Mechanochemical Understanding of Metal-Coordinated Polymers Using Simulation and Experiment

by

Eesha Khare

B.S., Harvard University (2017)
M.Phil., University of Cambridge Trinity College (2018)

Submitted to the Department of Materials Science and Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Materials Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2023

Author . . . . . .
Department of Materials Science and Engineering
May 19, 2023

Certified by . . . . . .
Markus J. Buehler
Professor of Civil and Environmental Engineering and Mechanical Engineering
Thesis Supervisor

Certified by . . . . . .
Niels Holten-Andersen
Associate Professor
Thesis Supervisor

Accepted by . . . . . .
Rob J. Macfarlane
Chair, Departmental Committee on Graduate Studies

Mechanochemical Understanding of Metal-Coordinated Polymers Using Simulation and Experiment

By

Eesha Khare

Submitted to the Department of Materials Science and Engineering on May 19, 2023
in Partial Fulfillment of the Requirements of the
Degree of Doctor of Philosophy in Materials Science and Engineering

ABSTRACT

Metal-coordination bonds have the capacity to reform after rupture, thereby enabling dynamic, tunable, and reversible (self-healing) mechanical properties. Several biological organisms, such as marine mussels (*Mytilus)* and marine worm jaws (*Nereis virens)*, have been found to take advantage of these unique properties of metal-coordinated complexes to produce loadbearing materials with complex mechanical functions. Inspired by these biological materials, metal-coordination bonds have been incorporated into synthetic materials to produce a range of mechanical properties. However, efforts in engineering such metal-coordinated materials have been highly empirical, limiting the full design potential of these bonds. Developing an understanding between the microscopic metal-coordination bond properties and resulting macroscopic mechanical properties of metal-coordinated materials would enable an *a priori* prediction for optimized utilization of coordination bonds to build materials with advanced mechanical functions.

This dissertation systematically characterizes metal-coordinated polymers and proteins with the aim of developing a mechanistic understanding of the relationship between microscopic bond chemistry and resulting macroscopic dynamic mechanical properties. We begin with a well-studied model system using an idealized polymer network where individual metal-coordination complexes control the macroscopic relaxation dynamics of the network. We use metadynamics simulations to show that the free energy landscape of metal-coordination bonds can be related to the macroscopic dynamic relaxation of these bonds in ideal polymer hydrogels as measured through experimental rheology. We then expand beyond single coordination complexes and use single molecule force spectroscopy to show that clusters of coordination bonds in model metal-coordinated protein dimers can rupture cooperatively, thereby synergistically increasing the rupture strength of the proteins. We resolve this rupture behavior mechanistically by using steered molecular dynamics simulations to show that metal-coordination bond rupture is highly heterogenous and undergoes several rupture pathways, even with the same initial conditions. This indicates that metal-coordination bonds may have evolved in natural materials for primarily dissipative functions.

The above insights are subsequently evaluated within the context of the Nvjp-1 protein, a major component of the *Nereis* worm jaw with high amounts of metal coordination. We find that increasing quantity of metal ions makes the protein more compact, whereas increasing the spatial distribution of metal ions is found to increase the protein toughness. We then briefly demonstrate how machine learning methods can be developed for similar systems to predict

materials properties. The methodology and insights developed in this thesis have important implications for understanding the molecular mechanisms of metal-coordination bond-based stabilization of proteins and polymers and the *a priori* design of new metal-coordinated materials with desired mechanical properties.

Thesis Supervisor: Markus Buehler
Title: Professor of Civil and Environmental Engineering and Mechanical Engineering

Thesis Supervisor: Niels Holten-Andersen
Title: Associate Professor

ACKNOWLEDGEMENTS

I would like to begin the thesis by acknowledging all the people, funding sources, and organizations that have made this work possible. I would first like to thank my thesis advisors, Professors Markus Buehler and Niels Holten-Andersen, without whom this work would have been impossible. Markus is an extremely creative thinker, and his level of support for my personal growth and development is incredible. Markus has spent countless hours helping me think through scientific ideas and discuss research questions, and I really appreciate his thoughtfulness and supportiveness throughout this journey. Niels is also a wonderful, supportive advisor, and I hope to have imbued some of his ability to think deeply about complex topics. Niels has also spent many hours working through challenging research concepts with me, and he has provided me unlimited academic freedom and ownership over my research. They say the most important decision in your PhD is who you select as an advisor, and I am extremely grateful to have been advised by Markus and Niels.

I have also been lucky to have had such an active thesis committee. Professor Kerstin Blank at the Max Planck Institute for Colloids and Interfaces opened her lab and time for me to explore concepts of cooperativity in metal-coordinated systems. I appreciate her willingness to take me on as a student, and have grown incredibly from her attention to detail, rigorous and careful approach to scientific thinking, and extensive knowledge of biomaterials and mechanics. Professor David Kaplan at Tufts University similarly opened up his lab and ideas for me to conduct much of the bacterial expression work related to producing metal-coordinated proteins. I've been meeting with David every month through a collaborative exchange in our labs, and have greatly appreciated how he always pushes me to think of the real-world relevance of my work. My final committee member, Professor Alfredo Alexander-Katz, directed the polymers academic program and taught me so much about polymer physics. I really appreciate that I have been able to go to him at any time to learn and grow from his immense knowledge in the field.

My former academic mentors and advisors are also critical people in my academic and personal trajectory. From my first academic mentor, Professor Elaine Collins at San Jose State University, I began to appreciate how to conduct scientific research, especially as a first-year student in high school. I also had an opportunity to conduct research under Professor Yat Li at the University of California Santa Cruz, where he gave me my first appreciation of the chemistry research that continued on to become my Intel International Science Fair project. Thank you also to my high school chemistry teacher Mr. Penner at Lynbrook, who really excited me about the fields of chemistry and materials science. Professors Stoyan Smoukov and David R. Clarke were my research advisors in college, and helped me grow as a scientific thinker. Both gave me independent projects to work on, even though I was in college, and enabled me to learn how to construct new ideas to contribute to the field. I really appreciate their hard work in preparing me for graduate school and beyond. I also wanted to acknowledge my mentors outside of research, from Kevin Hettrich and Tim Holme at QuantumScape, to David Hochschild at the California Energy Commission, to Michael Kearney, Vinny Beranek, and Chloe Holzinger at The Engine. All of these mentors encouraged me to think beyond the

academic sphere of science to try to understand how I could make an impact on the world at large.

My PhD labs, the Laboratory for Atomistic and Molecular Mechanics, the Laboratory for Bio-inspired Interfaces at MIT, the Tufts Molecular Biology group, and the MechanoBioChemistry group at the Max Planck Institute, were all welcoming and friendly. I have loved being a part of these groups and have learned an incredible amount both academically and socially from my group members. I would like to give a special thanks to Amadeus Alcantara, Jonny Yu, Constancio Ito Gonzalez, Zaira Martin-Moldes, Ayesha Talib, Jake Song, Seth Cazzell, Mario Milazzo, Diego Lopez, Bo Ni, Francisco Martin-Martinez, Nic Lee, Andrew Lew, and Sabrina Shen for being such excellent collaborators and teachers throughout my PhD journey.

More broadly, I would like to thank all of my friends from California, Harvard, Cambridge and MIT. My friends have always been excellent sources of fun and personal reflection. I would also like to give a special thanks to Team Taxie, with whom I worked closely over two years to propose a startup idea to address climate change. MIT has provided me with incredible opportunities such as this entrepreneurial experience, including financial support from the National Science Foundation Graduate Research Fellowship, the MIT Office of Graduate Education Collamore-Rogers Fellowship, the MISTI-Germany grants, and support for the DAAD German fellowship. Through my experience at MIT, I was also able to engage in the Graduate Student Council, Graduate Materials Council, diversity and equity initiatives, and the DMSE-DAAP program.

Of the people that I met in the past five years, I feel extremely lucky to have met Elad Deiss-Yehiely. Elad and I first met during orientation of the polymers program, and since that moment, he has provided me with so much laughter, support, kindness, and thoughtfulness. He is an incredibly reflective person, and he truly makes me become a better person every day. I'm happy I came to MIT, because I got to meet Elad.

Finally, and most importantly, I would like to thank my family and extended family, who are the reason I started a PhD in the first place, and have truly helped me get to where I am today. My parents Manoj and Reena Khare have made incredible sacrifices to help my sister and I have many opportunities. They have also always encouraged me to pursue many ideas, even beyond science, which I believe has helped me become a broader and more motivated thinker. My sister Somya, who is on track to becoming a Dr. -Dr., is also an incredibly caring person, and I appreciate her ability to have such a fresh perspective on the world, while encouraging me to do the same. Thank you so much to all my family, friends, and mentors for being such a great part of my PhD journey.

TABLE OF CONTENTS

6. Conclusions

7. Appendix and Glossary

8. Bibliography

LIST OF FIGURES AND TABLES

LIST OF SUPPLEMENTAL MATERIAL

LIST OF JOURNAL PUBLICATIONS

JOURNAL PUBLICATIONS REFERENCED IN THESIS

1. **E. Khare**, C. G. Obeso, Z. Martin-Moldes, A. Talib, D. L. Kaplan, N. Holten-Andersen, K. G. Blank. M. J. Buehler. Cooperative and heterogeneous rupture of rationally-designed metal-coordinated proteins. *Submitted.*

2. **E. Khare*,** J. Luo, M.J. Buehler, Localization of Zn^{2+} ions affects the structural folding and mechanics of Nereis virens Nvjp-1. *Soft Matter (*2023).

3. **E. Khare**, D. S. Grewal, M. J. Buehler, Bond clusters control rupture force limit in shear loaded histidine-Ni^{2+} metal-coordinated proteins. *Nanoscale* (2023).

4. J. Song*, **E. Khare***, L. Rao, M. J. Buehler and N. Holten-Andersen, Coordination stoichiometry effects on the binding hierarchy of histamine and imidazole-M^{2+} complexes. *Under review.*

5. **E. Khare,** S. A. Cazzell, J. Song, N. Holten-Andersen, M. J. Buehler, Molecular understanding of Ni2+-nitrogen family metal-coordinated hydrogel relaxation times using free energy landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2213160120 (2023).

6. **E. Khare,** C. Gonzalez-Obeso, D. L. Kaplan, M. J. Buehler, CollagenTransformer: End-to-End Transformer Model to Predict Thermal Stability of Collagen Triple Helices Using an NLP Approach. *ACS Biomater. Sci. Eng.* **8**, 4301–4310 (2022).

7. **E. Khare*,** C. H. Yu*, C. Gonzalez-Obeso* *et al.*, Discovering design principles of collagen molecular stability using a genetic algorithm, deep learning, and experimental validation. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2209524119 (2022).

8. C. H. Yu*, **E. Khare*,** *et al.*, ColGen: An end-to-end deep learning model to predict thermal stability of de novo collagen sequences. *J. Mech. Behav. Biomed. Mater.* **125**, 104921 (2022).

9. **E. Khare**, N. Holten-Andersen, M. J. Buehler, Transition-metal coordinate bonds for bioinspired macromolecules with tunable mechanical properties. *Nat. Rev. Mater.*, 1–16 (2021).

JOURNAL PUBLICATIONS NOT REFERENCED IN THESIS

10. M. Wysokowski, R. K. Luu, S. Arevalo, **E. Khare**, W. Stachowiak, M. Niemczak, T. Jesionowski, M. J. Buehler, Untapped Potential of Deep Eutectic Solvents for the Synthesis of Bioinspired Inorganic-Organic Materials. *Submitted.*

11. E. Khare, X. Peng, Z. Martín-Moldes, G. Genin, D. L. Kaplan, M. J. Buehler. Application of the Interagency and Modeling Analysis Group model verification approach for scientific reproducibility in a study of biomineralization. *ACS Biomater. Sci. Eng.* (2023)

12. S. C. Shen, **E. Khare**, N. A. Lee, M. K. Saad, D. L. Kaplan and M. J. Buehler, Computational Design and Manufacturing of Sustainable Materials through First-Principles and Materiomics. *Chem. Rev.* **123**, 2242–2275 (2022).

13. J. H. Ting*, **E. Khare***, *et al.*, Role of methylene diphenyl diisocyanate (MDI) additives on SBS-modified asphalt with improved thermal stability and mechanical performance. *Energy and Fuels* **35**, 17629–17641 (2021).

* Indicates equal contribution

CHAPTER 1

INTRODUCTION—ENGINEERING MATERIALS, TRANSITION METAL IONS, AND THESIS OVERVIEW

1.1 Engineering materials systems with control

Biology has evolved over millennia to develop materials with remarkable mechanical properties. Human bone and deer antlers, just two of many examples, exhibit both high stiffness and high resistance to fracture, properties that are often traded off in synthetic materials.[1,2] Inspired by such materials, scientists and engineers have sought to both uncover the wealth of mechanisms underpinning the properties of biological materials and build materials that mimic or surpass the properties demonstrated in biology. This rich relationship between engineering and biology is not limited to the disciplines of bioengineering or biomimicry; it permeates through engineering as a whole. One needs to look no further than the wing of an airplane or the porous design of a construction crane to find the vibrant dialogue between nature and engineering.[3]

In large part due to such natural inspiration, the development of new materials technologies has skyrocketed in the past decades, revolutionizing industries such as energy, robotics, and plastics. New materials in these sectors have traversed the mechanical boundaries between strength, stiffness, and toughness, all while achieving the unique electrical or optical properties required to make these materials functional for human use, thereby enabling economic and technological progress. However, the rapid advancement of materials development has also come with a large cost. The fast rate of materials growth has been concomitant with the production of waste due to massive resource use and environmentally unfriendly practices. The development of materials engineering with a strong focus on sustainable practices is now the forefront of materials engineering.

The sustainability of materials engineering can be addressed through several methods. These methods can be broadly divided into various stages of a material's lifetime from sustainable sourcing and synthesis, to materials property development, to end-of-life recycling or reuse. For example, building materials with more abundant and sustainably harvested constituents could reduce the dependence on oil-based chemicals that have large environmental consequences. At the back-end, developing effective collection and processing techniques for recycling or reusing materials could drastically reduce the initial harvesting of virgin materials. The middle ground, between the birth and the rebirth of a material, requires developing materials that maintain their desired properties over an extended lifetime. Not only is this component critical to materials circularity, but also, this component can be entirely controlled by materials scientists and engineers, unlike the birth and rebirth processes that are heavily dependent on political or economic infrastructures.

Increasing the lifetime of materials while maintaining their desired properties could address the waste or disposal of materials due to the degradation of the material itself or due to defects

that cause the material to break. In this thesis, we focus on the latter: improving the structural mechanical properties of materials so that they do not fracture and can thus have a longer usable period. Engineering materials with prolonged lifetimes through resistance to fracture requires harnessing the chemical properties of materials. The key question in the design of such materials is understanding how the properties at a microscale influence the macroscale. In other words, we are still learning how to control matter from the bottom-up: using chemistry to engineer materials with precision and control. This is the key question at the heart of materials design. This is the key question at the heart of this thesis.

In order to address the need for harnessing chemistries for improved structural properties, this thesis focuses on a broadly emerging class of bioinspired chemistries—metal-coordination bonds. Metal-coordination bonds are noncovalent bonds between metal cations and organic ligands and have been found in several biological organisms to yield remarkable structural mechanical properties such as high toughness or self-healing, due to their capacity to break and reform. If metal-coordination bonds could be implemented in a similar manner into synthetic materials, we could imagine a future of materials that self-heal structural defects without external input and never fracture, dramatically reducing waste from damaged materials. Despite the increasing adoption of metal-coordination bonds for mechanical function in synthetic materials, most literature to date has been heavily empirical, limiting the full design potential of these bonds. The aim of this thesis is twofold. First, we investigate how the chemistry of metal-coordination bonds can control the macroscopic properties of metal-coordinated polymers and proteins. Second, we translate this insight into clear design principles of how to use these bonds to achieve desired properties. Using computational and experimental approaches, we establish various design criteria to enable the *a priori* rational design of metal-coordinated materials for continued progress in sustainable materials development.

1.2 Transition metal ions for tunable properties

Transition metal ions are found in natural organisms, in large part owing to their unique ability to catalyze enzymatic processes (**Box 1**).[4–6] Although this catalytic role of metal ions in biology has been extensively researched, evidence is accumulating that metal-coordination complexes also play a prominent role in determining the mechanical properties of biological materials.[7–11] Several organisms, such as marine mussels and marine worms, have been found to take advantage of metal-coordination bonds in the assembly of hierarchical material structures, resulting in an exceptional range of strength, elasticity, hardness and other mechanical properties.[12–16] Two such properties afforded by metal coordination are high stiffness and high extensibility, resulting in high toughness.[2,17–19] In synthetic materials, there is often a tradeoff between these properties.[2] For example, increasing the number of crosslinks increases stiffness, but also decreases toughness by limiting extensibility.[20] Conversely, in biological materials, dynamic metal-coordinate bonds contribute to stiffness by acting as additional crosslinking points, but can also break and reform over time to enable extensibility and energy dissipation.[20] Therefore, mimicking natural metal-coordinate crosslinked molecular structures and learning from them could enable a wealth of new tough and stiff materials for use in impact resistance, self-healing or biomedical applications.

BOX 1. Other uses of metal coordination

Metal coordination is an old and deep field with several applications across traditional scientific disciplines. Metal coordination has been studied extensively in bioinorganic chemistry for enzymatic processes such as in heme metal centers in several proteins in which the local protein environment substantially affects metal reactivity,[488] or in systems in which metal coordination has a role in stabilizing protein structure.[489–491] The unique electronic properties of metal-coordination bonds have also been used in photochemistry to induce metal-to-ligand charge transfer in excited states.[492] Applications of metal coordination include its use in metal–organic frameworks for hydrogen storage,[493] adsorption with controlled porosity[494] and catalysis[495]. In these systems, metal coordination offers a large range of geometries, sizes and functionalities.[496] Metal coordination has also been used in the field of supramolecular chemistry [497–499] to control the topology of dynamic supramolecular polymers,[500] assembly of small molecules[501,502] and polymers,[503,504] and activity of enzymes.[505] Although metal coordination applied to mechanical properties has not been studied as broadly, we encourage readers to explore the inorganic and bioinorganic literature to understand the fundamental chemical features of metal-coordination bonds. In particular, more information about coordination geometry, binding constants and other properties of metal-coordination complexes is found in Refs [111,126,130].

Metal ions coordinate with ligands to form metal-coordination bonds that involve the donation of two electrons from the ligand to the metal ion (**Figure 1-1a**). As a result, compared to traditional covalent bonds, these non-covalent bonds generally have increased kinetic lability, or capacity to break and reform, at the cost of decreased thermodynamic stability. The lability of each metal-coordinate bond can be quantified in terms of a bond dissociation rate—the frequency at which these bonds are likely to break—which is inversely related to bond dissociation time. When incorporated as a stress-bearing crosslink into a polymer network, these bonds enable unique dynamic mechanical properties because the repeated breaking and reforming of bonds cause the network to continuously flow and remodel. Furthermore, if dynamic metal-coordinate bonds are incorporated alongside permanent load-bearing bonds, they can act as sacrificial bonds by breaking instead of the permanent bonds that hold the polymer network together, thereby increasing the energy required for network failure (**Figure 1-1b**).[21–23] This dissipation of energy through sacrificial bond breaking can act as a molecular mechanism underlying bulk nonlinear mechanical properties such as toughness (**Figure 1-1b)**. Because metal-ligand bonds re-associate over time, these hybrid networks also exhibit self-healing.

In addition to reversibility, metal-coordination bonds have the advantage that their bond strength and dissociation rate can be tuned to a greater extent than those of standard covalent bonds or other transient bonds such as hydrogen bonds by varying simple parameters (**Figure 1-1c**). For example, to change the strength and network dynamics of a traditional covalently crosslinked polymer material, one typically needs to synthesize entirely new polymers with different molecular weights or architectures to change the polymer network density and timescale of any physical entanglements. Transient bonds such as dynamic covalent and hydrogen bonds can be tuned by changing the donor or acceptor chemistry, by applying strong stimuli (for dynamic covalent bonds) or by changing pH (for hydrogen bonds). In contrast, metal-coordination bonds can be more easily tuned by changing the metal ion, metal oxidation state or counterion, which changes the coordination state and bond dynamics.[24]

Further, unlike hydrogen bonds and dynamic covalent bonds, which are confined to their positions by the polymer backbone, metal ions can freely diffuse to find binding partners, but can also be easily removed by standard chelators to 'switch off' coordination complexes. Through these simple levers, the transient nature of the metal-ligand bond enables highly dynamic properties that can be controlled over a wide time scale (**Figure 1-1c,d**). Finally, because metal–ligand bonds have different dissociation time scales, different metal ions can be used to institute a temporal hierarchy of energy dissipation in a material.[25] These properties, combined, suggest that metal-coordination bonds may be an emerging class of crosslink chemistries to tune static and dynamic mechanical properties in advanced materials.

