *PLOS Comput. Biol.* 14.1 (Jan. 2018), pp. 1–23 (cit. on pp. 41, 96).

[140]Zakarya Benayad, Sören von Bülow, Lukas S. Stelzl, and Gerhard Hummer. “Simulation of FUS Protein Condensates with an Adapted Coarse-Grained Model”. In: *J. Chem. Theory Comput.* 17.1 (Jan. 2021), pp. 525–537 (cit. on pp. 41, 45, 58, 61).

[141]Giulio Tesei, Thea K. Schulze, Ramon Crehuet, and Kresten Lindorff-Larsen. “Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties”. In: *Proc. Natl. Acad. Sci. U. S. A.* 118.44 (2021). eprint: https://www.pnas.org/content/118/44/e2111696118.full.pdf (cit. on pp. 41, 43).

[142]Roshan Mammen Regy, Gregory L Dignon, Wenwei Zheng, Young C Kim, and Jeetain Mittal. “Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations”. In: *Nucleic Acids Research* 48.22 (Dec. 2020), pp. 12593–12603. eprint: https://academic.oup.com/nar/article-pdf/48/22/12593/34906162/gkaa1099.pdf (cit. on p. 41).

[143]Steven Boeynaems, Alex S Holehouse, Venera Weinhardt, et al. “Spontaneous driving forces give rise to protein- RNA condensates with coexisting phases and complex material properties”. In: *Proc. Natl. Acad. Sci. U. S. A.* 116.16 (2019), pp. 7889–7898 (cit. on p. 41).

[144]Shiv Rekhi, Cristobal Garcia Garcia, Mayur Barai, et al. “Expanding the molecular language of protein liquid-liquid phase separation”. In: *bioRxiv* (2023), pp. 2023–03 (cit. on p. 41).

[145]Lara A Gruijs da Silva, Francesca Simonetti, Saskia Hutten, et al. “Disease-linked TDP-43 hyperphosphorylation suppresses TDP-43 condensation and aggregation”. In: *EMBO J.* 41.n/a (), e108443 (cit. on p. 41).

[146]Stephen E Farr, Esmae J Woods, Jerelle A Joseph, Adiran Garaizar, and Rosana Collepardo-Guevara. “Nucleosome plasticity is a critical element of chromatin liquid–liquid phase separation and multivalent nucleosome interactions”. In: *Nature communications* 12.1 (2021), p. 2883 (cit. on p. 41).

[147]Arya Changiarath, David Flores-Solis, Jasper J Michels, et al. “Promoter and Gene-Body RNA-Polymerase II co-exist in partial demixed condensates”. In: *bioRxiv* (2024), pp. 2024–03 (cit. on pp. 41, 47–49, 51, 52, 54, 57, 59–62, 64, 66).

[148]Theodora Myrto Perdikari, Nina Jovic, Gregory L. Dignon, et al. “A predictive coarse-grained model for position-specific effects of post-translational modifications”. In: *Biophys. J.* 120.7 (2021), pp. 1187–1197 (cit. on pp. 42, 96).

[149]Gregory L. Dignon, Wenwei Zheng, Young C. Kim, and Jeetain Mittal. “Temperature-Controlled Liquid–Liquid Phase Separation of Disordered Proteins”. In: *ACS Central Science* 5.5 (May 2019), pp. 821–830 (cit. on pp. 42, 43).

[150]J. A. Anderson, J. Glaser, and S. C. Glotzer. “HOOMD-blue: A Python Package for High-Performance Molecular Dynamics and Hard Particle Monte Carlo Simulations”. In: *Comput. Mater. Sci.* 173 (2020), p. 109363 (cit. on pp. 42, 96).

[151]Viktor Hornak, Robert Abel, Asim Okur, et al. “Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters.” In: *Proteins* 65.3 (Nov. 2006), pp. 712–725 (cit. on p. 43).

[152]Robert B. Best and Gerhard Hummer. “Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides”. In: *J. Phys. Chem. B* 113.26 (2009), pp. 9004–9015 (cit. on p. 43).

[153]Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, et al. “Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field”. In: *Proteins: Struct., Funct., Bioinf.* 78.8 (2010), pp. 1950–1958 (cit. on p. 43).