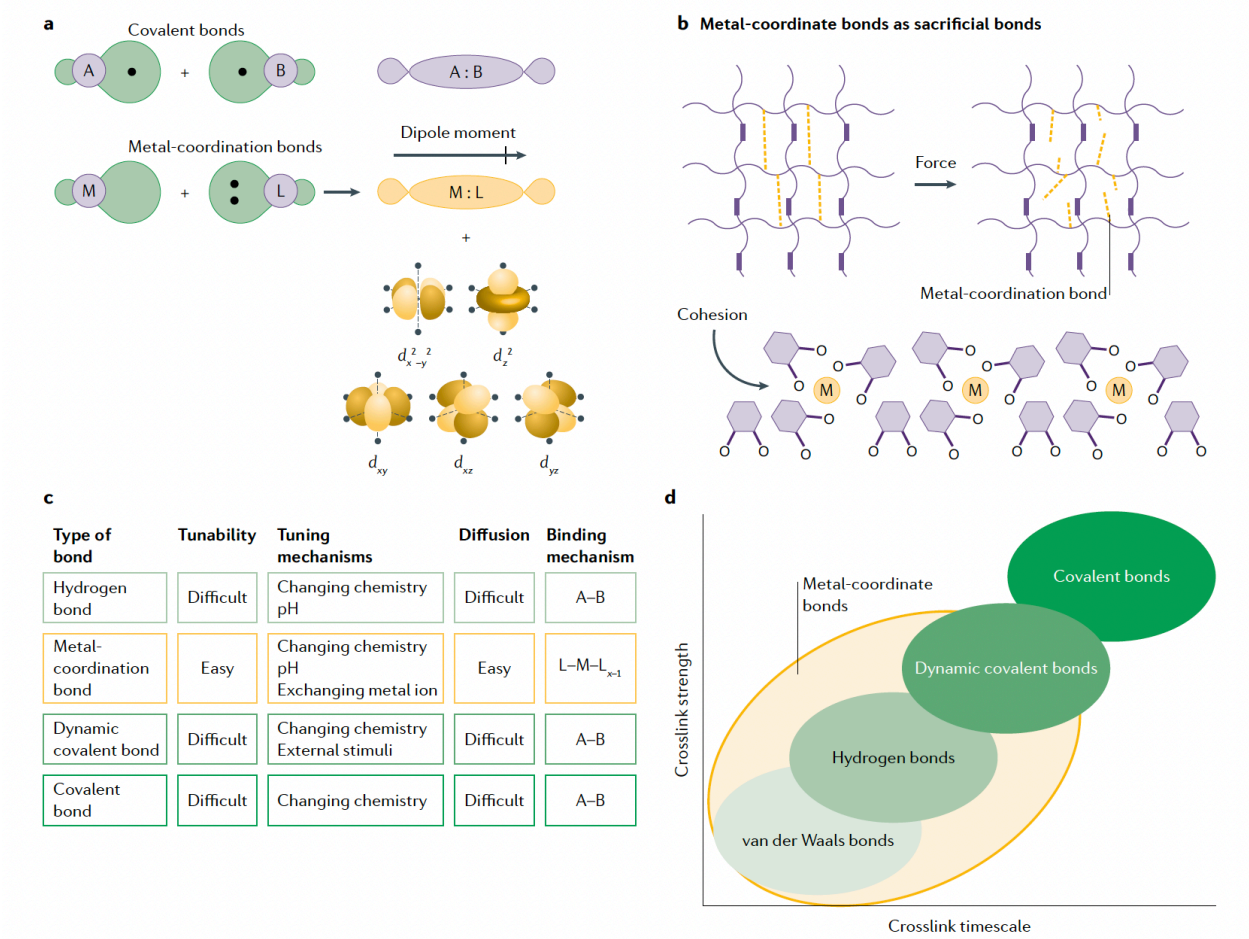


Figure 1-1. Properties enabled by metal-coordinated bonds. a) In covalent bonds, one electron from each hybrid orbital is found in the molecular orbitals. In coordination bonds, both electrons from the ligand (L) are donated to the metal (M) to form a molecular orbital with a dipole moment. In addition, the d orbitals of the metal ion become distorted. b) Metal coordination bonds can form sacrificial crosslinks that break before stronger bonds rupture, thereby increasing fracture energy and preventing catastrophic material failure (top). Metal coordination may also be used in cohesion, where metal-coordinate complexes bind to each other (bottom). c) Comparison of metal-coordination bonds with hydrogen bonds and covalent bonds. d) Metal-coordination bonds span a range of crosslinking strengths and timescales. Figure from Ref.[26]

Although substantial progress has been made in the translation and application of bio-inspired metal coordination for material function,[9,25,27–31] mimicking the complex features of biological metal-coordinated molecular materials such as proteins is nontrivial, because the local dynamics of their biochemical milieu is typically intimately coupled to a higher-order structural

hierarchy.[32–34] Further, although substantial work has focused on the relationship between the timescale of metal-coordinate bonds and linear mechanical material properties such as stress relaxation time, less work has explored how bonds affect non-linear mechanical properties in the large-deformation regime to tune properties such as toughness and the ability to self-heal.[35–37]

A fundamental understanding of the microscopic origins of the macroscopic behavior of materials with metal-coordination bonds, of how such bonds are affected by different local chemical environments, how they are structurally organized, and how they consequently affect nonlinear mechanical properties is still missing. Further, the remarkable nonlinear mechanical properties, including toughness and adhesiveness, of metal-coordinated biological materials have only recently started to be explored in synthetic materials. Harnessing the advantages of metal coordination to build novel, reversible, dynamic materials necessitates the convergence of knowledge on metal coordination, biology and materials science.

In the remainder of the Introduction, we survey metal-coordination bonds with the aim to stimulate cross-disciplinary discussion between biologists, chemists, materials scientists and engineers to understand and exploit metal coordination to impart new mechanical functions to soft materials. To this end, we start by exploring metal-coordination structural design strategies found in biological materials. Then, we investigate how natural binding motifs have been translated into synthetic contexts and the role of polymer and protein chemistry, physics, computation and of the experiment–modeling nexus in that translation.[22,38] More specifically, we focus on tunable parameters enabled by metal-coordinate chemistry and their relationship to material mechanical properties. Last, we briefly review some interesting applications to illustrate how these bonds have been creatively employed.

1.1 Metal coordination in biological systems

Several aquatic organisms have been found to use metal coordination with histidine, catechol and aspartate residues as ligands to produce hard, unmineralized or tough materials (**Box 2**).[17] We highlight materials in two organisms with widely differing hydration states to illustrate how metal-coordination bonds in specific structures can be used to increase toughness in hydrated materials and hardness in hard condensed materials: mussels and *Nereis* worms.

Aquatic mussel threads: tough and adhesive

Marine mussels (in the genus *Mytilus*) have been heavily explored as a biological inspiration for metal-coordinated underwater adhesives and tough materials.[16,39,40] Mussels typically adhere to surfaces such as rocks by secreting byssal threads, which have a tough, energy-dissipative core protected by a cuticle and a terminal adhesive plague that prevents threads from being dislodged by crashing ocean waves (**Figure 1-2a**).[16,41] The byssal thread is divided into two mechanically distinct regions: the corrugated, compliant and highly extensible proximal region[42] and the fibrous, stiff, strong and extensible distal region. [43] The resulting mechanical gradient created by the structural fusion of these regions effectively mitigates the mechanical mismatch between the rock and the soft mussel tissue[44]. Specifically, by matching tensile experiments with theoretical modeling, a 80:20 stiff (distal) to soft (proximal) material ratio was found to effectively dissipate impact energy from wave motion.[39,42,43,45]

BOX 2. Examples of coordinating ligands in biological materials

Histidine and aspartate are naturally occurring amino acids, and L-3,4-dihydroxyphenylalanine (DOPA) is a post-translationally modified amino acid. In addition to metal coordination, these amino acids can engage in multiple types of bonding, including cation–pi interactions, pi–pi stacking, hydrogen bonding and covalent bonding.[58,190,195]

Histidine
Histidine consists of an imidazole functional group in which the electronegative pyridine nitrogen functions as the atom coordinating with the metal ion. The alpha amino group and carboxyl group have also been shown to coordinate in the free ligand.[506] The sensitivity of histidine to deprotonation is highly dependent on the anions present and can result in pi–pi stacking, as shown in suckerin hydrogels.[134]

Catechol
DOPA is synthesized from the naturally occurring amino acid tyrosine, and consists of a catechol functional group with two hydroxyl oxygens as metal-coordinating atoms.[507,508]

Aspartate
Aspartate consists of a carboxylic acid side chain that is deprotonated under physiological conditions.

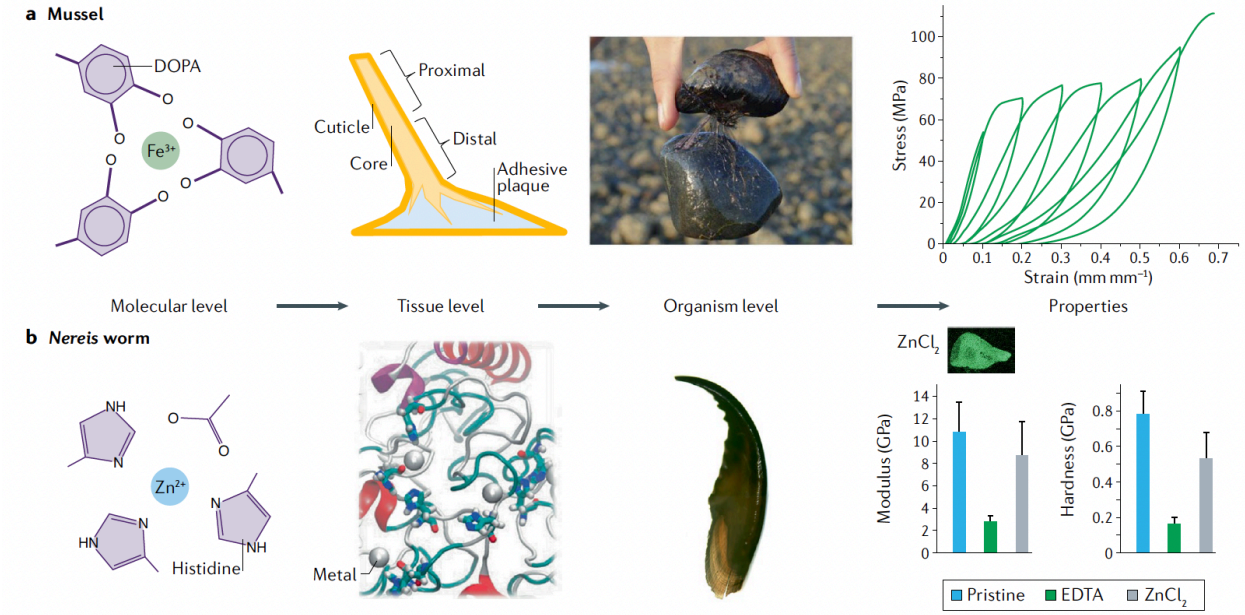


Figure 1-2. Hierarchical organization and resulting mechanical properties of mussels and Nereis virens. a) In marine mussels, at the molecular level, L-3,4-dihydroxyphenylalanine (DOPA) coordinates with Fe3+; at the tissue level, metal coordination occurs in the mussel protective cuticle, byssal thread core and adhesive plaque; at the organism level, mussels secrete byssal threads that attach well to surfaces;[45] and the resulting mechanical properties are illustrated by the characteristic stress–strain curves of a single distal byssal thread. As the thread is stretched to increasing strain values (10–70%), yield is followed by a loss of stiffness in the subsequent cycle.[46] b) In the Nereis worm, at the molecular level, histidines coordinate with Zn2+; at the tissue level, metals organize within

histidine-rich Nereis Nvjp-1 proteins;[47] at the organism level, the combination of this structural organization and histidine–Zn2+ coordination imparts stiffness to the Nereis worm jaw.[48] In terms of properties, the metal coordination increases the stiffness of the worm's jaw: the modulus is higher for the pristine and Zn2+-modified jaw than for the jaw with no Zn2+ (which was removed by chelation with ethylenediaminetetraacetic acid, EDTA).[14] Figure from Ref.[26]

Histidine–Zn^{2+} interactions within the histidine-rich domains of the collagenous proteins in the byssal thread distal region are primarily responsible for the byssal core toughness.[46,49–53] This energy-dissipative role of histidine derives from both its hierarchical organization in the mussel byssal core and its atomistic binding properties, both of which prevent the thread from breaking under large deformations. Histidine is hierarchically organized into tightly folded partially crystalline protein domains that not only contribute to the stiffness of the natural thread but also enable fast, elastic long-range structural recovery and bring into contact histidine–Zn^{2+} sites to facilitate bond reformation.[17,54,55] The distribution of zinc in a gradient further helps to endow specific regions with different degrees of energy dissipation and stiffness, resulting in the mechanical gradient mentioned earlier.[8,56] Atomistically, Zn^{2+} interacts in specific coordination spheres with histidine and aspartate. X-ray absorption spectroscopy and X-ray absorption near-edge structure studies have shown that upon tensile loading under low deformation in the linear regime, zinc coordination crosslinks contribute to initial thread stiffness.[52] Upon increased deformation, the coordination sphere around the metal ion ruptures and dissipates energy, increasing the toughness of the thread.[52] However, the bonds reform quickly once the load is removed and ligand exchange reactions lead to the recovery of a more stable protein configuration over time.[52] Thus, histidine–Zn^{2+} crosslinks and the structural organization of the protein underlie the toughness of the byssus core.

Surrounding the byssus core is the byssus cuticle, which is a hard, yet extensible protective coating. DOPA–Fe^{3+} coordination crosslinks strategically organized in the coating material have been proposed to be responsible for this unique mechanical behavior, as demonstrated through chelation[57] and in-situ resonance Raman spectroscopy experiments.[28] Specifically, the cuticle is organized in a granular composite-like structure, in which densely crosslinked granules could provide hardness and less densely crosslinked matrix might provide extensibility.[28] It was speculated that mussels use DOPA, a post-translationally modified amino acid, rather than standard amino acids, despite the increased energy expenditure required to biosynthesize them, because of DOPA's increased affinity for metal ions in certain environments.[28] The complex role of DOPA in biological and bio-inspired material adhesion has been extensively discussed in several reviews, which we refer the readers to for a more complete understanding.[58–60]

Nereis worm jaw: hard and stiff

Polychaetes, including the *Nereis* and other marine worms, have hard mandibular structures, required for biting or cutting, that in part derive their mechanical properties from metal coordination crosslinks (**Figure 1-2b**).[15,61,62] Unlike the hydrated byssal threads of marine mussels, polychaetes jaws are hard condensed materials and don't contain as much water. Like the marine mussel, *Nereis* utilizes histidine–Zn^{2+} binding in a gradient distribution in the jaw, with material stiffness and hardness correlating with zinc content.[13–15,61,63] To understand how Zn^{2+} contributes to the stiffness of one component of the histidine-rich *Nereis* jaw, the *Nvjp-1*protein, a multiscale computational and experimental work studied the differences between histidine binding with Na^{+} and Zn^{2+}.[47] As expected, Na^{+} binds weakly to histidine because of purely electrostatic interactions and symmetry breaking of histidine's molecular

orbitals, whereas Zn^{2+} binds strongly because of additional electronic interactions with histidine.[47] As a result, *Nvjp-1* proteins coordinating with Zn^{2+} become more compact and harden.[47] In addition to the contribution of metal-coordinate crosslinks to the stiffness of the protein networks, the close proximity of glycine residues allows histidine to coordinate with Zn^{2+} without steric penalties imposed by the structural rearrangements required for Zn^{2+} binding.[47] Together, the structural organization of the protein and binding of histidine–Zn^{2+} impart stiffness to the *Nereis* worm jaw.

Less explored organisms

Several other biological organisms are known to use metal coordination to build material structures, including spiders (for their fangs)[64] and ants and scorpions (for their mandibles),[65] as well as less studied organisms such as terrestrial slugs (for their mucus-based adhesive secretion).[66–68] Further, in addition to the select few mentioned above, several other amino acids can coordinate with metal ions *in vivo,* including single hydroxyl-functionalized (serine, threonine and tyrosine), carboxylate-functionalized (glutamic acid) and thiol-functionalized amino acids (cysteine and methionine). The structural roles of these and other post-translationally modified amino acids, including phosphoserine, should be further explored.[48,69,70]

The mussel byssal thread and *Nereis* worm jaw demonstrate that protein structure and metal-coordinate crosslinks can cooperate to impart remarkable material mechanical properties. Advanced spectroscopic characterization of more metal-coordinated proteins may also help elucidate other coordination complex chemistries, rupture mechanisms, and structural organizations of crosslinks that contribute to mechanical function. Such biological questions are addressed in Ref.[17]

1.2 Synthetic Translation

Inspired by studies of metal-coordination complexes in biological materials as described above, researchers have implemented coordination complexes as loadbearing crosslinks in polymer and protein biomimetic networks with resulting mechanical properties that can be controlled via the choice of metal ion, coordinating ligands and physical-chemical processing conditions. Through such studies it has become increasingly clear that the physical-chemical environment in which the metal-coordination bond sits (protein or polymer backbone, pH, metal-ligand stoichiometry) strongly influences its local bond dynamics and thereby the bulk material mechanics. In this section we focus on hydrogels, not only because they are simple model systems to study metal-coordinate bond interactions within, but also because tuning the mechanical properties of hydrogels can enable medical applications, such as controlled drug release. We start by surveying metal coordination in polymer hydrogel networks, then discuss how it has been used in protein-based hydrogel materials, which are structurally more complex than polymer systems.

Gelation of coordination polymers

Early experiments with simple copeptides showed that polymers conjugated with metal-coordinating amino acids can assemble into dynamic materials.[71,72] Following on these initial demonstrations, seminal work showed that incorporating specific catechol or histidine coordination ligands into polymer networks can result in hydrogels with viscoelastic properties

that can be easily tuned with simple levers such as pH (**Figure 1-3a)**.[910] Despite their simplicity, these synthetic polymer mimics of metal-coordinated biological materials already offer a range of potential applications due to their relative ease of synthesis, assembly and tunability (**Figure 1-4**).

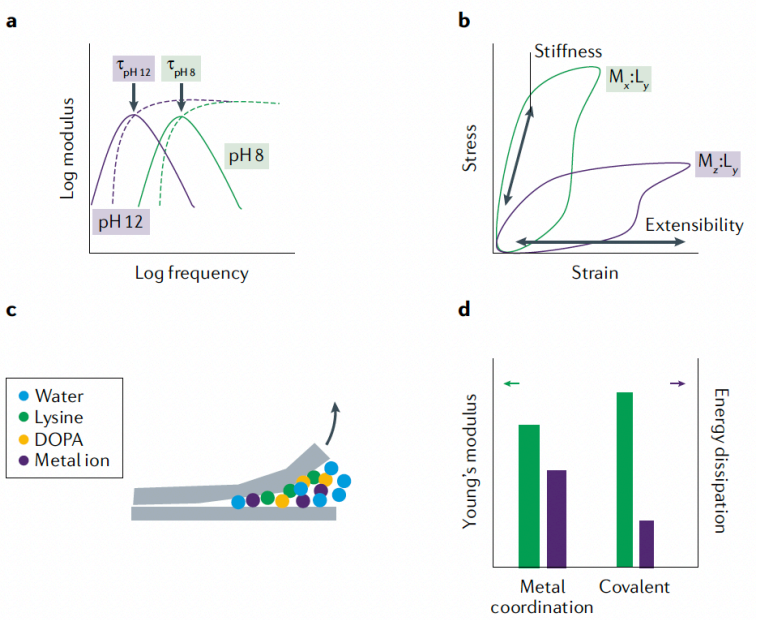


Figure 1-3. Mechanical signatures of metal-coordination bonds. a) Rheological frequency sweep of metal-coordination crosslinked gels demonstrates dynamic properties with relaxation times, τ, that change depending on the pH conditions. The relaxation time is the crossover between the two curves, which represent the storage (solid lines) and loss modulus (dashed lines) of the network. b) Different metal-to-ligand stoichiometries result in different extensibility properties of the network. The graph shows the case of imidazole-Co^{2+} melts; Mz is bigger than Mx. c) Lysine, compared with other amino acids, enhances the adhesive strength of L-3,4-dihydroxyphenylalanine (DOPA) in synthetic polymer networks. d) Metal-coordination gels can achieve a higher Young's modulus and energy dissipation than covalently crosslinked gels. Figure from Ref.[26]

Coordinating ligands can be attached as terminal or pendant ligands on polymer backbones, depending on the intended polymer architecture. Telechelic polymers (polymers end-functionalized with coordinating ligands) have been utilized in efforts to assemble ideal networks to isolate the relationship between crosslink dynamics and network mechanics in gels. An ideal network is constructed from polymers with well-defined polymer chain lengths below the entanglement molecular weight, so that the bulk relaxation time of the network is dominated by the kinetic behavior (the bond dissociation time) of the metal-ion crosslinks.[73] By contrast, polymers with pendant coordinating ligands enable multiple ligands on the backbone to coordinate with the same metal ions, increase network stability in water, and prolong the relaxation time of the network by increasing the number of coordination sites along the backbone.[27,37,74] To synthesize polymers functionalized with metal-coordinating ligands, several traditional polymerization methods have been used, including ring-opening polymerization,[74–76] grafting[77] and reversible addition-fragmentation transfer.[78–80]

When crosslinking and forming gels or metal-coordinated networks, divalent and trivalent transition metal ions and specific pH conditions are needed to ensure that the ligand is in a correct protonation state and that enough coordinate crosslinks are formed to percolate and form a polymer network. Increasing pH deprotonates coordinating ligands, thereby enabling

metal coordination, and changes the type of coordination complexes present in the system.[9,31] For example, increasing pH in Fe^{3+}–catechol systems results in one, two or three bidentate ligands coordinated to one metal, forming a 'mono', 'bis', or 'tris' complex respectively.[9] For a simple bifunctional linear polymer with one catechol coordinating ligand at each end, tris complexes are required for forming a fully percolated network. By contrast, for a multiarmed polymer backbone both bis and tris complexes contribute to network formation. The amount of metal added, or metal-to-ligand stoichiometry, can also change the distribution of mono, bis or tris complexes.[31] The influence of the coordination number, or number of molecules bound to a metal ion, on macroscopic dynamics is discussed in the next section.

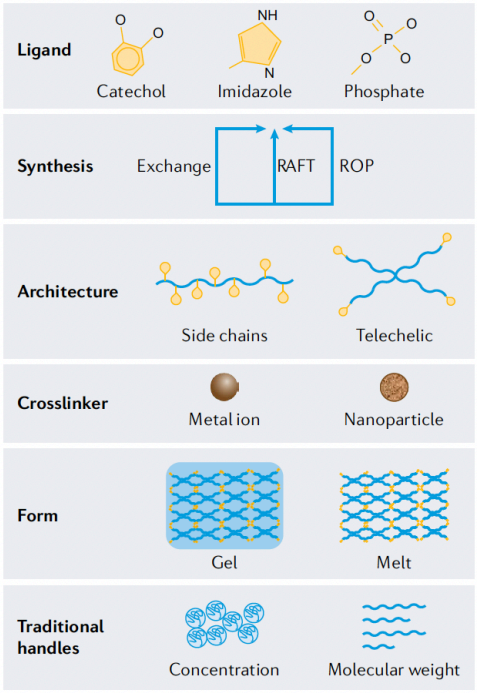


Figure 1-4. Engineering metal-coordinated polymers. Control levers in the synthesis of metal-coordinated polymers include coordinating ligands, synthesis polymerization methods (such as reversible addition–fragmentation chain-transfer polymerization (RAFT) or ring opening polymerization (ROP)), chain architecture, type of crosslinker and form of network, which can be a gel (with solvent) or a melt (without solvent). These levers can also be combined with traditional polymer engineering handles, including changing the concentration or molecular weight of the polymers. Figure from Ref.[26]

Understanding how traditional levers of polymer engineering[81] and nanostructural topology impact metal-coordination dynamics is critical in the development of more complex polymeric materials.[82,83] Traditional approaches to tune polymer network dynamics include entanglement length, volume fraction, molecular weight and crosslink density. For example, increasing the polymer volume fraction increases the likelihood that a metal ion finds binding partners to form a percolated network; increasing crosslink density increases stiffness; and increasing the entanglement reptation time increase the bulk material relaxation time. These traditional polymer engineering approaches can also directly affect metal–ligand crosslinks by inducing swelling forces.[73,84,85]

Protein-based synthetic materials

Metal-coordination bonds can also be incorporated into protein-based synthetic materials, which have more complex secondary and tertiary structures compared to random coil polymers. As we discuss here, this additional structural complexity results in gel properties different than those of polymers, such as different rheological behavior and formation of hierarchical structures.

Although most natural proteins are difficult to extract or express in large quantities, advances in protein expression have enabled the fabrication of protein hydrogels with recombinant mussel proteins and histidine-rich worm jaw proteins.[86–90] Optimized bacterial vectors such as *E. coli* allow the expression of proteins with tyrosine or histidine.[86,89,91] To synthesize post-translationally modified amino acids, enzymes such as tyrosinase are used to convert tyrosine into DOPA.[86,89] Solid-phase peptide synthesis has also emerged as a viable technique in the synthesis of designer metal-coordinating peptides.[92,93] It is still challenging to incorporate nonnative ligands, such as multidentate terpyridine, into proteins. Despite these challenges, these peptide synthesis methods offer ways to build specific protein sequences and structures.

When integrated in polymer and protein gels metal-coordination bonds often result in different network mechanical properties because protein structure imparts additional mechanical effects on networks. This was demonstrated in protein hydrogels synthesized from consensus decapeptides from mussel foot proteins and reversibly crosslinked with Fe^{3+}.[90] The protein hydrogel has a relaxation time of the same order of magnitude as that of a simple polymer–catechol–Fe^{3+} hydrogel (~10^0 s),[9] which indicates the critical role of metal–ligand bonds in determining relaxation timescales. However, the gels exhibit more viscous rheological behavior in the low-frequency regime where gels flow, because the stiff helical nature of the peptide is more complex than the random coil flexible polymer chains, and because additional mechanisms, including cation–π and π–π interactions, enhance the charge-transfer interaction between metal and ligand.[90]

Because the protein tertiary structure geometrically constrains the metal–ligand interactions, metal ion identity (Zn^{2+}, Cu^{2+}, Ni^{2+}) influence mechanical reinforcement differently in peptide-based networks than in polymer-only systems.[25,94,95] For example, Zn^{2+} induces a 10-fold stiffness increase in histidine-rich β-amyloid films, but Cu^{2+} has little effect,[95] even though Zn^{2+} and Cu^{2+} have similar relaxation times in histidine–polyethylene glycol networks.[25]

Protein-based metal-coordinated materials can also be used for creating hierarchical structures, because although metal ions are not required for protein structure assembly,[94–97] they can influence their organization.[98–101] For example, selectively incorporating carefully spaced intermolecular histidine–Zn^{2+} crosslinks induced the formation of extended assemblies of alpha-helical coiled-coil proteins.[102]

Different ways of influencing dynamic mechanical behavior, such as metal coordination and peptide structure, can also be combined to yield materials with two distinct stimuli-responsive relaxation modes.[94,95,103] Coiled-coil protein hydrogels naturally behave as yield-stress fluids due to the intrinsic dissociation time of coiled-coil network crosslinks. However, sequencing specific histidine modifications into well-defined coiled-coil proteins induces intermolecular histidine–Zn^{2+} crosslinks that affect the gel relaxation time by over three orders of magnitude.[102]

If the Zn^{2+} ions are removed via chelation, the hydrogel recovers the characteristic dissociation time of the coiled-coil crosslinks.[102] If the bonds are incorporated intramolecularly instead of intermolecularly, the bonds thermodynamically stabilize the coiled-coil protein by increasing the barrier height for dissociation of the coiled-coil domain, which directly increases the network's relaxation time.[104] Such modifications are particularly interesting because mutating specific residues on proteins to coordinate with metal ions could help control the dynamic relaxation properties of protein assemblies.[105]

The above examples illustrate that metal-coordination bonds can be readily incorporated in polymer or protein hydrogels and thereby offer tunable dynamic mechanical properties. Substantial research in this field has focused on the development of coordinating polymers due to their comparative ease of synthesis and the possibility to isolate the mechanical contributions of coordination bonds within the resulting networks. However, incorporating these bonds into protein hydrogel materials offers more hierarchical and complex mechanical effects that necessitate further exploration. Specifically, studies that attempt to decouple the synergistic roles of protein structure and metal ions in the determination of mechanical properties and assembly are still nascent.[94,95,97] Machine learning may be a suitable way to garner additional insights and design de novo proteins to understand the role of metal-coordinated bonds within a protein sequence.[106,107]

1.3 Understanding Design Parameters

We next examine how the kinetic properties of metal-coordination bonds depend on local physical-chemical aspects such as electronic bonding interactions, coordination geometry, ligand exchange pathways, coordination partners, chemical redox potential and water coordination (**Figure 1-5**). Several of these kinetic bond properties are not observed in hydrogen or dynamic covalent bonds and we discuss how this affects the bulk material's dynamic mechanical properties in unique ways. We also discuss how new ligand chemistries have expanded the range of mechanical properties accessible via metal-coordination crosslinking in polymeric materials. Direct connections between the thermodynamic properties of metal-coordination bonds and bulk material mechanical properties such as strength and Young's modulus are still poorly understood, thus we only touch upon them briefly.

Physical-chemical influence of coordination complexes on mechanical properties

Specific electronic interactions between metal ions and ligands dictate the mechanistic connection between microscopic metal-coordinate bond dynamics (bond dissociation time) and macroscopic material mechanics (such as material relaxation time, **Figure 1-5a**). Specifically, an accumulating body of evidence suggests that the activation energy of bond dissociation dictates the bulk material relaxation time as characterized through an empirical Arrhenius relationship, where higher activation energy scales with slower relaxation (see **Box 3** for more details on the fundamental coupling between bond kinetics and material crosslink dynamics). Various classical inorganic chemistry series may start to explain the trends in mechanical properties. To explain why Ni^{2+} ($3d^8$) relaxes more slowly than Cu^{2+} ($3d^9$) or Zn^{2+} ($3d^{10}$) in a histidine-based network,[25] inorganic chemistry concepts including histidine affinity for metal ions ($Zn^{2+} < Ni^{2+} < Cu^{2+}$),[108] the Irving–Williams complex stability series ($Ni^{2+} < Cu^{2+} > Zn^{2+}$), which is related to ionic radius and second ionization potential,[109] the Gibbs monohydrate interaction energies calculated through gas-phase density functional theory ($Ni^{2+} < Cu^{2+} <$

Zn^{2+}),[110] the increase in acidity of histidine in presence of metal ions ($Ni^{2+} < Cu^{2+} < Zn^{2+}$)[110] and equilibrium binding constants have been used.[111] However, in our opinion, these series do not adequately explain different relaxation times, and further quantum chemical studies that consider the role of water as an active coordination partner and complex bond dissociation pathways are needed.[112] Though density functional theory has emerged as a useful technique,[113–117] modeling the local coordination environment of metal ions and the dynamic nature of metal–water exchange in aqueous environments requires the application of quantum mechanical principles in a manner less computationally intensive than DFT. To this end, the reactive force field (ReaxFF), in which empirical interatomic potentials depend on a bond order formalism, enables the modelling of systems in which bonds break and form[118–120] and has already reproduced Cu–water coordination properties[121] and metal-organic framework stability.[122] Hence, this force field could potentially help develop a deeper understanding of the specific role of bond electronic properties in determining metal-coordination complex dynamics to assist their application in the future engineering of polymer material mechanics.

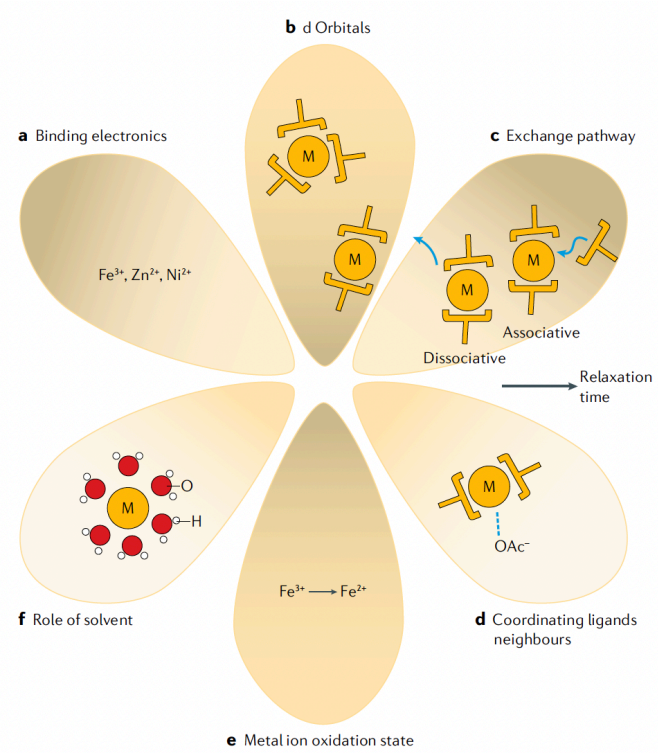


Figure 1-5. Chemical factors influencing the relaxation time of the network. The levers are binding electronics, as different numbers of electrons in metals produce different bond lifetimes (panel a); d orbitals of metals (M), which result in different preferred coordination geometries of binding (panel b); metal preference for associative or dissociative exchange pathways, which determines whether surrounding ligands influence kinetics (panel c); coordination with counterions and neighboring ligands, which influences the bond (panel d); multivalency, which allows the metal ions to be reduced to form different bond types (panel e); and the hygroscopic water-binding nature of metal ions, which is critical for gelation conditions and, potentially, for cohesive properties (panel f). Figure from Ref.[26]

Electronic properties also dictate preferred metal-coordination complex geometries and coordination numbers, which affect bulk network relaxation time and strength (**Figure 1-5b**).[123–125] To minimize repulsion between electrons in the valence shell, Zn^{2+} generally prefers tetrahedral, Cu^{2+} tetrahedral or square planar, and Ni^{2+} octahedral ion arrangements.[126] These coordination geometries may also explain why protein-based hydrogels exhibit different trends

in metal-dependent relaxation times compared to polymer networks, because proteins impose a coordination geometry that may not be preferred depending on the metal ion. In aqueous solutions and hydrogels, metal ions satisfy these preferred coordination geometries by binding to different numbers of ligands, for example to one, two or three bidentate ligands to form mono, bis or tris complexes. These complexes exhibit different bond dissociation times.[9,127] For example, rheological experiments show that bis Fe^{3+}–catechol systems have a relaxation time of 10^{-2}s and tris systems of 10^{0} s.[9] This trend has been observed for other metals with catechol systems; V^{3+}–catechol exhibits a larger degree of tris-complexation and thus an even longer relaxation time than Fe^{3+} gels at the same pH.[24]

BOX 3. Concepts from transient network physics in the design of coordination materials

Throughout this Introduction, we discuss the importance of dynamic mechanical properties enabled by metal coordination. Whereas thermodynamic properties (such as the equilibrium constant K) determine whether network formation is possible,[31] bond kinetics (such as the rate constant k) and frequency of crosslinking determine the macroscopic dynamic properties and relaxation modes once the network has formed.[217,509] Seminal work with polymers crosslinked via metal coordination to pincer ligands revealed that ligands with the same thermodynamic equilibrium constant and different kinetic constants show substantially different dynamic properties measured through viscosity.[217] For example, bulkier ligands slow ligand-exchange dynamics and increase polymer solution viscosity.[217]

To understand these kinetic constants, studies of transient bonds in biological systems have helped provide quantitative models for how bonds dissociate. For instance, the Bell model, based on ligand receptors in biological molecules, describes how the lifetime of a transient bond and rate of bond formation are dependent on k–, the dissociation rate constant of the bond.[335] Although metal-coordination bonds also dissociate due to thermal fluctuations, the Evans model proposed that the application of force decreases the activation energy barrier required to break the bond.[128,510,511] Tuning the activation barrier of the dissociation of metal-ligand complexes can substantially affect the frequency of dissociation of metal complexes and thus the relaxation time of an ideal bulk network.[104] Though the Bell–Evans model has been successfully applied to transiently bonded systems such as hydrogen-bonded alpha helices and beta sheets,[326,511,512] these studies have focused on ligand placement along a backbone rather than coordination geometry or multiple coordination partners in metal binding.[103,112,128,513] A greater understanding of quantitative models that describe specific kinetic properties of coordination complexes will facilitate the design of new molecules for metal coordination.

The dependence of mechanical properties on the coordination number has also been observed through single-molecule force spectroscopy. Atomic force spectroscopy experiments coupled with first-principles quantum chemical studies demonstrated that different complexes have different rupture forces, attributed to their mechanical rupture pathways. For example, bis Fe^{3+}–catechol complexes require higher forces to rupture than the corresponding tris complexes, perhaps surprisingly given that gels with primarily bis Fe^{3+}-catechol relax faster than gels with tris Fe^{3+}-catechol.[128] In addition, compared to a covalent crosslink, the coordination bond features more peaks in its force–distance profile as it is pulled, because of more degrees of freedom involved in bond deformation, including bond stretching and bond angle bending.[129]

Further experimental and computational studies on how coordination geometries and numbers affect bond dissociation rate and strength in aqueous environments are needed.

Coordination complexes break and reform through associative or dissociative pathways,[130–132] which dictate whether other dangling ligands, tethered to a network but not bound to a metal complex, can affect bond dynamics (**Figure 1-5c)**. For metals with associative character, to which dangling ligands associate before the next ligand dissociates, the presence of additional dangling ligands increases the rate of ligand exchange and therefore the kinetics of crosslinker remodeling.[27] Metal complexes with dissociative character are unaffected by dangling ligands, because steric hindrance prevents additional ligands from complexing.[133] By contrast, other transient bonds such as hydrogen bonds and ion pairs generally exhibit dissociative ligand exchange reactions. This may be because metal centers allow more positions for ligands to coordinate (for example six sites in an octahedral arrangement that are not always filled with ligands), whereas hydrogen bonds normally involve one donor and one acceptor atom. In imidazole-functionalized polymer melts, these differences between ligand exchange mechanisms explain the observed rheological behavior. Co^{2+} engages in a dissociative mechanism in which the ligand exchange rate is independent from dangling ligand concentration. Therefore, increasing the concentration of dangling ligands by decreasing the metal-to-ligand (M:L) ratio results in a steady decrease in viscosity as expected because of the decrease in the number of crosslinks in the network. By contrast, Zn^{2+} and Cu^{2+} engage in associative exchange mechanisms in which the ligand exchange rate is directly proportional to the concentration of dangling ligands. Increasing the concentration of dangling ligands by decreasing M:L in Zn^{2+} or Cu^{2+} networks results in a surprising constant, low viscosity because dangling ligands rapidly exchange and remodel the network.[27] These exchange pathways also affect bulk tensile mechanical properties. With increasing M:L ratios, extensibility in Co^{2+} melts was reduced due to increased crosslinking density but Zn^{2+} and Cu^{2+} melts maintained their extensibility owing to frequent remodeling due to dangling ligands (**Figure 1-3b)**.[27] Through an understanding of the associative or dissociative mechanism of the metal-ligand coordination site, mechanical properties can thus be tuned by changing the number of dangling ligands present.

In addition to dangling ligands, counterions and other ligands that do not directly coordinate to the metal can also affect the stability and kinetics of metal-coordination complexes (**Figure 1-5d**). This has been demonstrated in several biological materials in which counterion impacts material properties.[15,134] Progressive exposure of synthetic *Nereis* worm jaw protein hydrogels to Zn^{2+} and different counterions (such as Cl if $ZnCl_2$ salts are used) sclerotized and increased the gel's elastic modulus in a highly counterion-influenced manner.[91] In particular, acetate was suggested to facilitate the local deprotonation of histidine residues to enable stronger coordination with Zn^{2+}. Further, in contrast to chloride ions, which facilitate in coordination bonds confined to a local area, acetate ions enabled the diffusion and crosslinking of Zn^{2+} through the entire material.[91] The Zn-acetate-containing hydrogel also had a lower equilibrium water content than gels with other counterions.[91] This same decrease in water content has been observed in other histidine– Zn^{2+} based hydrogels.[102] Neighboring residues can have a similar effect; lysine synergistically improves DOPA's wet adhesion by displacing cations to enable catechol–metal binding (**Figure 1-3c**), though it may not directly act as a metal-coordination partner.[135–137] The influence of counterions and neighboring residues emphasizes that the local chemical environment of the bonds is critical to their mechanical behavior.

The oxidation state of transition metal ions is yet another source of metal-coordinate bond state changes. For example, metal ion reduction can enable coordinating ligands to oxidize and crosslink via permanent covalent bonds, significantly affecting the material's mechanical properties (**Figure 1-5e**). In Fe–catechol systems, a catechol–catechol covalent bond is formed when Fe is reduced and catechol is oxidized to a quinone under acidic conditions, resulting in a permanently crosslinked gel.[138,139] Electrochemical potentiometric titration can also transition reversible crosslinks from strongly binding Fe^{3+} to nonbinding Fe^{2+} in acrylate-sulfonate hydrogels.[140] By using a range of metal ions with different valency and redox potentials, electrochemical cells can control the ratios of coordination: covalent bonds, thereby varying the balance of transient to permanent network crosslinks.[141] Oxidizing agents such as hydrogen peroxide can also be used to manipulate the oxidation state of metal ions.[142] In a polymer–histidine–cobalt ion network, Co^{2+} oxidizes to Co^{3+} upon the addition of hydrogen peroxide. As a result, the network transitions from being reversible (with Co^{2+}) to permanently crosslinked and swelling resistant (with Co^{3+}).[142] We note that the biochemical literature has several examples of the mechanical role of different metal ion oxidation states.[143,144] We encourage readers to refer to this literature to understand further strategies for chemically tuning the metal ion oxidation state to modify the bond type and kinetics.

A significantly less understood topic in metal-coordination polymer material design is the role of water or other solvent molecules on network mechanical properties (**Figure 1-5f**).[145] However, evidence of the profound effects of water on metal-coordinated materials is abundant. For example, metal has a more pronounced effect on the hydrated *Nereis* worm jaw state than on the dry state: the modulus and hardness of hydrated *Nereis* proteins increase about three times upon the addition of Zn^{2+}, compared to about 1.5 times in dry samples.[14] An increasing number of studies have started elucidating why the role of solvent is critical in metal-coordinate crosslinked polymer systems, apart from the general effect of swelling on network density. In fully swollen hydrogel networks, water can both coordinate with the metal ion or protonate the ligand[146] and thus affect the thermodynamics of the coordination complex.[131] Modeling the equilibrium constants of different binding events including water self-ionization, ligand protonation, hydroxide ion competition and metal–ligand reactions in solvated metal-coordinated networks revealed that excess amounts of metal can still enable robust hydrogel gelation, because competing hydroxide ions in water can buffer excess metal ions and thereby protect the gel network from dissolution.[31] The insights from this model greatly expand the possible metal–ligand stoichiometries and pH conditions required for gelation and can be tailored to different hydrogel chemistries by using the equilibrium binding constants of relevant ligands.[31,111] Furthermore, even in dehydrated hydrogels, small amounts of water remain locally bound to metal-coordination complexes.[147] Despite increased network density and thereby stiffness upon dehydration, these tightly bound water molecules appear to enable the metal-coordinate complex to remain dynamic, which significantly increases energy dissipation in these dehydrated networks (**Figure 1-3d**), compared to dehydrated networks with permanent crosslinks.[147] These studies highlight the important role water plays in metal coordination.

A more detailed molecular picture of how microscopic bond properties impact macroscopic mechanical properties will undoubtedly help in developing predictive models relating coordination bond chemistry to mechanics. One approach to achieve this goal is using multiscale simulation methods, which thus far have not been used extensively in metal-coordinated networks but may reveal important mechanistic insights into their dynamic properties. For example, density functional theory can help determine transition state

distortions when pulling the metal-coordinate bond[129,148] or provide an electronic basis for activation energies; molecular dynamics can help solve for the structures of various metal-coordinated proteins;[47] coarse-graining can be used to translate microscopic associative interactions into larger network behavior[149–151]; and continuum mechanics may eventually be used to predict overall mechanical properties (**Figure 1-6**). These simulations may explain interchain and intrachain bond dissociation under different mechanical loading conditions[152] or network percolation and cluster formation.[153]

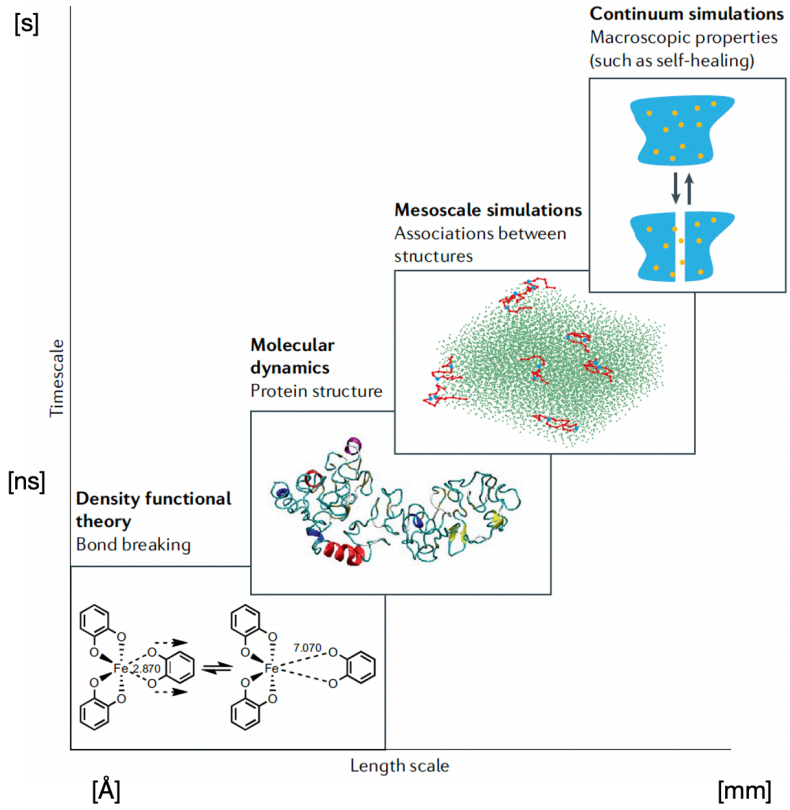


Figure 1-6. Multiscale modelling of metal-coordinated materials. Multiscale modelling has been underutilized in this field but offers significant advantages for materials design, complementing experimental analysis and helping to understand interactions in various classes of bonding in materials. Metal-coordinate bonds can be used to engineer a new degree of freedom by exploiting their dependence on time. Density functional theory can be used to capture the quantum physics of metal-coordinate bond breaking at the atomic level.[148] Replica exchange molecular dynamics can be used to model the structures of metal-coordinated proteins.[47] Coarse-grained non-equilibrium molecular dynamics simulations unveil the configurations of multisticker associative polymer solutions.[149] Continuum methods that map local transient behaviors of metal-coordinate bonds to bulk macroscopic properties such as self-healing are yet to be developed. Figure from Ref.[26]

Less explored chemistries

The understanding of chemical interactions between metal and ligand has enabled an expanded tunability of metal–ligand thermodynamics and kinetics by employing a growing number of variants of naturally occurring amino acids and artificially synthesized ligands (**Figure 1-7**).[14,154,155] As an example, a 3-hydroxy-4-pyridinonone (HOPO) ligand (**Figure 1-7a**) can form stable gels at physiological instead of alkaline pH, bind to Al^{3+}, Ga^{3+} and Cu^{2+}, and resist oxidative degradation to a quinone species because the electron-withdrawing ketone and hydroxyl functional groups allow the phenol to strongly bind the metal ion.[156] As a result, in

contrast to a gel assembled with catechol ligands, the resulting gel does not evolve from a transiently to a covalently crosslinked gel with time.