[154]Robert B Best, David De Sancho, and Jeetain Mittal. “Residue-Specific α-Helix Propensities From Molecular Simulation”. In: *Biophys. J.* 102.6 (2012), pp. 1462–1467 (cit. on p. 43).

[155]Stefano Piana, Alexander G. Donchev, Paul Robustelli, and David E. Shaw. “Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States”. In: *J. Phys. Chem. B* 119.16 (2015), pp. 5113–5123. eprint: NIHMS150003 (cit. on p. 43).

[156]Giovanni Bussi, Davide Donadio, and Michele Parrinello. “Canonical sampling through velocity rescaling”. In: *J. Chem. Phys.* 126.1 (2007), p. 014101. eprint: https://doi.org/10.1063/1.2408420 (cit. on p. 43).

[157]Michele Parrinello and Aneesur Rahman. “Polymorphic transitions in single crystals: A new molecular dynamics method”. In: *J. Appl. Phys.* 52.12 (1981), pp. 7182–7190 (cit. on p. 43).

[158]Mark James Abraham, Teemu Murtola, Roland Schulz, et al. “GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism From Laptops to Supercomputers”. In: *SoftwareX* 1-2 (2015), pp. 19–25 (cit. on p. 43).

[159]Paul J Flory. *Principles of polymer chemistry*. Cornell university press, 1953 (cit. on pp. 44, 55).

[160]Roland Horst. “Calculation of phase diagrams not requiring the derivatives of the Gibbs energy demonstrated for a mixture of two homopolymers with the corresponding copolymer”. In: *Macromol. Theory Sim.* 4.3 (1995), pp. 449–458. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mats.1995.040040304 (cit. on p. 44).

[161]D. Janka Bauer, Lukas S. Stelzl, and Arash Nikoubashman. “Single-chain and condensed-state behavior of hnRNPA1 from molecular simulations”. In: *J. Chem. Phys.* 157.15 (Oct. 2022), p. 154903. eprint: https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/5.0105540/16554276/154903_1_online.pdf (cit. on pp. 45, 46, 56, 63).

[162]JR Henderson and J Lekner. “Surface oscillations and the surface thickness of classical and quantum droplets”. In: *Mol. Phys.* 36.3 (1978), pp. 781–789 (cit. on p. 45).

[163]Marc Boehning, Claire Dugast-Darzacq, Marija Rankovic, et al. “RNA polymerase II clustering through carboxy-terminal domain phase separation”. In: *Nat. Struct. Mol. Biol.* 25.9 (2018), pp. 833–840 (cit. on pp. 49, 52).

[164]David Flores-Solis, Irina P. Lushpinskaia, Anton A. Polyansky, et al. “Driving forces behind phase separation of the carboxy-terminal domain of RNA polymerase II”. In: *Nat. Commun.* 14.1 (2023), p. 5979 (cit. on pp. 49, 50).

[165]Feiyue Lu, Bede Portz, and David S Gilmour. “The C-Terminal Domain of RNA Polymerase II Is a Multivalent Targeting Sequence that Supports Drosophila Development with Only Consensus Heptads”. In: *Mol. Cell* 73 (6 Mar. 2019). doi: 10.1016/j.molcel.2019.01.008, 1232–1242.e4 (cit. on pp. 50, 67).

[166]Xiangze Zeng, Alex S Holehouse, Ashutosh Chilkoti, Tanja Mittag, and Rohit V Pappu. “Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins”. In: *Biophys. J.* 119.2 (2020), pp. 402–418 (cit. on p. 50).

[167]Kiersten M Ruff, Stefan Roberts, Ashutosh Chilkoti, and Rohit V Pappu. “Advances in understanding stimulus-responsive phase behavior of intrinsically disordered protein polymers”. In: *J. Mol. Biol.* 430.23 (2018), pp. 4619–4635 (cit. on p. 50).

[168]Stefan Roberts, Tyler S Harmon, Jeffrey L Schaal, et al. “Injectable tissue integrating networks from recombinant polypeptides with tunable order”. In: *Nature Mat.* 17.12 (2018), pp. 1154–1163 (cit. on p. 50).