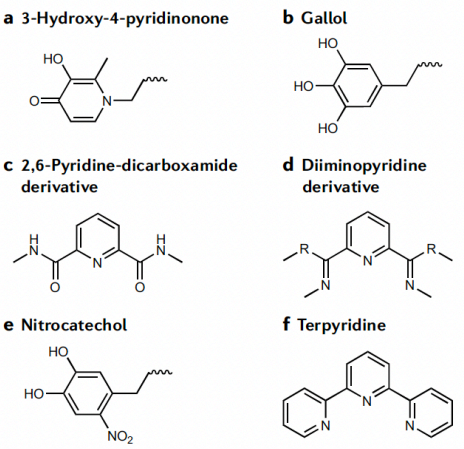


Figure 1-7. Examples of ligands used for metal coordination with different metal ions. a) 3-Hydroxy-4-pyridinonone (HOPO) is a catechol analogue used to resist oxidation degradation and maintain the transient nature of the HOPO–metal bond.[156] b) Gallol is a catechol analogue with three coordinating alcohol groups for coordination instead of two.[157] c) Carboxamido ligands coordinate weakly with and pyridyl group coordinates strongly with Fe3+.[11] d) Diiminopyridine derivatives can coordinate well with Zn^{2+}.[158] e) Nitrocatechol is a catechol analogue used to resist oxidation degradation and maintain the transient nature of the nitrocatechol–metal bond.[156] f) Terpyridine is a tridentate ligand that provides strong coordination with several transition metal ions.[159] Figure from Ref.[26]

Other biologically inspired ligands are currently under investigation in the context of structural materials design.[160,161] Gallol, which similarly to catechol coordinates with Fe^{3+} but contains three hydroxyl groups (**Figure 1-7b**), is naturally found in the tough, flexible bodies of marine invertebrates called tunicates. Polymers functionalized with gallol have been synthesized as adhesive hydrogels[157,162] Combining different types of coordinating ligands can also produce interesting properties. For example, an elastomer with two weak carboxamido–iron ligands reversibly breaks when a neighboring strong pyridyl-iron ligand localizes the metal ion for faster reversible unfolding and refolding of chains (**Figure 1-7c**).[11] This cooperativity of coordination properties enables highly stretchable and self-healing elastomeric materials.[11] These novel chemistries open up previously unattainable tuning of bond strength and dynamics.[158,163]

In addition to new ligands, different transition metal ions or nanoparticles can also be used to tune mechanical properties.[24,164] For example, other transition metal ions such as Cu^{2+}, Mn^{2+} and Cd^{2+}, which do not coordinate in the native *Nereis* worm jaw, can induce hardening of the jaw and jaw protein.[14,91] Like metal ions, nanoparticles can also be used as crosslinkers in metal-coordination systems. Inserting Fe_3O_4 nanoparticles into catechol gels results in strikingly slower relaxation times than with single metal ion crosslinkers, but still produces a fully transient, reversible gel.[165] Stress relaxation can only occur through the dissociation of all the polymer chains attached to the nanoparticle surface, resulting in a slow process.[165] Nanoparticles also offer an avenue for thermal and magnetic actuation.[166] Additional research is required to elucidate how these less explored metal-ion, nanoparticle and ligand chemistries function to enable their potential applications.

Overall, metal-coordination complexes offer exceptional control on the dynamic mechanical properties of polymer materials, such as relaxation time and stiffness, through a multitude of physical-chemical levers. Increasing experimental and computational efforts are constructing

multiscale design principles to connect these organic-inorganic chemical dynamics to the resulting material mechanics. These efforts have already paid off for example by using this new understanding to develop new ligand chemistries that have been designed to influence specific mechanical properties.[11,163] Significantly more effort is needed to both solidify the relationship between metal-coordinate bond chemistry and mechanics and extend it to nonlinear properties such as toughness, which would enable a wealth of new applications.

1.4 Applications in mechanical materials

In this section, we survey how metal-coordination complexes have been applied to build tough and tunable materials. The results discussed here serve as good examples for how to use these complexes in combination with various materials design principles to tune mechanical properties.

Tough self-healing materials for increased material utility

Metal-coordination complexes have been utilized to build tough materials.[123,167–170] Fe3+ can be used to crosslink catechol-functionalized polyethylene glycol chains in a loosely crosslinked amorphous epoxy network to obtain a tough and stiff yet extensible material.[169] The catechol groups organize into nanoscale ionomeric domains, which enhance stiffness by restricting polymer chain movement and distribute stress away from the crack tip.[169,171] Upon the addition of Fe^{3+}, network toughness, stiffness and strength are amplified beyond what is expected from the existence of ionomeric domains alone because metal-coordination bonds contribute additional crosslinking points, but also break and reform to dissipate energy.[169]

The crosslinking and self-healing behavior of metal-coordination bonds can also be used to assemble materials that do not easily fracture. Metal-coordination complexes can crosslink silk fibroin hydrogels under physiological conditions so that they can flow to fill irregularly shaped tissue defects and gel without fragmenting.[172] This moldable yet mechanically robust gel enables stem cell proliferation and bone regeneration for tissue engineering applications.[172] The same principle can be applied *ex vivo*, where light-excitable metal-ligand complexes that release heat enable the flow of a polymeric material to fill cracks.[173]

The properties enabled by metal coordination are being exploited in tough, self-healing electronic materials.[174] Metal-coordination sites incorporated into a dielectric elastomer polydimethylsiloxane polymer resulted in a highly stretchable, tough, self-healing elastomer that could be actuated by an electric field.[11] Hysteresis could be controlled by changing the counterions coordinating with the cation.[175] For good electrical and self-healing properties, strong interaction with the counterion and lability of the coordination geometry are required.[11,175]

Tunable dynamic actuators

The ease of tunability of metal-ligand bonds has led to the engineering of materials with structural or dynamic property changes that can be triggered using simple levers like temperature and pH.[47,176] For example, Fe3+ ionoprinted onto specific regions of a catechol-functionalized methacrylamide hydrogel can be actuated using pH.[177] Tris complexes induce bending deformation and therefore cause a high stress differential despite low catechol

concentrations,[177] resulting in fast actuation.[178] This same tunability enables plasticity and shape change in metal-coordinated shape-memory polymers over a wide range of temperatures.[179]

In biomedical applications, metal ions can tune drug-bearing nanoparticles or gel viscoelastic properties to produce specific drug release kinetics.[180–182] For example, bioengineered mussel adhesive protein nanoparticles crosslinked with Fe^{3+} were filled with doxorubicin and shown to be effective for locoregional cancer therapy. In this system, dissolution of the metal-coordinated networks in the low pH of the cellular environment releases the drug.[183,184] This same principle can be applied to dissociate Cu^{2+} crosslinks under low pH in block copolymer micelles.[185] Applying ultrasound outside of the body can also disintegrate hydrogel networks lightly crosslinked with Fe^{3+}–DOPA.[186] These tunable dynamic properties are also attractive for growing cells on hydrogels, because temporal mechanical cues from the environment dictate cell response.[187]

Adhesives

A discussion of the structural functions of metal coordination is remiss without an acknowledgment of the large body of research focused on understanding mechanochemical processes in adhesion and designing new metal-coordination adhesive materials.[58,59,160,188–196] As mentioned earlier, Fe^{3+}–DOPA metal coordination enables the strong adhesive behavior of the mussel byssal thread adhesive plaque.[197] The plaque adhesive behavior can also be actuated by changes in pH, which change the adhesion mechanism from surface adhesion to cohesion.[198] These properties have been practically used, for example, in DOPA–Fe^{3+} hydrogels, which adhere to tissue to arrest bleeding in rats,[199] or in phosphonate–Mg^{2+} complexes, which act as *in-situ* nanoswitches to induce the adhesion of bioactive nanostructures that regulate cellular functions *in vivo*.[200] Although adhesion behavior is not the focus of this discussion, we would like to highlight the role of structural mechanical properties on the performance of metal-coordination adhesives.[201] In particular, peak tack stress and energy-dissipative volume were found to depend on the elastic plateau modulus and relaxation time of metal-coordinated hydrogel networks.[201] Through this relationship between tack properties and traditional linear viscoelasticity theory, our understanding of metal-coordination mechanics may influence our mechanistic understanding of adhesion.[201]

Other applications

Metal-coordination crosslinks have also been implemented in systems whose mechanical properties have still not been thoroughly explored. In ion-conducting polymers, for example, metal coordination can be used to decouple polymer viscosity and ion conduction[202] because of fundamental differences in the mechanisms that dictate polymer diffusion and ionic conductivity.[76,203] Metal-coordination complexes have also been used in a range of applications, including metal ion chelators,[204,205] rewritable security displays[206–209] and phosphate detectors compatible with HeLa cells.[210] Although the explicit mechanical properties of these devices have not been characterized, it is likely that metal-coordination imparts specific mechanical behaviors relevant to these applications.

In summary, initial applications of metal-coordinated materials are promising, as they demonstrate improved and tunable mechanical properties such as toughness and adhesion.

The diverse properties of metal–ligand coordination also enable a wide range of applications that are still being developed.[211–213]

1.5 Overview of thesis

In this thesis, we will systematically characterize, predict, and design metal-coordinated proteins and polymers with desired dynamic mechanical properties. The systematic characterization in this work serves as an important foundation in the rational design of metal-coordinated materials for mechanical function. We start with an ideal polymer network crosslinked together with individual metal-coordination complexes in Chapter 2. We use the chemical energy landscape of individual coordination complex to predict the macroscopic dynamic mechanical properties of the metal-coordinated ideal polymer network. We then shift investigate protein systems with multiple coordination bonds in Chapter 3. We characterize the dynamic strength of proteins coordinated with multiple metal-ions and show the existence of cooperative and heterogeneous rupture. In Chapter 4, we contextualize the insights on the mechanical properties of simple metal-coordinated proteins and polymers by studying the structural and mechanical properties of metal ions in native biological proteins with high amounts of metal-coordination. Finally, in Chapter 5, we show how large-scale computational methods can be applied to predict the mechanical properties of biomaterials systems when large amounts of data are collected, applying machine learning to collagen mechanical properties prediction as one example. We close the thesis with conclusions and outlooks in Chapter 6.

Throughout this thesis, we use multiscale simulation methods, including molecular dynamics (MD), steered MD, replica exchange MD, coarse grain MD, and metadynamics. We also use quantum chemical simulation methods such as density functional theory. These simulation efforts excellently complement the experimental data we collect using rheology, atomic force microscopy-single molecule force spectroscopy, and isothermal titration calorimetry, to lend fundamental mechanistic insight into the properties we observe.

CHAPTER 2

PREDICTING MACROSCOPIC METAL-COORDINATED HYDROGEL DYNAMIC PROPERTIES FROM INDIVIDUAL COORDINATION COMPLEXES

2.1 Significance Statement

We begin the results part of the thesis by developing a fundamental understanding of the dynamic mechanical properties of metal-coordinated ideal polymer networks, as control over such properties is one of the key advantages of metal-coordination bonds. A framework to effectively predict the mechanical properties of metal-coordinated materials would significantly help the rational design and incorporation of metal-coordination bonds in new materials. In this chapter, we report how such a framework is built: we relate the fundamental first-principles calculations of the metal-coordination bonds to the macroscopic properties of ideal polymer networks crosslinked by these bonds. In ideal polymer networks crosslinked by metal-coordination bonds, the dynamic properties of the network are directly related to the dynamic properties of the individual coordination bonds themselves. Therefore, by experimenting on such a polymer system, we can extract macroscopic network behavior that derives from the microscopic properties of the chemical bond.

Based on this principle, first, we use MD simulations to quantitatively relate the simulated energy landscape of metal-coordination bonds in hydrogels to the experimentally measured macroscopic relaxation time of metal-coordinated hydrogel networks. We show how this relationship can predict the relaxation time of other chemically similar metal-coordinated networks. Second, we use quantum chemical density functional theory calculations to show how the coordination number of the metal-coordination complexes relates to the viscosity of the network. This serves as another strong example of the link between chemical properties of the coordination bond and the macroscopic network.

By using idealized metal-coordinated polymer networks and demonstrating several relationships between computed fundamental principles and macroscopic mechanical properties, this work is a first-of-its-kind critical advance in helping provide new tools for a priori design of dynamic materials.

2.2 Introduction

The dynamics of metal-coordination bonds are intimately coupled to their local microscopic environment, making their smart incorporation into synthetic materials non-trivial. For example, despite the extensive metal-coordination found in both marine mussel threads and marine worm jaws, the marine mussel thread is adhesive, soft, and tough, whereas the marine worm

jaw is hard, stiff, and fracture resistant, differences in material properties likely caused by distinctive protein structures and levels of hydration. A fundamental understanding of the microscopic origin of the macroscopic behavior of metal-coordinate crosslinked materials would enable better utilization of metal-coordination bonds in the future design of loadbearing materials.

Molecular modeling methods offer an opportunity to elucidate the microscopic origin of metal-coordinate bond behavior. Xu and Vidavsky *et al.* used first-principles calculations to correlate the strength of a coordination complex to the macroscopic stiffness of a metal-coordinated polymer network.[129,214] Li et. al also used first-principles calculations to compute force-displacement data to compare against experimental single molecule force spectroscopy data on coordination bonds.[215] Others have also proposed relationships between bond strength and macroscopic material strength.[103] However, these approaches have focused primarily on static properties, or strength of coordination complexes, rather than dynamic bond properties that significantly affect time-dependent macroscopic material properties, which also have important effects on toughness, stretchability and self-healing. The computational exploration of metal-coordination bond dynamics is a new frontier and is especially challenging due to the limitations of conventional metal ion force fields and long sampling methods needed to probe relevant timescales.

In this chapter, we use several computational techniques coupled with experimental rheological experiments on metal-coordinated ideal-network polymers to probe the dynamics of metal-coordinated systems. We find that there is an intimate relationship between the chemical dynamics of the metal-coordinate bond and the macroscopic properties of a metal-coordinated network, and that computational modeling efforts can begin to probe this relationship to enable *a priori* design of properties.

2.3 Molecular understanding of Ni^{2+}-nitrogen family metal-coordinated hydrogel relaxation times using free energy landscapes

In this section, we discover the microscopic behavior of metal-coordination bonds using simulated free energy landscapes and quantify for the first time how this energy landscape relates to dynamic mechanical properties of a bulk metal-coordinate crosslinked material. Selecting biologically relevant metal-coordinate complexes imidazole-Ni^{2+} and histidine-Ni^{2+} as a demonstration model system, we correlate bond lifetimes estimated from simulations of coordination complexes with experimentally measured bulk network relaxation times of ideal-network metal-coordinated gels.[25] Through a quantitative empirical relationship, we show how the energy landscape of individual coordination bonds can be related to the macroscopic viscoelastic behavior of metal-coordinated polymer networks, despite their vastly different time and length scales of investigation (**Figure 2-1**). Specifically, we discuss how key features of the metal-coordinate bond energy landscape may contribute to bulk gel relaxation time and we use these features to expand the conventional Arrhenius equation to predict the relaxation time of other Ni^{2+}-nitrogen-containing metal-coordination chemistries in ideal networks. To our knowledge, this is the first study of its kind to empirically relate the energy landscape of metal-coordinate bonds with the macroscopic dynamic mechanical behavior of metal-coordinated

hydrogels and posit a quantitative empirical relationship between bond energy landscape and bulk network relaxation time. We anticipate the quantitative relationship presented here to be a starting point for the development of more sophisticated models that can predict relaxation timescales of materials with programmable viscoelastic properties.

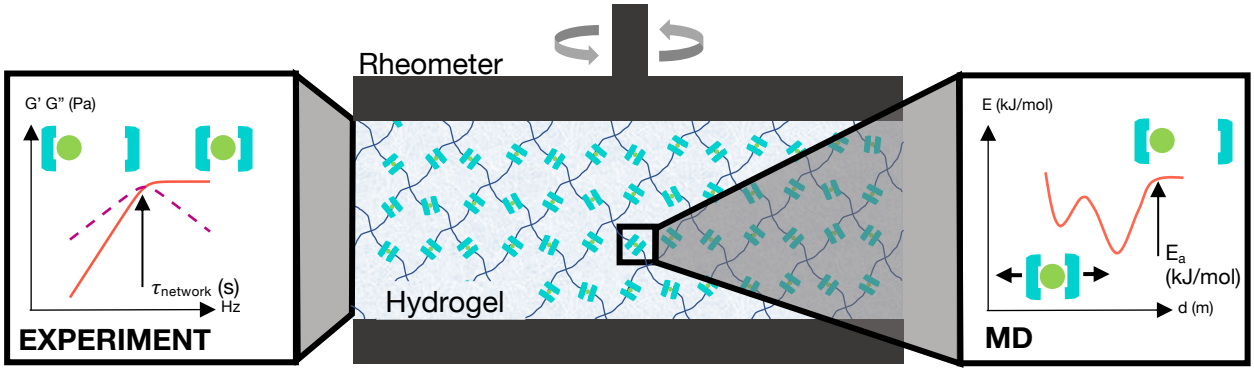


Figure 2-1. Time and length scales in simulation and experiment. The macroscopic network relaxation time (τ) of the hydrogel can be experimentally measured via a frequency sweep with a rheometer. This network relaxation time measurement is an experimental manifestation of the collective dissociation and re-association dynamics of all loadbearing metal-coordination bonds within the network. MD simulation calculates the energy landscape of an individual metal-coordination bond, through which bond lifetime can be inferred. Figure from Ref.[216]

2.3.1 Dynamic behavior of Ni^{2+}-nitrogen family hydrogels

As a demonstration model system, we focus on the Ni^{2+}-nitrogen family (imidazole and histidine) complexes due to their prevalence in biology and the development of appropriate Ni^{2+} force fields. This focus is also in large part inspired by the original study on histidine-metal coordinate crosslinking in hydrogels by Fullenkamp *et al.*, which documented that coordination crosslinks involving histidine, but not imidazole, lead to more solid-like hydrogels.[10] Both histidine and imidazole ligands are nitrogen-containing chelators with similar structures. Imidazole consists of a 5-membered heterocycle where the pyridine-like nitrogen coordinates in a monodentate interaction to metal ions under basic conditions. Histidine consists of the same imidazole heterocycle and an additional amine group which can also coordinate metal ions, thereby composing a more stable bidentate interaction (**Figure 2-2, inset**), as originally proposed by Fullenkamp *et al*.

To isolate the mechanical effect of these stereochemical differences in metal-coordinating ligands, we incorporated imidazole and histidine as telechelic ligands in 4-arm-PEG polymers and established hydrogel networks crosslinked via Ni^{2+} ions. To maximize the mechanical signal-to-noise ratio of metal-ligand dissociation, we employ nearly ideal polymer networks in a regime below entanglements (see calculations in **Appendix 2**), such that bulk relaxation time of the network is dominated by the kinetics of the dynamic metal-ligand crosslinks.[25,29,73]

As demonstrated in **Figure 2-2**, 4-arm-PEG-histidine hydrogel networks display a bulk relaxation time (τ) of τ ~ 400 s, whereas 4-arm-PEG-imidazole networks display a τ ~ 0.06 s at 5 °C based on a Maxwell model fit. This observation supports the hypothesis that the monodentate imidazole-Ni^{2+} bond acts as a more labile crosslink whereas bidentate histidine-Ni^{2+} bond results in a slower crosslink dissociation rate in synthetic hydrogels. A similar large difference in dissociation rates between nearly identical metal-coordinating ligands has been

observed for other stereochemistries,[217] but a mechanistic explanation for these differences remains speculative.

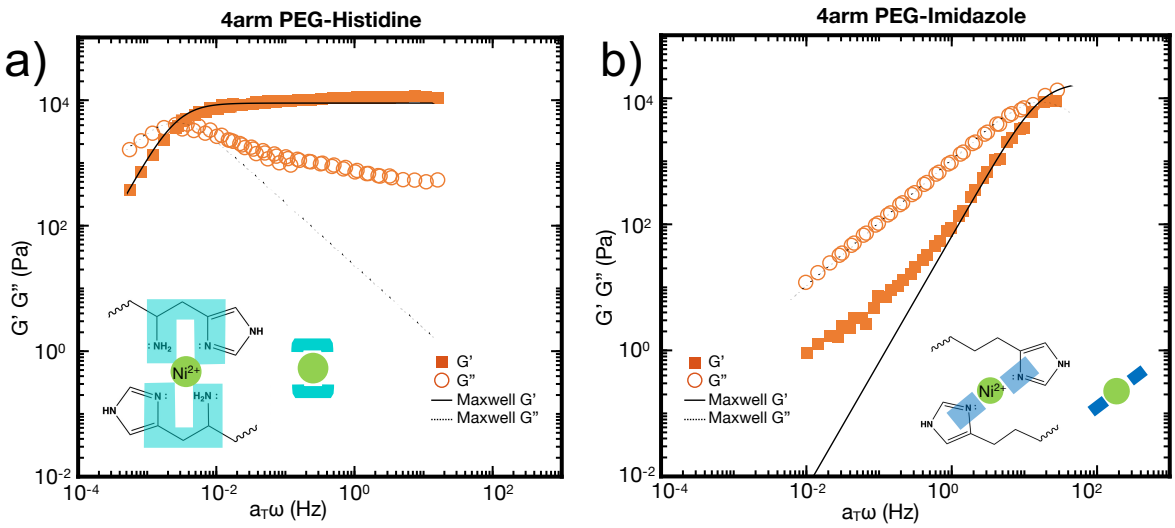


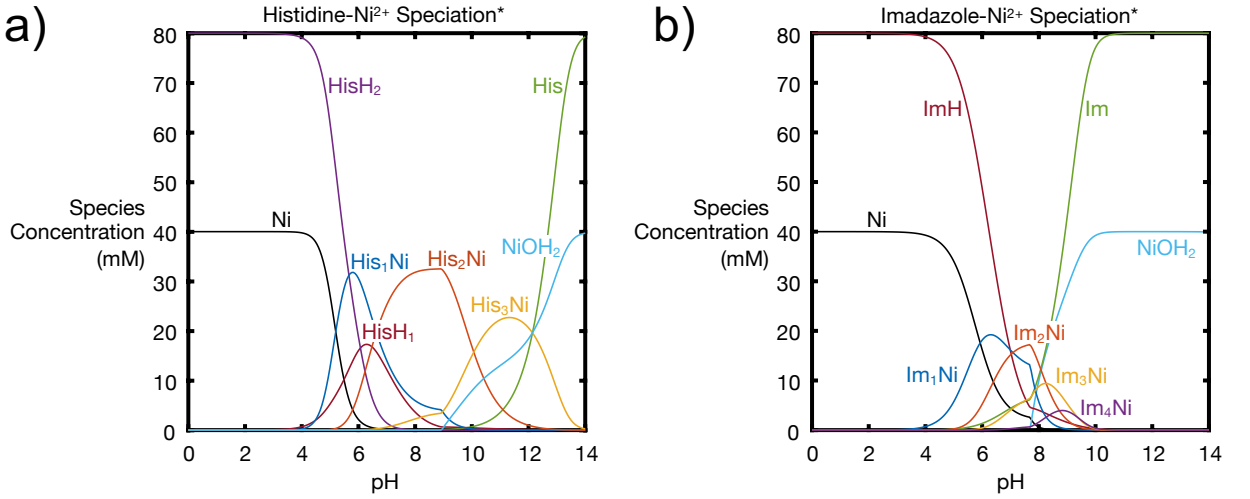
Figure 2-2. 4PEG-histidine gels are slower relaxing than 4PEG-imidazole gels. Time-temperature superpositioned storage (G') and loss moduli (G'') at 5°C from frequency sweeps collected over several temperatures for a) 1M:2L ratio of Ni^{2+}–4arm PEG-histidine and b) 1M:2L ratio of Ni^{2+} –4arm PEG-imidazole at ~pH 8 at ~20% w/v. Schematic of dominant coordination complex in each gel based on equilibrium predictions is shown in insets. Each gel demonstrates Maxwellian behavior, with imidazole gels deviating in the low frequency regime likely due to the broader speciation of coordination complexes present in imidazole gels compared to histidine gels (**Figure 2-3**). Figure from Ref.[216]

2.3.2. Free energy landscapes of Ni^{2+}-nitrogen coordination complexes

To better understand the effect of ligand stereochemistry and local chemical environment on dissociation rate, or inversely bond lifetime, of the coordination complexes, the free energy landscapes of histidine and imidazole coordinating with Ni^{2+} ions were characterized via metadynamics. Metadynamics enables a computationally efficient evaluation of energy landscapes via the addition of bias potentials that overcome local minima.[218–220] To ensure a relevant comparison between experiment and simulation, we first determined which metal-coordinate complexes would be dominant in the hydrogels using thermodynamic equilibrium predictions that use bond chemistry and hydrogel pH to calculate species distributions of coordinate complexes (**Figure 2-3**).[31] Based on these calculations, we proceeded with computing the free energy landscapes of the dominant His_2Ni_1 and Im_2Ni_1 complexes, representing two histidine or imidazole ligands coordinated with one Ni^{2+}, in water solvated environments.

The selection of an appropriate reaction coordinate, or collective variable (CV), is critical in calculating an energy landscape which captures all relevant reactants, products, and intermediate states. For this simulation, the CVs are chosen as distance between Ni^{2+} and the center-of-mass of the coordinating nitrogen atoms on each ligand (**Figure 2-4**). The CV must be selected to enable an efficient exploration of the computational space, while exploring "slow" variables that might prevent the full free energy landscape evaluation. The

appropriateness of the CV is evaluated by plotting the "COLVAR" file to ensure that the collective variables are thoroughly explored.



*Other Ni_xOH_y species with concentration less than 70 μM not plotted

Figure 2-3. Modeling studies based on thermodynamic speciation predictions from binding constants enable selection of appropriate metal coordination complexes for comparison between simulation and experiment. Equilibrium constants allow the estimation of molar concentrations of the different ligand species a) His_nNi^{2+} or b) Im_nNi^{2+} in the 4-arm PEG hydrogels. Speciation curves are computed using the openly available MATLAB program from Cazzell and Holten-Andersen with a ligand concentration of 0.08M (equivalent to 20 wt% used in experimental gels) at pH 8.[31,221] a) Histidine and b) imidazole are plotted at a 1M:2L stoichiometry used in the experimental gels. ML_2 complexes clearly dominate in histidine gels and ML_2 complexes narrowly dominate in imidazole gels. These complexes are selected for simulation for appropriate comparison to experiment. While Im1Ni1 is also present at pH 8, it is not simulated because it does not form crosslinks between polymer chains. These crosslinks preserve the mechanical stability of the metal-coordinated network and contribute to the observed relaxation time from rheology experiments. Notably, the ML_4 imidazole-Ni^{2+} complexes are not dominant at any metal stoichiometry (not shown) or any pH (shown), even though this complex often exists in biological protein metal-coordination sites. This may be because in the hydrogel, the competition between ligand and hydroxide binding to the metal ion is too strong. In contrast, the protein structure can enable four imidazole ligands to exist in a well-defined binding pocket to favor ML_4 metal ion binding. Figure from Ref.[216]

The free energy landscapes are illustrated in **Figure 2-3a,b**. Note that the key features of the free energy landscapes of coordination complexes conjugated to PEG-polymer are preserved when compared to the free energy landscapes of coordination complexes alone (**Appendix S2-1**). Thus, the free energy landscapes of the polymer-conjugated coordination complexes were not simulated further because their large size requires significantly more simulation time.

Compared to Im_2Ni_1, His_2Ni_1 displays a deeper energy well by ~40 kJ/mol indicating stronger binding between histidine and Ni^{2+} than between imidazole and Ni^{2+}, as originally hypothesized by Fullenkamp *et al*. The lowest energy binding state is captured in **Figure 2-5a,b** and represents either 4 nitrogen atoms and 2 nitrogen atoms binding to Ni^{2+} for histidine and imidazole, respectively (**Figure 2-4)**. His_2Ni_1 also demonstrates a metastable binding state where 3 nitrogen atoms coordinate to Ni^{2+}.

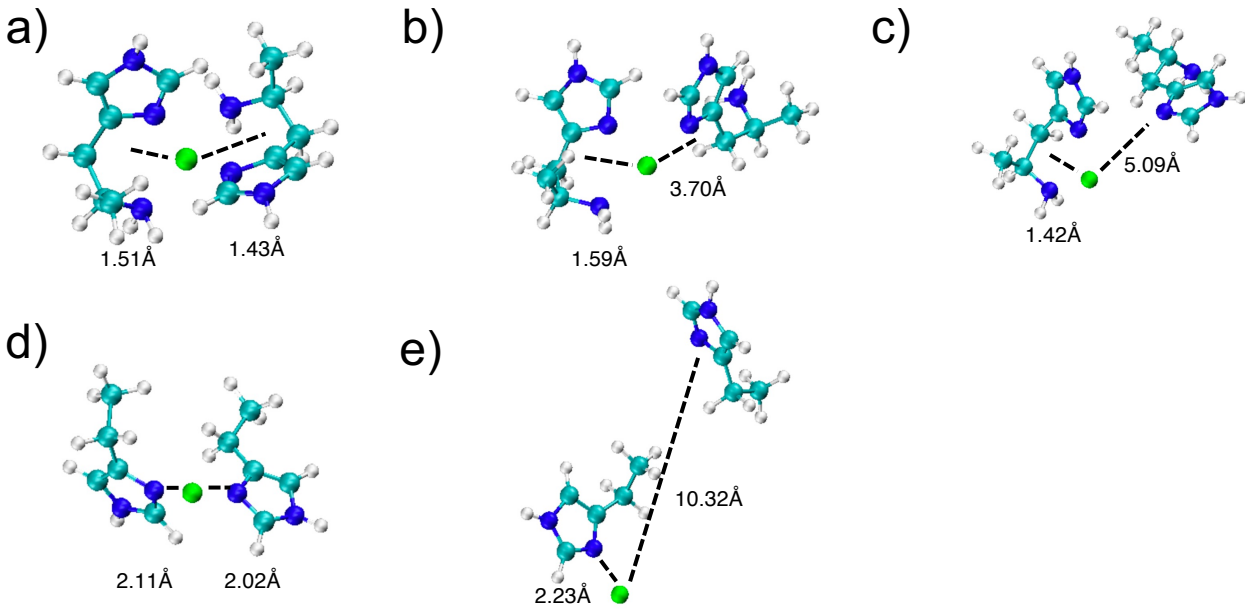


Figure 2-4. Atomistic representations of binding configurations from free energy landscape in Figure 3 showing the collective variables (distances between Ni2+ ions and the center of mass of histidine (a-c) and imidazole (d,e)). Water molecules not shown for clarity, green atom represents Ni^{2+} ion. Table shows distance between collective variable and Ni^{2+} ion for each subfigure. The collective variable, or the reaction coordinate, is selected to be the distance between the Ni2+ and the center of mass of the coordinating nitrogen atoms on each ligand, and is shown in the dotted black line. Figure from Ref.[216]

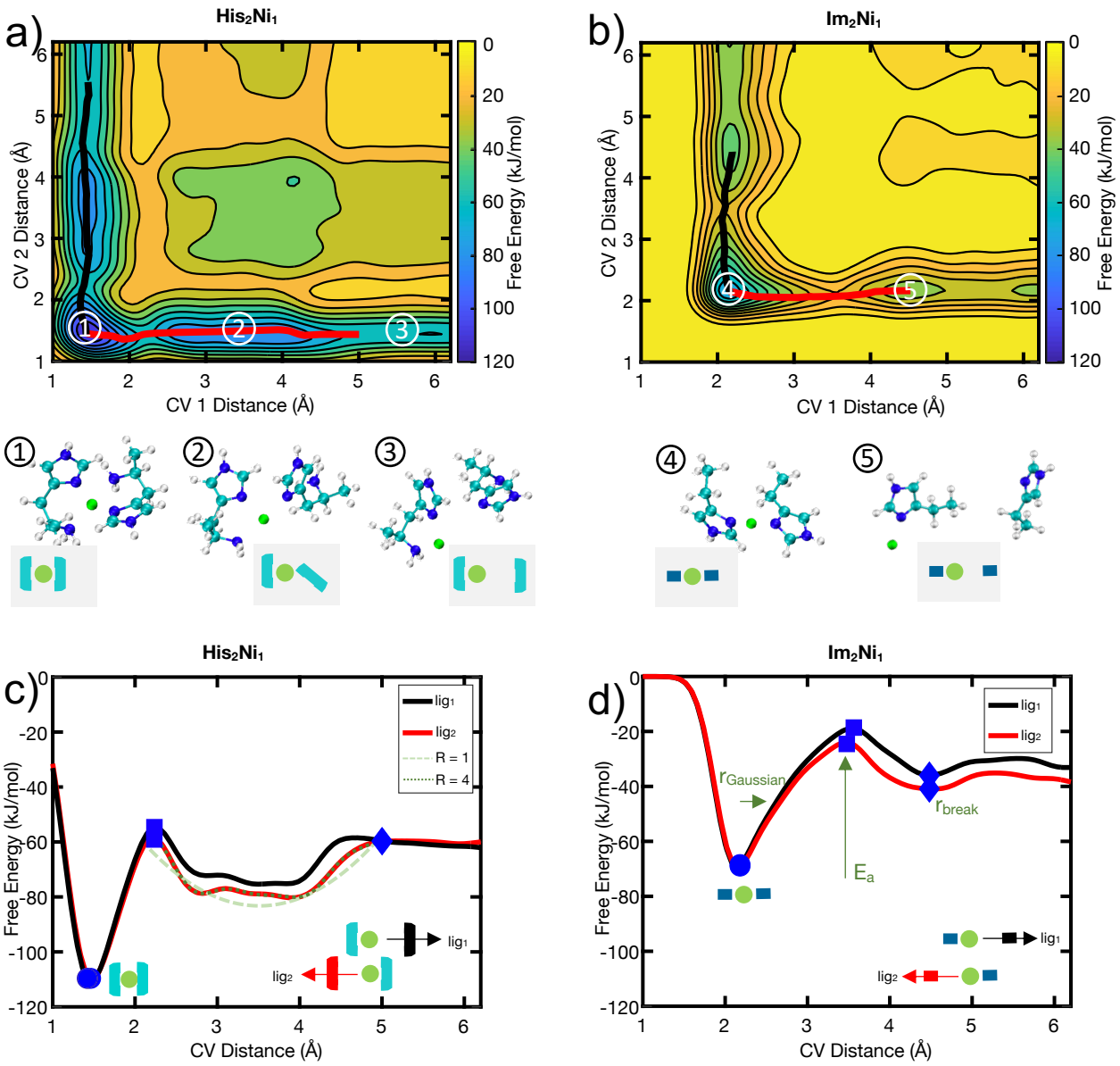


Figure 2-5. Free energy landscapes of His2Ni1 and Im2Ni1 complexes demonstrate clear differences in energies and shape. Free energy landscape of two a) histidine and b) imidazole ligands bound with Ni^{2+} with the collective variables (CV) of the metadynamics simulation plotted as the x and y-axis. His_2Ni_1 has a narrower and deeper well than Im_2Ni_1, indicating more stable coordination complex binding. Inset demonstrates schematics of the binding state of the coordination complexes with the atomistic representations drawn above. Red and black path lines are the minimum energy ligand dissociation path from the lowest energy minima (fully bound state) to the next energy minima (dissociated state with one ligand dissociated). The minimum energy paths are almost entirely parallel with the axes, representing the physical phenomena where one ligand is bound to the metal ion at a constant distance while the other dissociates. This ligand dissociation path is also the microscopic event hypothesized as the origin of the macroscopic relaxation process of the metal-coordinated network. The minimum energy path can be further visualized as 2D energy landscapes for c) His_2Ni_1 and d) Im_2Ni_1 where for the legend entry "lig_1," lig_2 is held constant while lig_1 dissociates and "lig_2" is vice versa (see inset). The landscape is described by E_a, R, and W. c) R is the minimum number of Fourier terms that best fit the metastable well (dashed lines). d) The activation energy (E_a) is computed as the difference between the minimum energy (blue circle) and highest transition energy (blue square) before the bond is broken (blue diamond). W is the Gaussian fitting distance (r_{Gaussian}) divided by the distance at break (r_{break}). His_2Ni_1 has a larger R and r_{break} and smaller r_{Gaussian} than Im_2Ni_1. Figure from Ref.[216]

2.3.3. Predicting macroscopic network dynamics from the energy landscape

To compute quantitative parameters on the energy landscape surface, we compute the minimum energy path, which is the parametric curve that connects two energy minima on the energy landscape surface while using the shallowest ascent to traverse the minimum energy barrier between the two minima (**Figure 2-5a,b** red, black lines). The minima are selected to be the fully associated state, where two ligands are bound to one metal ion, and the dissociated state, where one ligand is dissociated from the metal-coordinate complex while the other ligand remains bound. This minimum energy path from the fully associated to a dissociated state represents the microscopic event hypothesized as the origin of the macroscopic relaxation process of the metal-coordinated network.[52,222] Notably, the minimum energy path (**Figure 2-5a-d** red, black lines) is almost exactly the same as the reaction path where one ligand dissociates while the other remains bound to the metal ion at a fixed distance, which would be perfectly horizontal or vertical lines on the energy landscape. Along this reaction path, we can compute the bond dissociation activation energy (E_a) (**Figure 2-5d**), which plays a critical role in determining relaxation time through an Arrhenius dependence.

Crucially we find that despite the large differences in length and time scales probed in simulation on an individual metal-coordinate complex and in experiment on a metal-coordinated hydrogel network, the estimated E_a barriers governing the system dynamics on both the micro- and macroscopic scales follow the same ordering when comparing across different coordination complexes (**Figure 2-4a**). Specifically, His_2Ni_1 has a higher E_a in simulation than Im_2Ni_1 by ~7 kJ/mol. This is consistent with the slower experimentally measured bulk relaxation time of histidine-Ni^{2+} as opposed to imidazole-Ni^{2+} coordinated networks.

However, differences in E_a between His_2Ni_1 and Im_2Ni_1 alone are unable to account for the large magnitude of difference in relaxation time between the histidine-Ni^{2+} and imidazole-Ni^{2+} hydrogels (**Figure 2-6**). Our simulated energy landscapes allow us to access other critical parameters along the ligand dissociation pathways that may contribute to the observed differences in relaxation timescales.

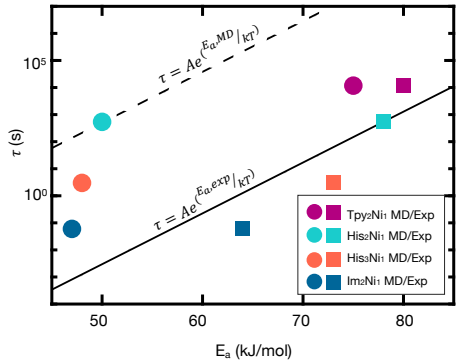


Figure 2-6. Arrhenius prediction of relaxation rates trained on His2Ni1 and Im2Ni1 does not show good alignment with MD or experiment activation energies. Arrhenius equation $\tau = A * \exp\left(\frac{\Delta E_a}{kT}\right)$ trained using the ΔE_a

of His_2Ni_1 and Im_2Ni_1 does not predict relaxation times of the coordination complexes reasonably. The fit is poor because the prefactor A and magnitude difference between $\Delta E_{a,His2Ni1}$ and $\Delta E_{a,Im2Ni1}$ are not enough to capture the magnitude of time difference between His_2Ni_1 and Im_2Ni_1. Circles represent MD data and squares represent experimental data. Figure from Ref.[216]

As such, we propose a simple heuristic relationship using key features of the microscopic energy landscape to explain macroscopic network relaxation time. The parameters are selected based on key observable differences between the energy landscapes of His_2Ni_1 and Im_2Ni_1. One parameter selected is the Gaussian width ($r_{Gaussian}$, as shown in **Figure 2-5d**) of the deepest binding energy well. $r_{Gaussian}$ serves as a proxy for bond stiffness, where a larger $r_{Gaussian}$ implies a less stiff bond. $r_{Gaussian}$ is normalized by the metal-ligand bond length at break (r_{break}, as shown in **Figure 2-5d**) to result in the dimensionless functional form W representing bond stiffness.

$$W = \frac{r_{\text{Gaussian}}}{r_{\text{break}}}$$

In addition to the large differences in E_a and W, His_2Ni_1 has more local minima around the deepest binding energy well than Im_2Ni_1 (**Figure 2-5a,b**). As we will discuss in the next section, these additional minima can be interpreted as metastable binding states which may also contribute to the longer bond lifetime of His_2Ni_1 compared to Im_2Ni_1. We express these additional minima, or ruggedness R, using the dimensionless number of Fourier terms that fit the local minima in the landscape (as shown in **Figure 2-5c**).

Building on the Arrhenius relationship for τ, we include the above stated W and R to yield the following proposed equation for average network relaxation time:

$$\tau \sim b_1 + b_2 * e^{Ea/kT} + b_3 * W^{b_4} + b_5 * R^{b_6} \quad (2\text{-}1)$$

By fitting the coefficients of this equation using the E_a, W and R of His_2Ni_1 and Im_2Ni_1 (**Table 2-1**), we find the coefficients listed in **Table 2-2**. Here the coefficients (b_2, b_3, b_5) represent intrinsic times associated with each aspect, E_a, W, and R respectively, of the energy landscape. The scaling factors (b_1, b_4, b_6) capture time and length scale differences in MD and experiment.

Table 2-1. Table of E_a, W, and R values used in fitting equation

	ΔE_a (kJ/mol)	W	R
His_2Ni_1	54	0.4082	4
Im_2Ni_1	46	0.2944	0
His_3Ni_1	48	0.3622	3
Tpy_2Ni_1	75	0.4524	1

Table 2-2. Coefficients of fitting equation trained on the Ea, W, and R of His2Ni1 and Im2Ni1. Note that the coefficients remain similar when the fitting equation is trained on either His_2Ni_1 or Im_2Ni_1 individually, rather than His_2Ni_1 and Im_2Ni_1 together as presented here.

b1	223.13
b2	3.15e-09
b3	-0.72
b4	-2.92
b5	8.60
b6	3.08

τ increases exponentially with E_a as already expected from the Arrhenius equation (**Figure 2-7b**). We find that τ has a direct relationship with the number of local minima, likely because coordinate complexes such as His_2Ni_1 with additional local minima have more complex dissociation mechanisms, which thereby increase lifetime (**Figure 2-7c**). Further, we find that τ has a direct relationship with W, indicating that a wider bond potential has a longer bond lifetime (**Figure 2-7d**). Interestingly, τ is more sensitive to changes in R than W indicating that the number of metastable states plays a greater role in controlling relaxation time. Given this relationship, we show that for a given activation energy, chosen here as that of His_2Ni_1, the range of relaxation times can vary significantly based on ruggedness and bond stiffness (**Figure 2-7e**).

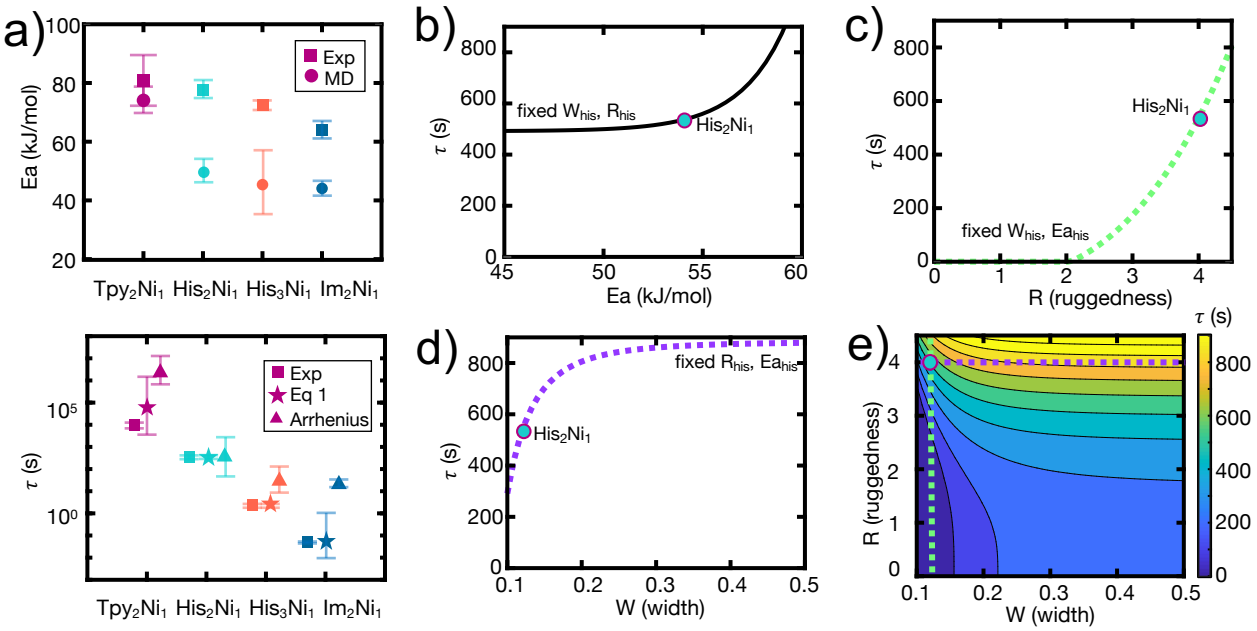


Figure 2-7. Comparison between experiment and simulation show matching qualitative trends and an empirical relationship. a) Experimental relaxation time, experimental activation energy, and simulated activation energies follow the same qualitative trends. The order of terpyridine (Tpy_2Ni_1) > His_2Ni_1 > His_3Ni_1 > Im_2Ni_1 kinetic stability is preserved in both experiment and simulation, even though experimental observations are made on a macroscopic polymer network with coordination complexes and simulation is based on a single coordination complex. The triangles represent the relaxation times predicted by the Arrhenius relationship alone (**Figure S4**), and

the stars represent the relaxation times predicted by **Equation 1**. Note some error bars are very small and difficult to visualize. Effects of b) E_a, c) R or d) W on relaxation time (τ) are plotted when the other two independent variables (W and R, E_a and W, or E_a and R of His_2Ni_1 respectively) are held constant. b) τ increases exponentially with E_a. c) τ increases with R. d) τ increases asymptotically with W. Plotted points represent the τ given the E_a, W, and R of His_2Ni_1. e) Given a fixed E_a of His_2Ni_1, τ can be varied significantly with changing width (W) and landscape ruggedness (R). Figure from Ref.[216]

The above empirical relationship enables a more facile *a priori* prediction of bulk relaxation times using the computational techniques demonstrated in this paper. We probed whether we could use our characterization of the microscopic His_2Ni_1 and Im_2Ni_1 energy landscapes to predict the macroscopic behavior of networks crosslinked with other Ni^{2+}-nitrogen-containing ligands, such as terpyridine (Tpy), a tridentate ligand, or other metal-ligand stoichiometries, such as His_3Ni_1. Remarkably, we find that **Equation 2-1** predicts the relaxation time of a His_3Ni_1 network to be τ ~3s and Tpy_2Ni_1 to be $\tau \sim 1 * 10^5$s which is reasonably consistent within the same order of magnitude as our experiments (**Appendix S2-2, S2-3**).