[169]Martin J Fossat, Xiangze Zeng, and Rohit V Pappu. “Uncovering differences in hydration free energies and structures for model compound mimics of charged side chains of amino acids”. In: *J. Phys. Chem. B* 125.16 (2021), pp. 4148–4161 (cit. on p. 50).

[170]Xiangze Zeng, Chengwen Liu, Martin J Fossat, et al. “Design of intrinsically disordered proteins that undergo phase transitions with lower critical solution temperatures”. In: *APL Mater.* 9.2 (2021) (cit. on p. 50).

[171]Kiersten M. Ruff, Stefan Roberts, Ashutosh Chilkoti, and Rohit V. Pappu. “Advances in Understanding Stimulus-Responsive Phase Behavior of Intrinsically Disordered Protein Polymers”. In: *J. Mol. Biol.* 430.23 (2018), pp. 4619–4635 (cit. on p. 50).

[172]Colin E. Delaney, Stephen P. Methot, Veronique Kalck, et al. “SETDB1-like MET-2 promotes transcriptional silencing and development independently of its H3K9me-associated catalytic activity”. In: *Nat. Stru. Mol. Biol.* 29.2 (2022), pp. 85–96 (cit. on p. 50).

[173]Adam G. Larson, Daniel Elnatan, Madeline M. Keenen, et al. “Liquid droplet formation by HP1suggests a role for phase separation in heterochromatin”. In: *Nature* 547.7662 (2017), pp. 236–240 (cit. on p. 50).

[174]Huasong Lu, Dan Yu, Anders S. Hansen, et al. “Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II”. In: *Nature* 558.7709 (2018), pp. 318–323 (cit. on pp. 53, 106).

[175]Agnieszka Pancholi, Tim Klingberg, Weichun Zhang, et al. “RNA polymerase II clusters form in line with surface condensation on regulatory chromatin”. In: *Mol. Syst. Biol.* 17.9 (2021), e10272. eprint: `https://www.embopress.org/doi/pdf/10.15252/msb.202110272` (cit. on p. 53).

[176]Yang Eric Guo, John C Manteiga, Jonathan E Henninger, et al. “Pol II phosphorylation regulates a switch between transcriptional and splicing condensates”. In: *Nature* 572 (7770 2019), pp. 543–548 (cit. on pp. 53, 67).

[177]Gregory L. Dignon, Robert B. Best, and Jeetain Mittal. “Biomolecular Phase Separation: From Molecular Driving Forces to Macroscopic Properties”. In: *Ann. Rev. Phys. Chem.* 71.1 (2020). PMID: 32312191, pp. 53–75. eprint: `https://doi.org/10.1146/annurev-physchem-071819-113553` (cit. on p. 53).

[178]Jasper J. Michels, Mateusz Brzezinski, Tom Scheidt, Edward A. Lemke, and Sapun H. Parekh. “Role of Solvent Compatibility in the Phase Behavior of Binary Solutions of Weakly Associating Multivalent Polymers”. In: *Biomacromolecules* 23.1 (Jan. 2022), pp. 349–364 (cit. on p. 55).

[179]A. Werner, F. Schmid, M. Müller, and K. Binder. ""Intrinsic" profiles and capillary waves at homopolymer interfaces: A Monte Carlo study". In: *Phys. Rev. E* 59 (1 Jan. 1999), pp. 728–738 (cit. on pp. 58, 60).

[180]S Torza and SG Mason. "Three-phase interactions in shear and electrical fields". In: *J. Colloid Interface Sci.* 33.1 (1970), pp. 67–83 (cit. on p. 58).

[181]Taranpreet Kaur, Muralikrishna Raju, Ibraheem Alshareedah, et al. "Sequence-encoded and composition-dependent protein-RNA interactions control multiphasic condensate morphologies". In: *Nat. Commun.* 12.1 (2021), pp. 1–16 (cit. on pp. 58, 90).

[182]Dongmei Liu, Ye Lin, Kai Gong, et al. "Phase behavior and interfacial tension of ternary polymer mixtures with block copolymers". In: *RSC. Adv.* 11.61 (2021), pp. 38316–38324 (cit. on p. 61).

[183]Ming-Tzo Wei, Yi-Che Chang, Shunsuke F. Shimobayashi, et al. "Nucleated transcriptional condensates amplify gene expression". In: *Nat. Cell Biol.* 22.10 (2020), pp. 1187–1196 (cit. on pp. 63, 67, 90).