2.3.4. Discussions, Implications, & Conclusions

Selecting Ni^{2+}-nitrogen family coordination complexes as a demonstration model system, we have shown how 4-arm-PEG-histidine and imidazole Ni^{2+}-coordinated networks differ in macroscopic relaxation times and how that difference is reflected in the energy landscapes of the coordination complexes His_2Ni_1 and Im_2Ni_1. Though histidine and imidazole differ by one coordinating nitrogen, when incorporated as load-bearing crosslinks in ideal polymer networks, they produce vastly different network relaxation times by a factor of ~6,000. In our simple 4-arm-PEG-imidazole hydrogels, imidazole-Ni^{2+} crosslinks are too labile to produce stable solid-like networks at room temperature. The lability of imidazole-Ni^{2+} in hydrogels is surprising because imidazole is a key ligand in coordinating with metal ions in proteins where it provides greater structural stability.[13,17,26,63] In proteins however, higher order protein structure may position imidazole ligands to enable stable crosslinks with metal ions. When taken out of this molecular context, the additional amine coordination provided by histidine is necessary for solid-like gel formation, as demonstrated by Fullenkamp *et al*.[10]

Our simulations indicate that this vast difference in histidine and imidazole's Ni^{2+}-coordinated network relaxation times may be correlated to several differences in the microscopic energy landscapes of their dominant coordination complexes His_2Ni_1 and Im_2Ni_1. First and foremost, as originally proposed based on an assumed Arrhenius-dependent relationship between network relaxation time and the energy barrier to ligand dissociation, we show that His_2Ni_1 indeed has a much deeper binding well with a higher bond dissociation activation energy than Im_2Ni_1. Further, His_2Ni_1 has additional metastable states, likely due to a more complex dissociation mechanism than Im_2Ni_1. The existence of metastable states is similar to protein folding energy landscapes that have several local minima, which often help guide specific folding pathways or regulate folding dynamics.[223] Like in these protein systems, the additional metastable minima in the polymer hydrogel may help slow relaxation dynamics. Potentially, these metastable wells may also enable the ligands to partially dissociate and then re-associate quickly to help in recovering the fully bound coordination complex. This hypothesis could be examined further in future work as an area of exploration, similar to the relationship between local minima in the energy landscape and material recovery suggested in Sun *et al*.[103]

Using these key energy landscape features that distinguish histidine from imidazole, we posited that the relaxation time of the experimental metal-coordinated network can be empirically calculated through an expanded Arrhenius equation that not only includes activation energy, but also bond stiffness, and number of local minima—all microscopic parameters determined from the energy landscape. Each of these microscopic parameters lends critical insights into the effect of energy landscape on coordination bond lifetime. As already established with the traditional Arrhenius equation, a larger bond dissociation activation energy results in a longer lifetime. Further, we find that a less stiff bond results in a longer lifetime, which aligns with Kramers' escape rate theory discussed in the following paragraph where a less stiff, or wider bond width, results in a longer time. The additional local minima also contribute to extending lifetime by enabling different configurations of the ligand-metal ion binding. Not only does the empirical relationship developed in this work lend insight into how different energy landscape parameters affect bond lifetime, but it also enables *a priori* predictions of relaxation times of other similar metal-coordinated gels. Here, the empirical relationship developed from His_2Ni_1 and Im_2Ni_1 allow us to predict the network relaxation time of His_3Ni_1 and Tpy_2Ni_1 reasonably.

We emphasize again that we are proposing an empirical relationship for the microscopic energy landscape and macroscopic relaxation time, rather than a physicochemical theory. This is in part due to limitations with metal ion force fields that currently do not fully capture details of metal-ligand charge transfer.[113,224,225] Given this limitation, we focus on the empirical relationship here as it serves our intended purpose: to both predict relaxation times for other nitrogen-based ligands coordinating with Ni^{2+} and observe the effects of different landscape parameters on bond lifetimes. We also considered other classical rate models,[226] such as Kramers' theory about the escape of a particle over an energy potential barrier, as a basis for our quantitative relationship. The detailed discussion of theories considered in model development is in the **2.3.5 Materials and Methods** section below. More simulations, theory and improved metal ion force fields are needed to derive a detailed physicochemical relationship between energy landscapes and resulting bond lifetimes beyond the empirical relationship proposed here.

The microscopic energy landscape of the individual coordination complex is also only one critical component in the determination of the macroscopic relaxation time of ideal metal-coordinated networks. Despite the ideal nature of the hydrogel network, there are still heterogeneous structures and defects that contribute to the network relaxation behavior.[83] Though the coefficients and scaling factors in **Equation 2-1** capture some of these network-related dynamics effects, additional mesoscale simulations would be required to include network heterogeneity to resemble real hydrogels more closely and probe the resulting range of relaxation times over an ensemble of simulations. More detailed quantum chemical simulations of these coordination complexes may also further help bridge the differences in E_a between MD on the single coordination complexes and experiment on coordination networks (**Figure 2-7a**).

This work dissects an energy landscape of a bond and relate it to experimental macroscopic kinetics of a system. The framework developed here opens several areas of exploration for designing new materials with a better understanding of how bond chemistries affect dynamic macroscopic properties. We strongly believe our approach to the demonstration model system of Ni^{2+}-nitrogen family of coordination complexes will be relevant to other coordination bond chemistries and more complex chemical environments, such as the metal-coordination bonds

within the protein structure of the mussel thread and marine worm jaw, to better characterize their natural energy landscapes. Next steps can be taken to explore the relevance of **Equation 2-1** to other coordination bond chemistries with better metal ion force field development. Different ligand chemistries can also be tested, particularly those that have a narrower speciation like histidine rather than imidazole. Such mechanistic insights hold the potential to guide the *a priori* design of metal-coordinated materials with specific relaxation times towards useful biological viscoelastic tissue mimics and self-healing sustainable construction materials.

2.3.5. Materials and Methods

Synthesis of telechelic 4PEG-Histidine hydrogel

4PEG-His was synthesized using modifications of the procedure by Cazzell *et al*.[31] Briefly, 3 g of 4-arm 10 kDa PEG-NH_2·HCl (0.25 equiv of PEG, 1.0 equiv of $-NH_2$ groups) (JenKem USA) was mixed with Boc-His(Trt)-OH (1.5 equiv) and BOP reagent (1.5 equiv) and dissolved in ~10 mL of dichloromethane. N,N-Diisopropylethylamine (DIPEA) (535 equiv) was added, and the reaction was allowed to proceed overnight under N_2. The product was purified by precipitation once in diethyl ether, thrice in methanol at −20°C, and once in diethyl ether. The product was then dried under vacuum. Protecting groups were removed by a cleavage cocktail of 95 mL of trifluoroacetic acid, 2.5 mL of triisopropylsilane, and 2.5 mL of H_2O for 2 h. An equal volume of methanol was then added, and the solvent was removed using a rotary evaporator. The final product was purified by dissolving it in methanol and precipitation in ether three times and vacuum drying. ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.67 (s, 4H), 7.43 (s, 4H), 4.44 (t, 4H), 3.89-3.33 (m, 848H), 3.29 (t, 8H). Coupling efficiency estimated to be 82%. See **Appendix S2-4** for characterization data.

Synthesis of 4Peg-Imidazole hydrogel

4PEG-Im was synthesized using modifications of the procedure by Fullenkamp *et al*.[10] Briefly, 3 g of 4-arm 10 kDa PEG-NH_2·HCl (0.25 equiv of PEG, 1.0 equiv of $-NH_2$ groups) (JenKem USA) was mixed with 3-(N-1-Trityl- imidazol-4-yl)propionic acid (0.69 g, 1.8 mmol, Chem-Impex International, Inc.) (1.5 equiv) and BOP reagent (1.5 equiv) and dissolved in ~10 mL of dimethylformamide. N,N-Diisopropylethylamine (DIPEA) (535 equiv) was added, and the reaction was allowed to proceed overnight under N_2. The product was purified by precipitation once in diethyl ether, thrice in methanol at −20°C, and once in diethyl ether. The product was then dried under vacuum. Protecting groups were removed by a cleavage cocktail of 95 mL of trifluoroacetic acid, 2.5 mL of triisopropylsilane, and 2.5 mL of H_2O for 2 h. An equal volume of methanol was then added, and the solvent was removed using a rotary evaporator. The final product was purified by dissolving it in methanol and precipitation in ether three times and vacuum drying. ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.61 (s, 4H), 7.76 (s, 4H), 3.89-3.33 (m, 848H), 3.05 (s, 8H), 2.69 (s, 8H). Coupling efficiency estimated to be 78%. See **Appendix S2-4** for characterization data.

Hydrogel formation

Hydrogels were formed by modifying the method described elsewhere.[31] 4-arm PEG polymer was dissolved in 18.2 MΩ cm water at a concentration of 400 mg/mL. 50 µL of this solution was dispensed onto Parafilm®. 0.2M nickel (II) chloride hexahydrate was added to achieve the appropriate metal-to-ligand stoichiometry. 1M sodium hydroxide was then added to reach the final pH and the resulting mixture was then homogenized by mixing with the spatula and by

kneading the material by folding the Parafilm®. The final concentration of the gel was ~20 % w/v, which is above critical overlap concentration and below the entanglement as confirmed by us and others (see **Appendix 2** for calculations).[29,227] The observed relaxation times are well above τ_{chain}.[29]

Experimental testing of hydrogel

The resulting hydrogels were measured for their viscoelastic properties and pH. The gels were loaded onto an MCR 302 stress-controlled rheometer from Anton Paar and tested using a 10 mm diameter parallel plate. Frequency sweeps were performed from 100 to 0.1 rad/s at temperatures from 0 to 35°C at a strain amplitude of 1%, within the linear viscoelastic regime of histidine and imidazole gels. After the 35°C frequency sweep, a strain sweep was performed from 0.01% to 1000% strain at a frequency of 1 Hz to confirm the linear viscoelastic regime of the gels at 5°C (**Appendix S2-5)**. Time temperature superposition was used to combine the frequency sweep results for various temperatures referenced to 5°C. The plateau modulus was determined as the maximum value of the storage modulus and the relaxation time was determined from the crossover of the storage and loss modulus using a Maxwell fit. The pH of the gels was determined using a SoilStik pH meter from FieldScout.

Computational modeling of energy landscapes

MD simulations implemented in LAMMPS[228] are used to investigate the energy landscape of imidazole, histidine, and pyridine ligands using the CHARMM22 force field[229,230]. Ni^{2+} metal ion parameters from Babu and Lim are used as they simultaneously replicate relative hydration free energies, first-shell coordination numbers, and average water-ion distances.[231] PEG parameters from the CHARMM c35r ether force field are used as they replicate persistence length, hydrodynamic radii, and scaling of radius of gyration with molecular weight well.[232] The ligands and metal ion are solvated in a 50Å x 50Å x 50Å water box using TIP3P water molecules. Solvated ligands are equilibrated using an NPT ensemble with a constant pressure of 1 atm at 300 K for 2 ns followed by equilibration with an NVT ensemble for 1 ns using the Nose/Hoover thermostat and barostat. A simulation timestep of 2 fs was chosen with a cutoff for nonbonded interactions at 12 Å. The particle-particle-particle-mesh (PPPM) method[233] was used to compute the long-range columbic interactions and the SHAKE algorithm[234] was applied to constrain high-frequency dynamics from hydrogen-related energy terms in the water molecules. Following equilibration, well-tempered metadynamics was used to study the free energy landscape of the ligands. The simulations are performed using a LAMMPS patch with the PLUMED v2 package.[219,235] The biased collective variables are the distances between the center of mass of the coordinating nitrogens on each ligand and the metal ion. Gaussian hills with a width of 0.2 Å are added at a rate of 1 kJ/mol for every 500 timesteps to bias the energy and a bias factor of 6 is applied. Simulations are run for 200 ns and convergence of the landscape is verified (**Appendix S2-6)**. See codes in **Appendix 2.**

Quantification of free energy landscape parameters

Binding energy is calculated from the difference between the free energy of the entire simulated system (ligands, metal ion, water, counterions) where one ligand is bound and the other is dissociated and the lowest well depth in the landscape where all ligands are bound to the metal ion. Activation energy is calculated from the highest peak before dissociation of the ligand from the metal ion. The breaking distance is defined as the maximum bond length between any coordinating nitrogen and Ni^{2+} is 2.11 Å corresponding to a collective variable distance of ~5 Å for histidine and ~4.5 Å for imidazole. Well width is characterized by a fitting to a Gaussian equation centered at the deepest well.

Alternative approaches to model development

Existing theories of the relationship between energy landscape and relaxation time were also explored as alternatives to the model presented in the main text. We specifically discuss rate k, or the transition rate out of the lowest energy binding well, here as opposed to bond lifetime τ used in the main text to better compare to existing theories that normally discuss rates, although $k = {}^{1}/_{\tau}$.

One of the most important relationships in physical chemistry, the Van't Hoff Arrhenius equation, suggests that the rate of a chemical reaction is exponentially dependent on activation energy ΔE_a. Here, the rate is modified by a pre-exponential factor A:

$$k = Ae^{-\Delta E_a/k_bT} \quad (2\text{-}2)$$

In rheology of transient networks, the Arrhenius equation (Equation 2-2) is commonly used to extract the ΔE_a of the relevant relaxation process after probing the rate of relaxation at several temperatures to extract A and ΔE_a. While the Arrhenius equation is used to determine ΔE_a from rheology experiments conducted at several temperatures, the prefactor A is system dependent and not universally applicable to all relaxation processes. As such, the prefactor A calculated for a His_2Ni_1 system for example cannot be used to calculate the rate of Im_2Ni_1 dissociation even if the ΔE_a of Im_2Ni_1 is known (**Figure 2-6**).

Transition state theory and Kramers' theory of reaction rate were also considered as potential candidates to describe the relationship between energy landscape and k. Kramers' theory in particular not only provides a closed form solution for the rate of escape of a Brownian particle over a potential energy barrier but also includes features of the energy landscape. Briefly, in this theory, a Brownian particle sits in a well in an energy field E(x). As the Brownian particle escapes from the well, it experiences a force from the potential gradient -dE/dx and a frictional force from the solvent. Kramers' theory says that the barrier crossing rate that is exponentially, or strongly, dependent on the energy barrier (ΔE_a), and polynomially, or weakly, dependent on the well curvature (E") at the minimum energy position (x_{min}) and barrier position (x_{max}) and solvent friction (ξ). The classic result for the barrier-crossing rate in high friction regime, which applies to our simulations and experiments where water is the solvent [236,237], is as follows:

$$k = \frac{\sqrt{E"(x_{min})|E"(x_{max})|}}{2\pi\xi} e^{-\Delta E_a/k_bT} \quad (2\text{-}3)$$

This theory suggests that increasing ΔE_a decreases transition rate out of the lowest energy well and that a sharper curvature at a fixed ΔE_a results in higher transition rates, given that a sharper curve generally implies a shorter distance between the well and barrier. More details about the development, assumptions, and application of Kramers' theory can be found in [226,238,239].

We initially calculated rates from the curvature and energy barriers in our simulated energy landscapes from **Equation 2-3**. However, we were unable to yield sufficient agreement between our experimentally determined rates and Kramers'-computed rates. We also modified ξ in the Kramers' equation and added an additional rate term to account for the metastable wells in the histidine and terpyridine systems. However, these rates also did not agree with our experimental relaxation times.

Given the lack of agreement between our experimentally determined rates and Kramers'-computed rates, we further modified Kramer's equation to account for alternative definitions of metastable wells and well curvature. We considered two definitions of metastability defined as ruggedness (R):

$$ruggedness = R = N_{wells}$$

the number of Fourier terms that make up the wells (N_{wells}) of the energy landscape, the definition used in the main text, or

$$ruggedness = R = \frac{d_{metastable}}{d_{break}} + 1$$

the distance in Å across the reaction coordinate that the metastable region ($d_{metastable}$) occupies normalized by the distance required to break the bond (d_{break}). It is offset by 1 to account for landscapes in which there are no metastable wells. We then considered two ways to incorporate the metastability into the rate equation, either through the multiplication or addition of the ruggedness term.

We also modified Kramers' equation with alternative equations of curvature to better capture the anharmonicity of the energy wells, as opposed to the harmonic equations typically used in Kramers' estimations. As such, we fit the binding energy potential well to both the Morse and Gaussian potentials to capture the well anharmonicity. We describe these curvatures (U'') using a dimensionless parameter $U'' * \xi * b_1$ where b_1 represents a natural frequency associated with the curvature. We also considered a dimensionless parameter $\frac{width}{d_{break}}$ where the Morse or Gaussian curvature width is normalized by the distance required to break the coordination bond (d_{break}).

The modified Kramers' equations accounting for the alternative definition of well curvature and ruggedness are presented below:

$$k = \mathrm{w_o} * [U'' * \xi * \mathrm{b_1}]^{\mathrm{b2}} * \exp\left(-\frac{E_a}{\mathrm{kT}}\right) * R^{\mathrm{b3}} \quad \text{(2-4)}$$

$$k = \mathrm{w_o} * [U'' * \xi * \mathrm{b_1}]^{\mathrm{b2}} * \exp\left(-\frac{E_a}{\mathrm{kT}}\right) + \mathrm{b_4} * R^{\mathrm{b5}} \quad \text{(2-5)}$$

$$\mathrm{k} = \mathrm{w_o} * \left[\frac{width}{d_{break}}\right]^{\mathrm{b2}} * \exp\left(-\frac{E_a}{\mathrm{kT}}\right) * R^{\mathrm{b3}} \quad \text{(2-6)}$$

$$\mathrm{k} = \mathrm{w_o} * \left[\frac{width}{d_{break}}\right]^{\mathrm{b2}} * \exp\left(-\frac{E_a}{\mathrm{kT}}\right) + \mathrm{b_4} * R^{\mathrm{b5}} \quad \text{(2-7)}$$

Several of these resulting equations gave us comparable rates to experimentally determined rates, with logical coefficients and prefactors.

Note that the above discussion is merely intended to suggest that there could be other logical functional forms of the empirical relationship between energy landscape and rate beyond the empirical equation presented in the main text. Further simulation and data would be required to arrive at a more robust physicochemical theory beyond the empirical relationships proposed here. We recommend other work that provides more derivations of Kramers' theory applied to

other model systems [240,241], extensions to multiwell [242], or applications to single molecule force spectroscopy experiments [243–245] for consideration. Further, other models besides Kramers' theory, transition state theory, and the Arrhenius equation, can also be considered.[246]

2.4 Coordination stoichiometry effects on the binding hierarchy of histamine and imidazole-M^{2+} complexes

In the previous section, we focused on the quantitative prediction of dynamic mechanical properties of nitrogen-family coordinating ligands (**Figure 2-8a**) with Ni^{2+}. In this section, we extend our analysis to different metal ions including Zn^{2+} and Cu^{2+} and the role of these chemistries on the corresponding macroscopic properties of the metal-coordinated networks. There is an unresolved understanding of the molecular origins of the different dissociation rates exhibited by imidazole motifs crosslinked with Ni^{2+}, Zn^{2+}, and Cu^{2+} ions in these materials. A complicating factor in understanding the "hierarchy" of binding between histidine and these transition metals is that the hierarchy appears to be strongly system or configuration dependent. For instance, the relaxation times of polymer networks crosslinked by coordination between histamine (an imidazole group with an additional amine group) and Ni^{2+}, Zn^{2+}, and Cu^{2+} ions follows a hierarchy of $Ni^{2+} > Cu^{2+} > Zn^{2+}$,[10,25] and the same order is observed in polymer melts crosslinked by acetate ligands and the same metals.[247] This behavior is contrasted by the mechanical reinforcement of synthetic hydrogels with imidazole-rich protein residues; the modulus of these imidazole-rich protein networks crosslinked with transition metal ions show $Zn^{2+} > Ni^{2+}$ for polymer networks with imidazole-rich coiled-coil cross-linking motifs,[102,104] and also for acrylamide copolymer networks with imidazole functional groups.[248] A similar trend of $Zn^{2+} > Ni^{2+} > Cu^{2+}$ is also observed in the case of hydrogels made with imidazole-rich *Nvjp-1* protein found in the jaws of the marine sandworm *Nereis*.[91]

A hypothesis to explain the apparent discrepancy in the metal ion hierarchy between histamine-containing synthetic materials and imidazole-rich protein residues is that both nitrogen atoms in the histamine ligand are available for coordination interactions with M^{2+} in synthetic materials, whereas only one nitrogen atom in the imidazole ligand is available coordination interactions in protein systems where the carboxyl and amine groups form the backbone of the protein. However, the cumulative binding stability constants of imidazole and histamine with M^{2+} follows a hierarchy of $Cu^{2+} > Ni^{2+} > Zn^{2+}$,[111] as does DFT calculation results of the interaction between a single imidazole ligand and M^{2+}.[249] The $Cu^{2+} > Ni^{2+} > Zn^{2+}$ hierarchy is also predicted by the Irving-Williams series, which ranks the stability of octahedral complexes of amines and thiols with first-row transition metal ions.[109] This relative ordering is in contrast with the hierarchy presented above for histamine-rich networks and imidazole-rich proteins. In addition, *in vitro* binding affinities of these transition metals with imidazole-rich Hpn proteins and *in vivo* selectivity of *H. pylori* which use Hpn proteins to sequester transition metal ions show different hierarchies of $Cu^{2+} > Ni^{2+} > Zn^{2+}$ and $Ni^{2+} > Cu^{2+} > Zn^{2+}$, respectively.[250] The exchange time of H_2O with these three transition metals, which is yet another measure of the dissociation rates of bound species to the metal ions, follow an order of $Ni^{2+} > Zn^{2+} > Cu^{2+}$. Thus, the reported binding hierarchies of imidazole-M^{2+} interactions are prevalently system-specific **(Figure 2-8b).**