[184]Eun Yong Shim, Amy K. Walker, Yang Shi, and T. Keith Blackwell. "CDK-9/cyclin T (P-TEFb) is required in two postinitiation pathways for transcription in the C. elegans embryo". In: *Genes Dev.* 16.16 (2002), pp. 2135–2146. eprint: `http://genesdev.cshlp.org/content/16/16/2135.full.pdf+html` (cit. on p. 63).

[185]Mina Farag, Samuel R. Cohen, Wade M. Borcherds, et al. "Condensates formed by prion-like low-complexity domains have small-world network structures and interfaces defined by expanded conformations". In: *Nature Commun.* 13.1 (2022), p. 7722 (cit. on p. 63).

[186]Leonidas Emmanouilidis, Laura Esteban-Hofer, Fred F. Damberger, et al. "NMR and EPR reveal a compaction of the RNA-binding protein FUS upon droplet formation". In: *Nat. Chem. Biol.* 17.5 (2021), pp. 608–614 (cit. on p. 63).

[187]Gregory L. Dignon, Wenwei Zheng, Robert B. Best, Young C. Kim, and Jeetain Mittal. "Relation between single-molecule properties and phase behavior of intrinsically disordered proteins". In: *Proc. Natl. Acad. Sci. U. S. A.* 115.40 (2018), pp. 9929–9934 (cit. on p. 63).

[188]Jiahui Wang, Dinesh Sundaravadivelu Devarajan, Arash Nikoubashman, and Jeetain Mittal. "Conformational Properties of Polymers at Droplet Interfaces as Model Systems for Disordered Proteins". In: *ACS Macro Letters* 12.11 (2023). PMID: 37856873, pp. 1472–1478. eprint: `https://doi.org/10.1021/acsmacrolett.3c00456` (cit. on p. 65).

[189]Anton Meinhart and Patrick Cramer. "Recognition of RNA polymerase II carboxy-terminal domain by 3-RNA-processing factors". In: *Nature* 430.6996 (2004), pp. 223–226 (cit. on pp. 70, 71).

[190]David Flores-Solis, Irina P Lushpinskaia, Anton A Polyansky, et al. "Driving forces behind phase separation of the carboxy-terminal domain of RNA polymerase II". In: *Nature Communications* 14.1 (2023), p. 5979 (cit. on pp. 69–71, 87).

[191]BR Sabari, A Dall'Agnese, A Boija, et al. *Coactivator condensation at super-enhancers links phase separation and gene control. Science*. 2018 (cit. on p. 69).

[192]Anton Meinhart, Tomislav Kamenski, Sabine Hoeppner, Sonja Baumli, and Patrick Cramer. "A structural perspective of CTD function". In: *Genes & development* 19.12 (2005), pp. 1401–1415 (cit. on p. 69).

[193]Joseph A Marsh and Julie D Forman-Kay. “Sequence determinants of compaction in intrinsically disordered proteins”. In: *Biophysical journal* 98.10 (2010), pp. 2383–2390 (cit. on p. 69).

[194]Lisa M Pietrek, Lukas S Stelzl, and Gerhard Hummer. “Structural ensembles of disordered proteins from hierarchical chain growth and simulation”. In: *Current opinion in structural biology* 78 (2023), p. 102501 (cit. on pp. 71, 72).

[195]Lisa M Pietrek, Lukas S Stelzl, and Gerhard Hummer. “Hierarchical ensembles of intrinsically disordered proteins at atomic resolution in molecular dynamics simulations”. In: *Journal of chemical theory and computation* 16.1 (2019), pp. 725–737 (cit. on pp. 73, 76).

[196]Chris Neale, Régis Pomès, and Angel E García. “Peptide bond isomerization in high-temperature simulations”. In: *Biophysical Journal* 110.3 (2016), 644a (cit. on p. 73).

[197]Sandro Bottaro, Giovanni Bussi, Scott D Kennedy, Douglas H Turner, and Kresten Lindorff-Larsen. “Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations”. In: *Science advances* 4.5 (2018), eaar8521 (cit. on p. 75).

[198]Markus Zweckstetter. “NMR: prediction of molecular alignment from structure using the PALES software”. In: *Nature protocols* 3.4 (2008), pp. 679–690 (cit. on pp. 75, 85).

[199]Lukas S Stelzl, Lisa M Pietrek, Andrea Holla, et al. “Global structure of the intrinsically disordered protein tau emerges from its local structure”. In: *Jacs Au* 2.3 (2022), pp. 673–686 (cit. on p. 76).

[200]Charles K. Fisher, Austin Huang, and Collin M. Stultz. “Modeling Intrinsically Disordered Proteins with Bayesian Statistics”. In: *Journal of the American Chemical Society* 132.42 (2010). PMID: 20925316, pp. 14919–14927. eprint: `https://doi.org/10.1021/ja105832g` (cit. on p. 76).

[201]Yang Shen and Ad Bax. “SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network”. In: *Journal of biomolecular NMR* 48 (2010), pp. 13–22 (cit. on p. 76).

[202]Jakob Toudahl Nielsen and Frans AA Mulder. “POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins”. In: *Journal of biomolecular NMR* 70 (2018), pp. 141–165 (cit. on p. 76).

[203]F Kramer, MV Deshmukh, H Kessler, and SJ Glaser. “Residual dipolar coupling constants: An elementary derivation of key equations”. In: *Concepts in Magnetic Resonance Part A: An Educational Journal* 21.1 (2004), pp. 10–21 (cit. on p. 77).

[204]Patrick Seyfried, Marcel Heinz, György Pintér, et al. “Optimal Destabilization of DNA Double Strands by Single-Nucleobase Caging”. In: *Chemistry–A European Journal* 24.66 (2018), pp. 17568–17576 (cit. on pp. 78, 81).

[205]Chenliang Xia, Wenbin Kang, Jun Wang, and Wei Wang. “Temperature Dependence of Internal Friction of Peptides”. In: *The Journal of Physical Chemistry B* 125.11 (2021), pp. 2821–2832 (cit. on p. 78).

[206]Rhiju Das and David Baker. “Macromolecular modeling with rosetta”. In: *Annual Review of Biochemistry* 77.1 (2008), pp. 363–382 (cit. on p. 81).

[207] Gerhard Hummer and Jürgen Köfinger. “Bayesian ensemble refinement by replica simulations and reweighting”. In: *The Journal of Chemical Physics* 143.24 (Dec. 2015), p. 243150. eprint: https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/1.4937786/15510660/243150_1_online.pdf (cit. on p. 81).

[208] Lilia M. Iakoucheva, Predrag Radivojac, Celeste J. Brown, et al. “The importance of intrinsic disorder for protein phosphorylation”. In: *Nucleic Acids Research* 32.3 (Feb. 2004), pp. 1037–1049. eprint: https://academic.oup.com/nar/article-pdf/32/3/1037/9489921/gkh253.pdf (cit. on p. 89).

[209] Wilton T Snead and Amy S Gladfelter. “The control centers of biomolecular phase separation: how membrane surfaces, PTMs, and active processes regulate condensation”. In: *Molecular cell* 76.2 (2019), pp. 295–305 (cit. on p. 89).

[210] Srivastav Ranganathan, Pouria Dasmeh, Seth Furniss, and Eugene Shakhnovich. “Phosphorylation sites are evolutionary checkpoints against liquid–solid transition in protein condensates”. In: *Proceedings of the National Academy of Sciences* 120.20 (2023), e2215828120. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2215828120 (cit. on pp. 89, 90).

[211] Lara A Gruijs da Silva, Francesca Simonetti, Saskia Hutten, et al. “Disease-linked TDP-43 hyperphosphorylation suppresses TDP-43 condensation and aggregation”. In: *The EMBO Journal* 41.8 (2022), e108443. eprint: https://www.embopress.org/doi/pdf/10.15252/embj.2021108443 (cit. on pp. 89, 90).

[212] Michael L Nosella and Julie D Forman-Kay. “Phosphorylation-dependent regulation of messenger RNA transcription, processing and translation within biomolecular condensates”. In: *Current opinion in cell biology* 69 (2021), pp. 30–40 (cit. on p. 89).

[213] David Zwicker. “The intertwined physics of active chemical reactions and phase separation”. In: *Current Opinion in Colloid amp; Interface Science* 61 (Oct. 2022), p. 101606 (cit. on p. 90).