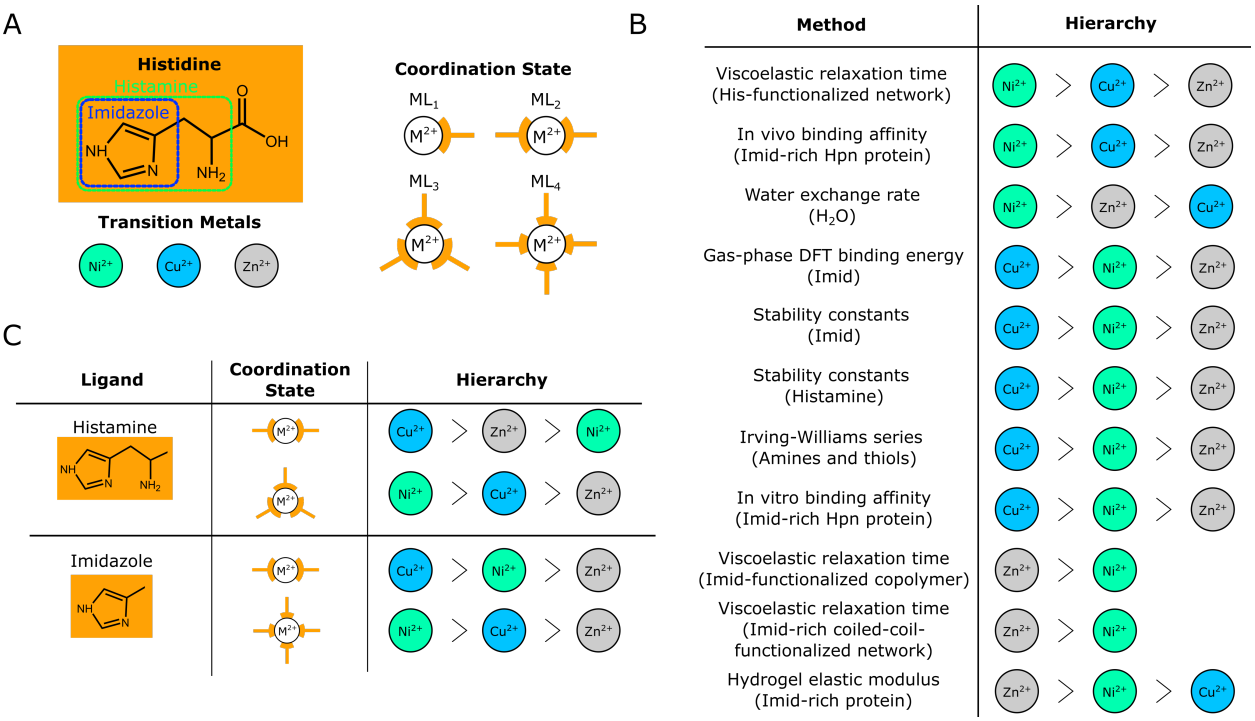


Figure 2-8. Diverse binding hierarchy between histidine and Ni2+, Cu2+, and Zn2+. a) Structure of histidine and its related functional groups histamine and imidazole. These histidine groups form complexes of various coordination numbers with first-row transition metals such as Ni^{2+}, Cu^{2+}, and Zn^{2+}. b) Reported binding hierarchy of Ni^{2+}, Cu^{2+}, and Zn^{2+} with histidine-containing materials and related systems. c) Results from this study for the binding hierarchy of histamine and imidazole with Ni^{2+}, Cu^{2+}, and Zn^{2+} at different coordination states. Figure from Ref.[251]

We suggest that developing an understanding of the binding hierarchy of imidazole with M^{2+} ions must begin with an assessment of the kinetics of the imidazole-M^{2+} interaction in an ideal environment in which the coordination interactions are not affected by other features of the system. An example of a nonideal environment might be the conformational restrictions on the imidazole-M^{2+} interactions in biological systems due to the structural self-assembly of the protein strands. Here we aim to study the binding hierarchy of histidine and imidazole with M^{2+} ions by utilizing a flexible telechelic polymer hydrogel as the model system. In this configuration, the telechelic end-groups of the polymers can be functionalized with either histamine or imidazole to induce transient cross-linking with Ni^{2+}, Zn^{2+}, and Cu^{2+} ions. The binding affinity between the ligands and the M^{2+} can thus be studied by monitoring the viscoelasticity of the hydrogel network, as the dissociation kinetics of such structures govern the relaxation of the hydrogel network.[217] Furthermore, this configuration allows control of the stoichiometric ratio of histidine and M^{2+}, which provides an easy handle for controlling the metal-ligand coordination number (ML_1, ML_2, ML_3, ML_4).[27,29]

Thus, here we characterize the viscoelasticity of transient networks formed with histamine or imidazole and Ni^{2+}, Zn^{2+}, and Cu^{2+} ions under different stoichiometric conditions and complement this effort with a density functional theory (DFT) investigation of the overall affinity of complexes at different coordination numbers. We show that the hierarchy of Ni^{2+}, Zn^{2+}, and Cu^{2+} binding stability is dramatically affected by the coordination number of the metal-ligand interaction (**Figure 2-8c**), and provide a structural explanation of this hierarchy.

2.4.1. Dynamic behavior of Ni^{2+}-nitrogen family hydrogels in buffer

The speciation of metal-ligand complexes is driven by the stochiometric quantities of the metal ions and ligands in the reaction media. This is illustrated by a speciation analysis of Ni^{2+} and histamine (see **2.4.4 Materials and Methods**), which has also been demonstrated in recent work[29,216] (**Figure 2-3, 2-9a**). This example analysis shows that the dominant coordination number of the metal-ligand complexes in the reaction media is strongly stoichiometry-dependent, such that the metal-ligand ratio can be adjusted to preferentially form *mono, bis,* or *tris* coordination complexes. The manipulation of the coordination number via the reaction media stoichiometry will in turn significantly affect the structure and stability of the complexes, which we aim to characterize in this work.

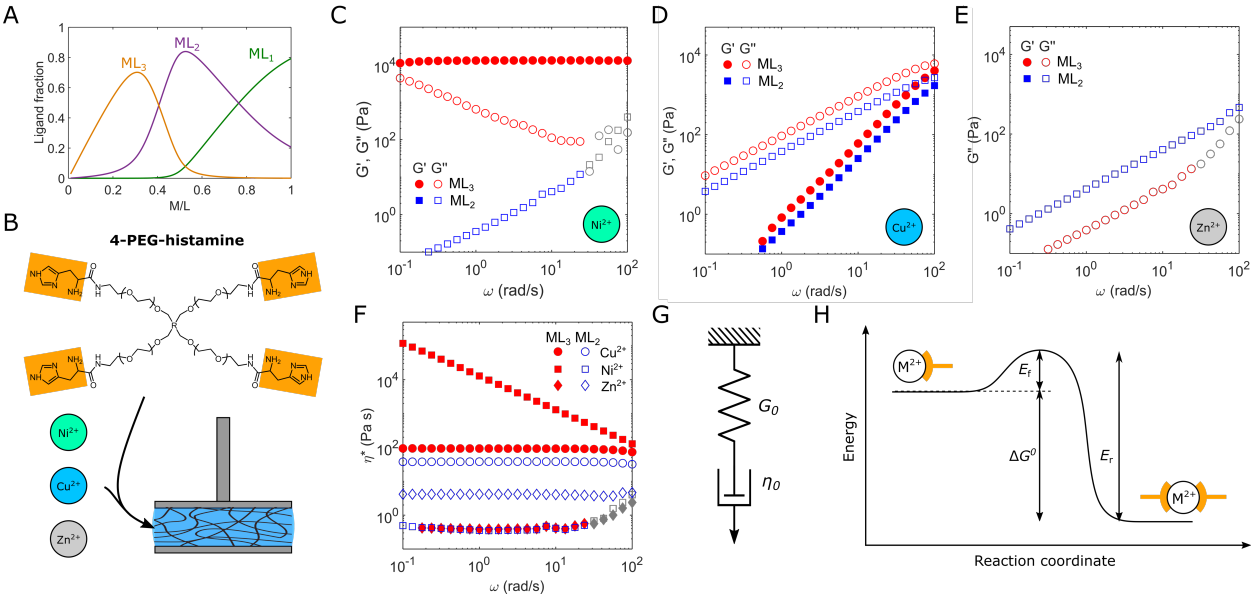


Figure 2-9. Coordination-number-dependent viscoelasticity of a histamine-functionalized polymer network. A) Speciation analysis of histamine with Ni^{2+}. Controlling the stoichiometric ratio of metal ions to ligands changes the population of ML_1, ML_2, and ML_3-coordinated species in the system. B) Structure of 4-PEG-histamine which is mixed with M^{2+} ions under stoichiometrically controlled conditions to design viscoelastic fluids, which are then measured on a rheometer. C), D), and E) Small-amplitude oscillatory shear analysis ($T = 5°C$) of 4-PEG-histamine with Cu^{2+}, Ni^{2+}, and Zn^{2+} at ML_3 and ML_2 coordination numbers, and F) its representation in terms of a complex shear viscosity. Data below the inertial limit of the rheometer is colored in grey. G) Schematic illustration of the Maxwell model of linear viscoelasticity, represented by a spring (with a spring constant G_0) and a dashpot (with a viscosity η_0) in series. H) Schematic illustration of the reaction coordinate underlying a spontaneous metal-coordination event. Figure from Ref.[251]

In our study, we investigate the metal-ligand stoichiometry-dependent rheological properties of 4-PEG-histamine and M^{2+} networks buffered with MOPS, which is a Good's buffer at physiological pH that shows minimal interaction with transition metals such as Cu^{2+}, Ni^{2+}, and Zn^{2+} and is thus used as standard buffers in binding assays.[252–254] This is in contrast to the previous study[29] in which the gels buffered by phosphate-buffered saline (PBS), where PO_4^- groups in PBS can interact significantly with the M^{2+}. The use of PBS can therefore pose complications in studying the stability of the resulting coordination complexes, especially since

interaction between M^{2+} and anionic species can cause a change in the coordination number of the metal centers.[91]

The storage modulus $G'(\omega)$ and loss modulus $G''(\omega)$ of 4-PEG-histamine-M^{2+} networks under different stoichiometries are shown in **Figure 2-9c-e** ($T = 5^oC$). The rheological data exhibits Maxwell-like viscoelastic behavior, as expected based on prior results[10,25] and as expected of transient networks.[255,256] In the ML_3-coordinated networks with Ni^{2+}, Cu^{2+}, and Zn^{2+}, we observe an ordering of the dissociation times of the 4-PEG network, indicated by the cross-over time $\tau_c = 1/\omega_c$, to be $Ni^{2+} > Cu^{2+} > Zn^{2+}$ in agreement with the refs.[10,25] However, this trend is completely different in a network formed with a ML_2 metal-ligand coordination. We find that the Ni^{2+} network completely loses its ability to form a viscoelastic network as indicated by the purely viscous response of the network. The Cu^{2+} network becomes slightly weaker, indicated by the slight lowering of the viscoelastic moduli of the network. The Zn^{2+} network actually becomes slightly more viscous, as indicated by an increase in $G''(\omega)$. Comparing just the complex shear viscosity $\eta^*(\omega) = G^*(\omega)/\omega$ of these networks (which is a measure of the zero-shear viscosity (η_0) of the network via the Cox-Merz rule), we find that for ML_2 stoichiometric conditions, the ordering of dissociation times is $Cu^{2+} > Zn^{2+} > Ni^{2+}$ (**Figure 2-9f**).

For a system following the Maxwell model of viscoelasticity (**Figure 2-9g**), the zero-shear viscosity η_0 of a material is a function of the network elasticity G_0 and the characteristic cross-over time τ_c, such that $\eta_0 = G_0\tau_c$. Here, we can recognize that G_0 depends on the density of cross-links in the network (ρ)[257] and that τ_c depends both on the dissociation rate (k_d) of the cross-links as well as ρ.[255,256] Thus, we can surmise that η_0 in our network is dependent on both ρ and k_d of the cross-linking interactions. ρ is a quantity related to the static equilibrium concentration of the cross-links, and thus related to the stability constant K, which is a measure of the *thermodynamic* free energy change ΔG^0 of the forward binding process between the metals and the histamine ligands. In contrast, k_d is a measure of the *kinetic* activation energy of the reverse binding process, E_r (**Figure 2-9h**).

We cannot decouple the contributions of G_0 and τ_c on the η_0 of the material, since the cross-linker dynamics for the polymer networks (except the $Ni^{2+}His_3$ networks) are too fast for ω_c to be measurable on the rheometer even at $T = 5^oC$. Thus, we are not able to decouple the hierarchies of thermodynamic affinity and kinetic stability of the different histamine-M^{2+} complexes which give rise to the different dissociation times. Nevertheless, we hypothesize that the observed η_0 is dominated by the thermodynamic affinity of the histamine-M^{2+} complexes (i.e. ΔG^0 in **Figure 2-9h**). This is because the binding of histamine with these transition metals is spontaneous[258] and occurs within seconds, evidenced by the rapid gelation of the system. Since the forward reaction is rapid, the forward activation energy E_f should be small, and thus we expect ΔG^0 to be similar to the reverse activation energy E_r (or at the very least, scale in a similar manner).

Thus, the experimental data indicates that there is a coordination number-dependent hierarchy of thermodynamic affinity between histamine and Ni^{2+}, Cu^{2+}, and Zn^{2+}, where the ordering follows $Ni^{2+} > Cu^{2+} > Zn^{2+}$ in a ML_3 configuration but the ordering changes to $Cu^{2+} > Zn^{2+} > Ni^{2+}$ in a ML_2 configuration. While these results provide some clarity on the different contexts in which Ni^{2+} and Cu^{2+} provide the strongest binding to histamine, we recognize that these results cannot be easily compared to results from biological systems. This is because the amine group of the histamine amino acid is incorporated into the backbone of proteins and peptides, and

thus it is rarely involved in metal-coordination interactions in metalloproteins,[126] except in a few some cases with Cu^{2+} or Cd^{2+}.[126,259,260] When the amine group cannot participate in the metal-coordination, the binding of histamine groups with M^{2+} species are likely to be dominated by the interaction with coordinating nitrogen atoms on imidazole.[126] Thus, a more suitable comparison to binding hierarchy exhibited in biological systems may come from studying the hierarchy of imidazole-M^{2+} binding affinity. We thus perform a repeat of our experimental characterization of the complex shear viscosity η^* with 4-PEG-imidazole-M^{2+} networks (**Figure 2-10a**). We study the dynamics of polymer networks formed under ML_2 and ML_4 imidazole-M^{2+} coordination stoichiometry, as metalloproteins rarely have more than four coordinated histidine amino acids unless there are two metals in the metal binding site.[42] For ML_4 coordination, a Ni^{2+} > Cu^{2+} > Zn^{2+} hierarchy of binding affinity is observed, while for ML_2 coordination, a Cu^{2+} > Ni^{2+} > Zn^{2+} hierarchy is observed (**Figure 2-10b-d**). A summary of the experimentally-observed binding hierarchies of histamine and imidazole with Ni^{2+}, Cu^{2+}, and Zn^{2+} are illustrated in **Figure 2-11a-b**.

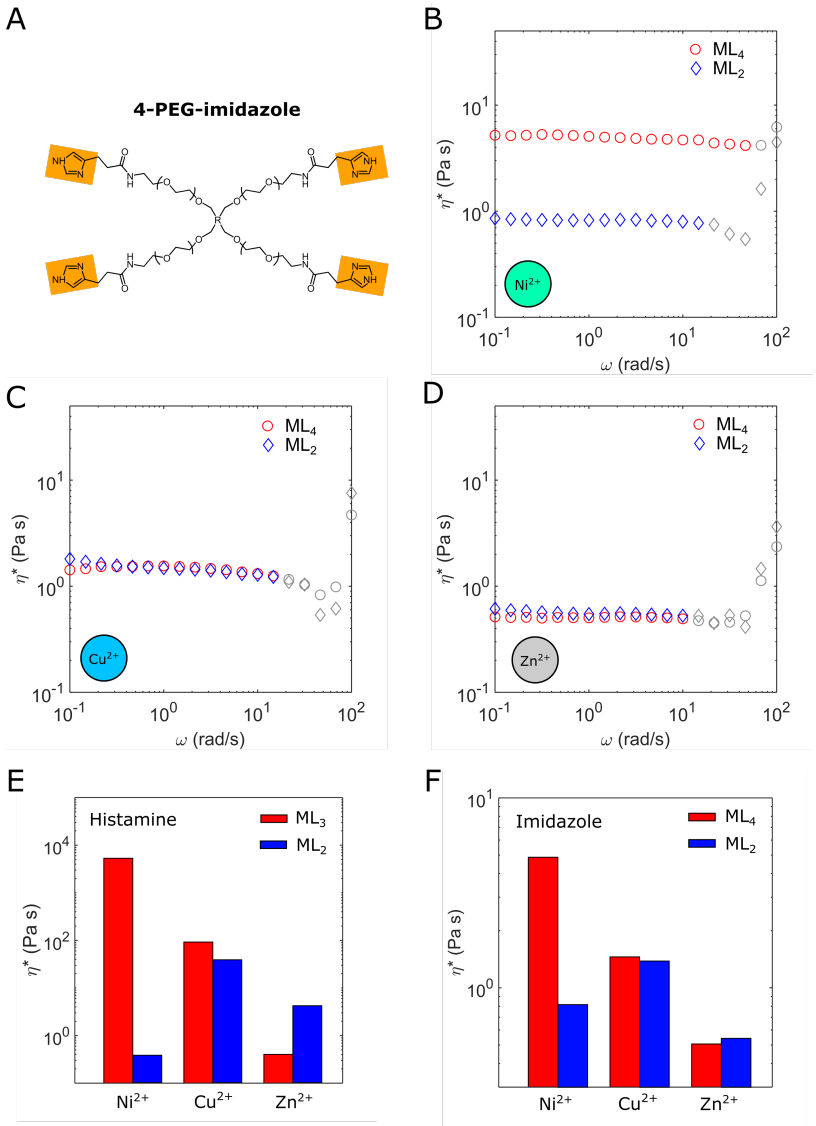


Figure 2-10. Coordination-number-dependent viscosity of an imidazole-functionalized polymer network. A) Structure of 4-PEG-imidazole. B), C), and D) Complex shear viscosity η^* measured via small-amplitude oscillatory

shear experiments ($T = 5℃$) on polymer networks mixed with Cu^{2+}, Ni^{2+}, and Zn^{2+} under stoichiometrically controlled conditions. Data below the inertial limit of the rheometer is colored in grey. E), F) Complex shear viscosity η^* of histamine- and imidazole-functionalized polymer networks with different coordination numbers. Figure from Ref.[251]

2.4.1. Relationship between coordination geometry and viscoelastic properties of network

To validate these experimental results, we perform DFT calculations to determine ΔG^0 of ligand-M^{2+} coordination interaction in stepwise manner beginning with a ML_1 binding state (**SI Methods**). Each metal ion starts with binding six explicit water molecules in an octahedral arrangement as $M(H_2O)_6$. Histamine or imidazole ligands are sequentially added while displacing two or one water molecules respectively. The energy for each ML, ML_2, ML_3, and ML_4 complex is computed as a sum of the ligand addition steps to reach the required binding coordination number. The results of the calculations for histamine (**Figure 2-11c**) and imidazole (**Figure 2-11d**) for the different M^{2+} were in good agreement with the hierarchy obtained via viscosity measurements (**Figure 2-11a-b**). For histamine, experiments and computation show that ML_3 complexes follow a $Ni^{2+} > Cu^{2+} > Zn^{2+}$ hierarchy, and ML_2 complexes follow a $Cu^{2+} > Zn^{2+} > Ni^{2+}$ hierarchy. For imidazole, experiments and computation show that ML_4 complexes follow a $Ni^{2+} > Cu^{2+} > Zn^{2+}$ hierarchy, and ML_2 complexes follow a $Cu^{2+} > Ni^{2+} > Zn^{2+}$ hierarchy.