[214] Denes Hnisz, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp. “A Phase Separation Model for Transcriptional Control”. In: *Cell* 169.1 (2017), pp. 13–23 (cit. on pp. 90, 106).

[215] Patrick Cramer. “Organization and regulation of gene transcription”. In: *Nature* 573.7772 (2019), pp. 45–54 (cit. on p. 90).

[216] Mark D. Driver and Patrick R. Onck. “Selective phase separation of transcription factors is driven by orthogonal molecular grammar”. In: *bioRxiv* (2024). eprint: https://www.biorxiv.org/content/early/2024/04/15/2024.04.12.589262.full.pdf (cit. on p. 90).

[217] Marc Boehning, Claire Dugast-Darzacq, Marija Rankovic, et al. “RNA polymerase II clustering through carboxy-terminal domain phase separation”. In: *Nat. Struct. Mol. Biol.* 25.9 (2018), pp. 833–840 (cit. on p. 90).

[218] Ibrahim I. Cisse, Ignacio Izeddin, Sebastien Z. Causse, et al. “Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells”. In: *Science* 341.6146 (2013), pp. 664–667. eprint: https://www.science.org/doi/pdf/10.1126/science.1239053 (cit. on p. 90).

[219] Brian A. Lewis, Subhendu Kumar Das, Rajiv Kumar Jha, and David Levens. “Self-assembly of promoter DNA and RNA Pol II machinery into transcriptionally active biomolecular condensates”. In: *Science Advances* 9.42 (2023), eadi4565. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.adi4565 (cit. on p. 90).

[220]Katerina Linhartova, Francesco Luca Falginella, Martin Matl, et al. “Sequence and structural determinants of RNAPII CTD phase-separation and phosphorylation by CDK7”. In: *Nature Communications* 15.1 (2024), p. 9163 (cit. on p. 90).

[221]Arya Changiarath, Jasper J. Michels, Rosa Herrera Rodriguez, et al. “Conditions for the co-existence of promoter and gene-body condensates”. In: *bioRxiv* (2024). eprint: https://www.biorxiv.org/content/early/2024/03/17/2024.03.16.585180.full.pdf (cit. on pp. 90, 106).

[222]Adiran Garaizar, Jorge R. Espinosa, Jerelle A. Joseph, et al. “Aging can transform single-component protein condensates into multiphase architectures”. In: *Proceedings of the National Academy of Sciences* 119.26 (2022), e2119800119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2119800119 (cit. on p. 90).

[223]Christopher Brasnett, Armin Kiani, Selim Sami, Sijbren Otto, and Siewert J Marrink. “Capturing chemical reactions inside biomolecular condensates with reactive Martini simulations”. In: *Communications Chemistry* 7.1 (2024), p. 151 (cit. on p. 90).

[224]Emanuele Zippo, Dorothee Dormann, Thomas Speck, and Lukas S. Stelzl. “Molecular simulations of enzymatic phosphorylation of disordered proteins and their condensates”. In: *bioRxiv* (2024). eprint: https://www.biorxiv.org/content/early/2024/09/14/2024.08.15.607948.full.pdf (cit. on pp. 90, 96).

[225]Benedikt Rennekamp, Fabian Kutzki, Agnieszka Obarska-Kosinska, Christopher Zapp, and Frauke Gräter. “Hybrid Kinetic Monte Carlo/Molecular Dynamics Simulations of Bond Scissions in Proteins”. In: *Journal of Chemical Theory and Computation* 16.1 (2020). PMID: 31738552, pp. 553–563. eprint: https://doi.org/10.1021/acs.jctc.9b00786 (cit. on p. 90).

[226]Martin Girard and Tristan Bereau. “Computer simulations of lipid regulation by molecular semigrand canonical ensembles”. In: *Biophysical Journal* 120.12 (2021), pp. 2370–2373 (cit. on pp. 90, 96).

[227]Megan J. Bywater, Deborah L. Burkhart, Jasmin Straube, et al. “Reactivation of Myc transcription in the mouse heart unlocks its proliferative capacity”. In: *Nature Communications* 11.1 (2020), p. 1827 (cit. on p. 91).