The DFT simulations also reveal the geometry of the metal-coordination complexes, which can provide insight into the structural origins of the binding hierarchy exhibited by the metals (**Figure 2-11a**). The identified coordination geometries of the complexes are generally in agreement with coordination geometries found in other work. For histamine complexes, Cu^{2+} and Zn^{2+} have been reported to undergo ML_2 coordination through square-pyramidal[261] and tetrahedral[262,263] structures, respectively, while Ni^{2+} is reported to undergo ML_2 and ML_3 coordination with histamine through an octahedral structure.[264–266] For imidazole complexes, Cu^{2+} has been reported undergo ML_4 or ML_6 coordination in square planar or octahedral geometries,[267–269] and Ni^{2+} and Zn^{2+} have been reported to undergo ML_4 coordination in a tetrahedral structure,[117] though the coordination number can be lower in biological protein systems due to metal ion interactions with other coordinating amino acids or water.[126,261,270]

The lowest energy structures from DFT (**Figure 2-11a**) can lend some insight into how the coordination geometries may influence binding hierarchies of the different transition metals. Ni^{2+} is found to preferentially bind in an octahedral geometry, whereas Cu^{2+} and Zn^{2+} are shown to preferentially bind in square planar or tetrahedral geometries.[271,272] These preferential bindings may explain the higher stability of histamine with Ni^{2+} compared to the Cu^{2+} and Zn^{2+} in the ML_3 complex, and the higher stability of histamine with Cu^{2+} and Zn^{2+} compared to Ni^{2+} in the ML_2 complex. This explanation does not explain the high ML_4 stability of imidazole with Ni^{2+} compared to Cu^{2+} and Zn^{2+}, and high ML_2 stability of imidazole with Ni^{2+} compared to Zn^2. In our calculations, we find that octahedral coordination is not preferred by imidazole ML_4 complexes due to the steric crowding of imidazole in octahedral geometries in the presence of water, and find that imidazole with Ni^{2+} shows a higher stability than Cu^{2+} and Zn^{2+} in the ML_4 complex, even though imidazole ML_4 is in a tetrahedral geometry. This might reflect the limitations of a purely structural perspective on the binding hierarchies, and the energetics of binding may also play a role in the binding hierarchy of the metal ions. For instance, the strong preference of imidazole to Ni^{2+} could be explained using crystal field theory, which evaluates the net stabilization of electrons in the d orbitals of metal ions depending on specific ligand

field geometry. From Ni^{2+} (d^8) to Cu^{2+} (d^9) and Zn^{2+} (d^{10}), the incremental electrons would occupy the T2g orbitals with higher energy, thus destabilizing the coordination compound.[273]

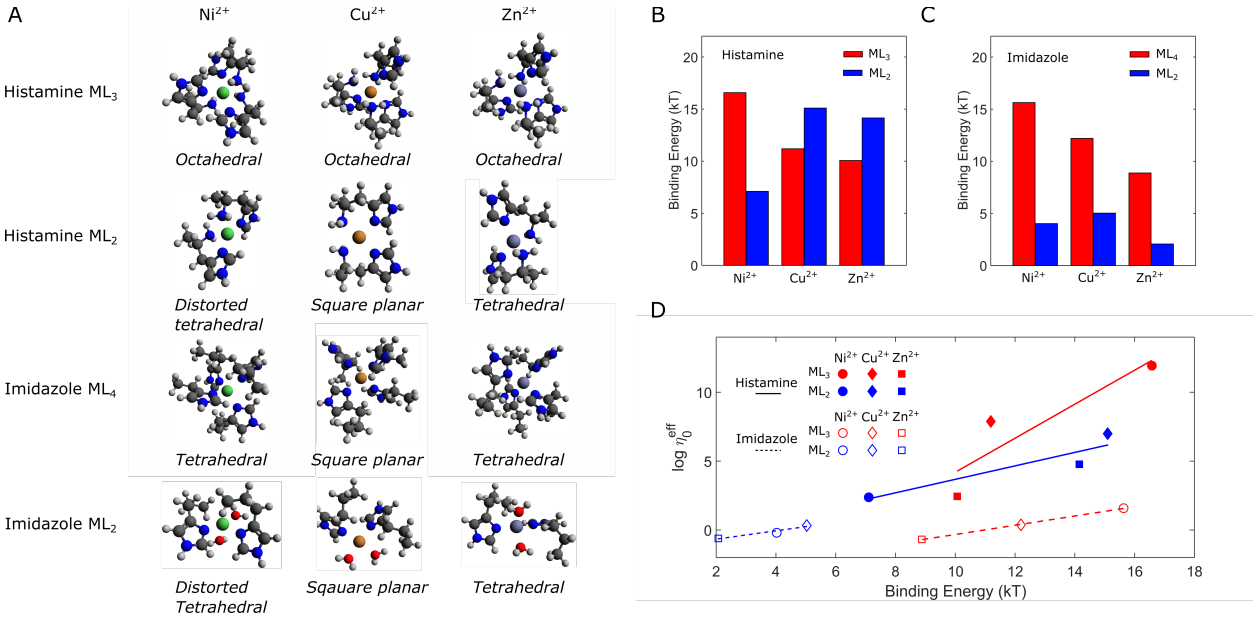


Figure 2-11. Density functional theory (DFT) calculations of the binding hierarchy of histamine and imidazole with Ni2+, Cu2+, and Zn2+ under different coordination states. A) Equilibrated structures of the histamine and imidazole with Ni^{2+} (green), Cu^{2+} (orange), and Zn^{2+} (purple) as computed by DFT. Structures are identified by visual observation. B), C) Binding energy calculations for histamine and imidazole with different coordination numbers. The results are summarized in Fig. 1C. D) Normalized viscosity η_0^{eff} (viscosity values in Fig. 3E and F, normalized by η_0 of the 4-PEG-histamine and 4-PEG-imidazole polymer solutions without ions) versus the binding energies obtained from DFT calculations for histamine in ML_3 and ML_2 complexes, and imidazole in ML_4 and ML_2 complexes for Ni^{2+}, Cu^{2+}, and Zn^{2+}. The lines (solid for histamine, dashed for imidazole) serve as a guide to the eye. Figure from Ref.[251]

A more quantitative comparison of the experimental and computational results is shown in **Figure 2-11d**, where η_0^{eff} (which is the measured viscosity values normalized by η_0 of the 4-PEG-histamine and 4-PEG-imidazole polymer solutions without ions) is compared with the binding energy calculations. As expected, a positive correlation between binding energy (ΔG^0) and η_0^{eff} exists, though the correlation appears to be system-specific. For example, we see that the viscosities of the ML_4-coordinated 4-PEG-imidazole networks are only marginally higher compared to the ML_4-coordinated 4-PEG-imidazole networks, despite the substantial difference in the binding energies of the two coordination states. The reason for this is not immediately clear, although a possible hypothesis for this might be that the weak and dynamic coordination between imidazole and M^{2+} species may not fully prevent the precipitation of $M(OH)_2$ species, resulting in an overall lower population of M^{2+} species especially in the ML_2 state. We also cannot neglect the fact that the simulations do not consider any of the counter-ions present in the experimental system such as the MOPS buffer, and that some quantitative differences may arise as a result.

2.4.2. Discussions, Implications, & Conclusions

We summarize our results in **Fig. 2-8c**. Altogether, these results reveal a strong metal-ligand coordination number dependence on the thermodynamic binding affinity of imidazole and histamine-M^{2+} complexes, in which the hierarchy of binding affinities are dictated by the coordination number of the complexes formed under different stoichiometric conditions. This hierarchy is particularly driven by the tendency of Ni^{2+} to form ML_3 or ML_4 complexes with histamine and imidazole, respectively, suggesting that Ni^{2+} binding dictates in systems where the formation of complexes with larger coordination numbers is allowed, whereas Cu^{2+} binding dictates in systems where primarily *bis* complexes are formed. This might provide insight into the preferred binding state of histidine-M^{2+} complexes in biological systems; for instance, the *in vivo* binding of M^{2+} in *H. pylori* in which Ni^{2+} dominates the metal-binding hierarchy may be a result of the stoichiometric binding of the transition metals with histamine or imidazole in a ratio greater than 2 (i.e., ML_3 or higher), whereas the *in vitro* binding may be dominated by ML_2 complexes in which Cu^{2+} exhibits the greatest affinity to histamine or imidazole.

Our results are relevant in conditions in which histamine or imidazole ligands freely and exclusively bind to the M^{2+} at a neutral pH, deviations from these predictions can occur in the presence of competing species in non-buffered conditions that may compete with the M^{2+} binding, such as counterions[91] and hydroxides,[31] or in systems that exhibit higher-order self-assembly (for instance clustering of the metal-ligand complex).[102,104] For instance, this higher-order clustering of the coordination complexes is reported to occur in imidazole-containing polymer systems reinforced by Zn^{2+}, which can give rise to multivalent cross-linking and thus a slower mode of viscoelastic relaxation in these systems.[102] This may explain the substantially higher reinforcement of viscoelasticity in imidazole-containing polymers by Zn^{2+} than other M^{2+} ions. The aggregate formation by Zn^{2+} is a fascinating behavior—perhaps attributable to the formation of well-controlled tetrahedral coordination structures—which warrants future studies.

Overall, our study reports the kinetic stability of histamine-M^{2+} and imidazole-M^{2+} complexes in their energetically preferred coordination structures, which can provide a starting point for understanding the origins of the different binding mechanics in histidine-rich systems observed in literature. These results also provide a basis for the rational design of mechanically robust histidine-rich metal-coordinate bio-mimetic soft materials. Extending the study to account for metal-ligand coordination in more complex scenarios, such as in the presence of competitive species, would be a natural extension of the study.

2.4.3. Materials and Methods

Materials

4-arm poly(ethylene glycol) amine (4-PEG-NH_2) (MW = 10 kDa) were purchased from JenKem Technology USA, Inc. (Allen, TX). Boc-histidine(trt)-OH (BOC), BOP reagent (BOP), N-N-diisopropylethylamine (DIPEA), dichloromethane (DCM), trifluoroacetic acid (TFA), triisopropylsilane (TIPS), MOPS buffer (MOPS), diethyl ether, and methanol were purchased from Milipore Sigma.

Synthesis of 4-PEG-histamine

4-PEG-histamine was synthesized following the method in the reference.[25] 4 g of 4-PEG-NH_2, 1.19 g of Boc-His(Trt)-OH, 1.06 g of BOP were dissolved in 15 mL of DCM in a Schlenk flask under continuous nitrogen gas flow. 0.86 mL of DIPEA was added to begin the reaction under magnetic stirring, which ran overnight under N_2. The product was precipitated in 50 mL of ether, purified in 50 mL of methanol three times (by dissolving in methanol, precipitating the product in the freezer, and then separating the supernatant via centrifugation), and precipitated again in 50 mL of ether. 95 mL of TFA and 2.5 mL of TIPS were added to the solid product, and the solution was mixed for 2 h under magnetic stirring. The solvent was removed using a roto-vap, and the solid product was precipitated in 50 mL of ether three times, after which the final product was vacuum dried. The final product is a white solid, with the histidine residue exhibiting ^{1}H-NMR peaks (300 MHz, $CDCl_3$) at ca. 7.45, 8.5, and 8.65,[29] allowing us to calculate an average coupling efficiency of 88%.

Synthesis of 4-PEG-imidazole

4-PEG-imidazole was synthesized following modifications to the method in the reference.[10] 1 g of 4-arm 10 kDa PEG-NH_2, 0.23 g of 3-(N-1-Trityl- imidazoleazol-4-yl)propionic acid, 0.32 g of BOP were dissolved in 10 mL of DMF in a Schlenk flask under continuous nitrogen flow. 0.21 mL of DIPEA was added to begin the reaction under magnetic stirring, which ran overnight under N_2. The product was precipitated in 50 mL of ether, purified in 10 mL of methanol three times (by dissolving in methanol, precipitating the product in the freezer, and then separating the supernatant via centrifugation), and precipitated again in 40 mL of ether. 95 mL of TFA, 2.5 mL of TIPS, and 2.5 mL of H_2O, were added to the solid product and the solution was left under magnetic stirring for 2 h. 100 mL of methanol was added, and the solvent was removed using a roto-vap. The solid product was precipitated in 10 mL of methanol, and 50 mL of ether three times, after which the final product was vacuum dried. The final product is a white solid, with the imidazole residue exhibiting ^{1}H-NMR peaks (500 MHz, $CDCl_3$) at ca. 8.61 and 7.76, allowing us to calculate the average coupling efficiency of 78%.

Material Synthesis and Rheology

All 4-PEG-histidine networks were made by a combination of 200 mg/mL 4-PEG-histidine, 0.2 M solutions of either $NiCl_2$, $CuCl_2$ or $ZnCl_2$ and 1.0 M MOPS buffer. The ratio was adjusted such that final viscoelastic fluid consists of 10 wt. % 4-PEG-histidine with a stoichiometric amount of the transition metal to the ligand (ML_2 or ML_3) and with a final buffer concentration of 0.2 M.

All 4-PEG-imidazole networks were made similarly to 4-PEG-histidine. The ratio was adjusted such that final viscoelastic fluid consists of 20 wt. % 4-PEG-imidazole with a stoichiometric amount of the transition metal to the ligand (ML_2 or ML_4) and with a final buffer concentration of 0.2 M.

The viscoelastic fluid was then transferred to an Anton Paar MCR-302 stress-controlled rheometer where all measurements were done on a 10 mm parallel plate geometry. The fluid was loaded and then sealed with mineral oil to prevent evaporation. All small-amplitude oscillatory shear experiments are performed with a shear stress of 5 Pa to stay within the linear viscoelastic regime of the material. Constant shear rate experiments are performed by with a logarithmic shear rate sweep from 0.1 rad/s to 100 rad/s.

Speciation Calculations

The metal-ligand stoichiometry-dependent equilibrium speciation calculations were performed by following the method of the reference.[29] Briefly, this method utilizes the Newton-Raphson method[274] to calculate the equilibrium distribution of metal-ligand species under different stoichiometric amounts of metals. Using Ni^{2+} as an example, we solve for the speciation under the chemical equilibria:

$$\mathrm{Ni^{2+}} + \mathrm{His} \overset{K_1}{\Leftrightarrow} \mathrm{Ni^{2+}His}$$
$$\mathrm{Ni^{2+}} + \mathrm{HisNi^{2+}} \overset{K_2}{\Leftrightarrow} \mathrm{Ni^{2+}His_2}$$
$$\mathrm{Ni^{2+}} + \mathrm{His_2Ni^{2+}} \overset{K_3}{\Leftrightarrow} \mathrm{Ni^{2+}His_3}$$

where we use the equilibrium constants *K* between Ni^{2+} and the related histamine[108] such that $K_1 = 10^{6.8}\ \mathrm{M^{-1}}$, $K_2 = 10^{5.05}\ \mathrm{M^{-1}}$ and $K_3 = 10^{3.12}\ \mathrm{M^{-1}}$.

Computational Methods

All metal coordination complexes were fully optimized using DFT methods as implemented in the Gaussion09 quantum chemistry package. The binding energies were estimated by employing ωB97xD///6-31G*(lanl2dz for metal) level of theory, followed by a higher level optimization using 6-311+G**(SDD for metal).[275] Solvation effects were considered employing SMD,[276] based on charge density of a solute molecule interacting with a continuum description of the solvent. The Gibbs free energy correction was obtained by frequency analysis. Gibbs free energies for histidine were calculated by computing the energy differences in the following equation:

$ML(H_2O)_x + L = ML_2(H_2O)_y + (H_2O)_z$

The metal cation starts as $M(H_2O)_6$. As each ligand is added, it displaces one or two water molecules if it is imidazole or histamine respectively. The binding energy for each complex is computed as the cumulative energies from going from an ML, to ML_2, to ML_3 complex.

2.5 Coarse-grained viscosity simulations with changing crosslinker potential

In the previous sections, we focused on experiments of hydrogels with computational simulations of individual coordination complexes. In alignment with previous research,[25] we found that the dynamic properties of the hydrogels are highly affected by the crosslinker coordination bond chemistry and we sought to uncover underlying chemical principles explaining this relationship. In this section, we present preliminary simulations expanding on our demonstration of the relationship between the energy landscape of individual metal coordination bonds with macroscopic mechanical properties of a metal-coordinated network. We construct a coarse-grained model of the ideal network hydrogel discussed in 2.3. The coarse-grained model allows us to access larger length and time scales to more directly relate the effect of the energy landscape of the individual crosslinker chemistry with the dynamic mechanical properties of the network, similar to the work done by Balazs et al.[277–279] Our preliminary simulations demonstrate that the crosslinker chemistry has an important effect on the resulting viscosity of the network, and we suggest future directions to continue establishing this relationship.

2.5.1 Construction of idealized polymer network coarse-grain model

In the hydrogel, we assume a monodisperse polymer distribution of 10 kDa 4-arm star-PEG molecules that are crosslinked by the metal ions. The polymer beads and water molecules are coarse-grained such that each polymer bead represents one monomer of PEG, and each water bead represents four molecules of water. To describe the chemical interactions of the polymer network, we use the force field parameters discussed in Lee et. al,[280] as these parameters have been validated for the conformation and hydrodynamics of PEG in water. The terminus of each of the arms of the PEG polymer is modeled as a second bead type, which represents the interaction between the metal ion and coordinating ligand. The bead pairwise interaction parameters are tuned to represent that interaction. Note that metal ions are not explicitly modeled, but rather are effectively modeled through changing the interaction type of the terminus (coordinating ligand) of the PEG star polymers.

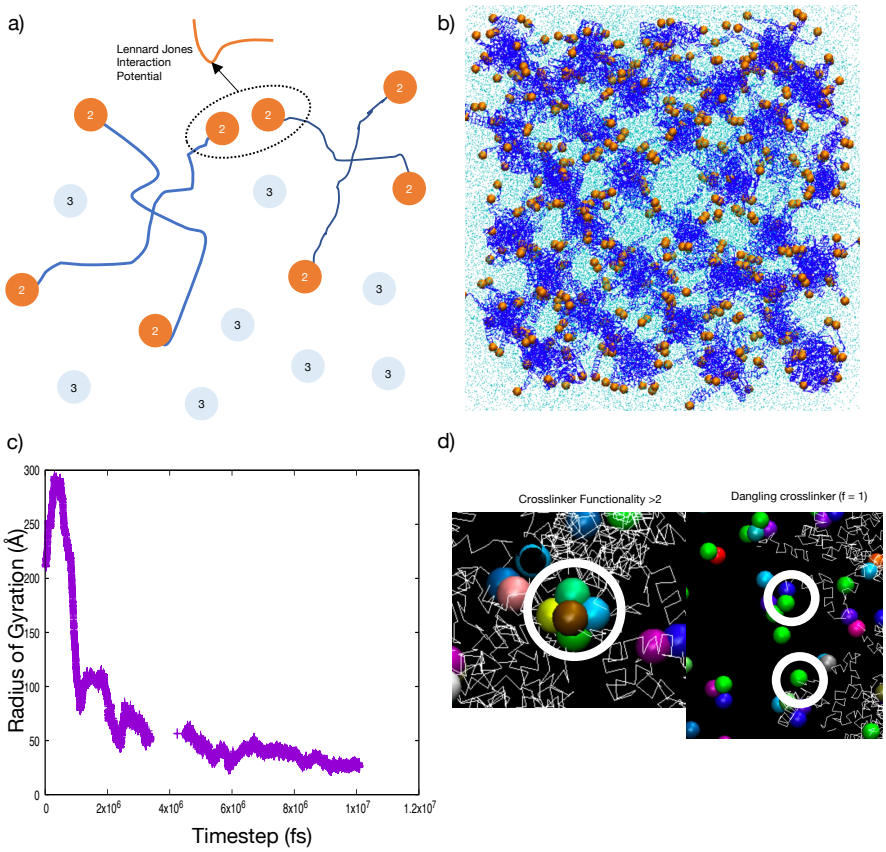


Figure 2-12. Polymer network and equilibration. a) Schematic showing the design of the simulations where type 1 atoms (not shown) represent the polymer, type 2 is the crosslinker, and type 3 is the water. The interaction between the crosslinker is modeled as a pairwise interaction Lennard Jones potential. b) Image of constructed polymer network where blue strands represent polymers, orange circles represent crosslinkers, and cyan lines represent water beads. c) Radius of gyration over equilibration time of one representative polymer in network reaches a stable value. d) Image of defects in network, where defects represent a crosslinker functionality of above 2 (clusters of crosslinkers) or below 2 (representing dangling bonds).

The constructed polymer network is shown in **Figure 2-12a,b,** where the blue strands represent the polymer bead type, and the orange circles represent the crosslinker bead type.

The network is then equilibrated under NVT and NPT conditions until it reaches a steady pressure, temperature, and energy, and the radius of gyration of the polymers also approaches a steady value (**Figure 2-12c**). No explicit bonded interactions are defined between the crosslinker bead types to allow the dynamic breaking and reforming of crosslinks. As such, to ensure that the network stably equilibrates while remaining percolated, an artificially high potential of 10000 kcal mol^{-1} between the crosslinker beads is used, before being switched to the potentials explored in this study after equilibration. Note that because the crosslinker interaction potentials are not explicitly defined as being bonded, there are some network defects that emerge even in our ideal network. Such defects are illustrated in **Figure 2-12d**, and to minimize defects, a shorter potential cutoff of 6 Å is used. For the ideal network, the totally number of defects is 2%, where there is 1 cluster of 3 bonds, and 9 broken bonds prior to shear simulations, and for the non-ideal network, the total number of defects is 30%. Both networks remain fully percolated despite the presence of defects, and the defects reported for the non-ideal network are slightly higher than other quantifications of defects in literature.[281,282]

2.5.2 Metal-coordination chemistry landscape effects on dynamic mechanical properties

Once the network is equilibrated, the pairwise potential of the crosslinker beads is changed to smaller interaction potentials to determine the effect of the interaction potential on the resulting viscoelastic properties of the network. The ideal and non-ideal networks were simulated under varying shear strain rates and the resulting steady state viscosities of the networks were evaluated. **Figure 2-13** shows the effect of crosslinker interaction potential on the resulting viscosity of the ideal (a) and non-ideal (b) network. Both types of networks demonstrate a decrease in viscosity as the shear rate increases. This is consistent with the behavior of a non-Newtonian shear-thinning hydrogels, where viscosity decreases due to the reversible crosslinking mechanisms.[283]

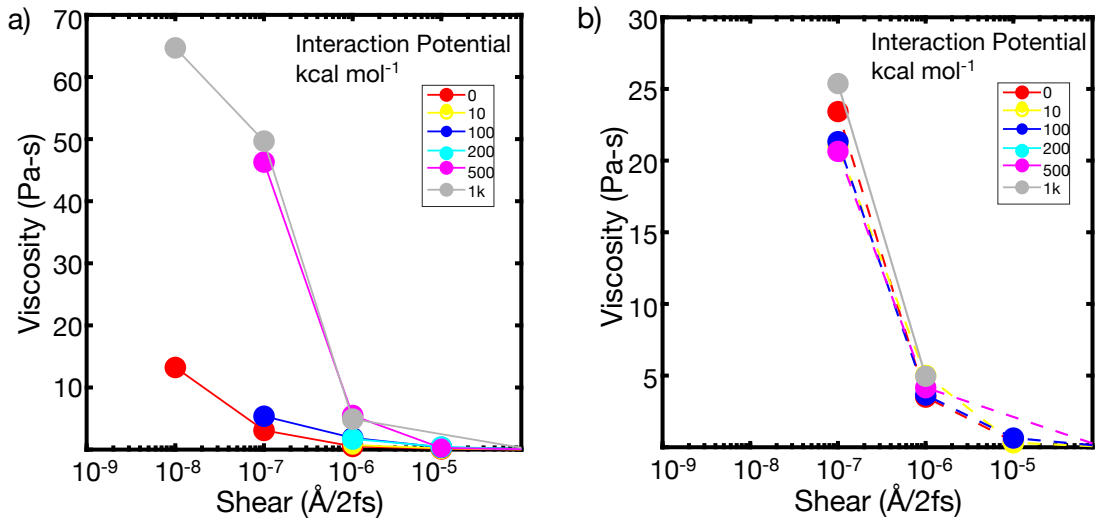


Figure 2-13. Effect of polymer crosslinker chemistry on dynamic mechanical properties of polymer network. Effect of depth of polymer energy well in the Lennard Jones interaction on a) ideal network and b) non-ideal network as a function of shear rate.

In both networks, it is found that increasing the interaction potential of the crosslinker increases the viscosity of the network. This difference in viscosity between the different interaction potentials increases as the shear rate decreases. This is expected, as the lower shear rates probe the zero-shear viscosity, where the hydrogel dynamic properties should be dictated by the breaking and reforming of the crosslinker. When the energy barrier between the crosslinker is higher, it takes more energy for the bond to break, resulting in a higher viscosity of the material. These relationships are less clear in the non-ideal polymer network. In the non-ideal network, the presence of defects and multifunctional sites seem to have a larger effect on the resulting viscosity of the network, which in turn makes it difficult to directly parse out the contributions of the crosslinker potential to the network dynamics. Such network defects may be why it is more difficult to predict the imidazole network dynamics in **Figure 2-7a**.

2.5.3 Discussions, Implications, & Conclusions

The goal of this section was to present preliminary simulations on the effect of the crosslinker potential on the dynamics of a polymer network. We found that for an ideal network, the energy potential of the crosslinker interaction has a strong effect on the resulting viscosity of the network, where a stronger interaction potential results in a higher viscosity. For a non-ideal network, these differences follow a less consistent trend, likely due to the presence of defects in the network. Importantly, these simulations are an initial attempt to directly link the energy landscape concepts presented in Section 2.3 of this thesis, with the resulting dynamic properties of hydrogel networks.

The simulations presented here are currently being further expanded, and several additional steps are required to further validate the relationship between the energy landscape of a metal-coordination bond and the resulting dynamic mechanical properties of ideal network metal-coordinated hydrogels. Due to the long simulation runtimes of the shear simulations, only a select number of shear-rates were tested. Running simulations at lower speeds will take significant computational time, but may start to show a plateau in viscosity to yield a zero-shear viscosity value. The zero-shear viscosity value can more directly be compared to the relaxation time τ measured in the experimental hydrogel networks. An alternative way to compute a more directly relatable quantity to experimental hydrogel networks would be to use oscillatory shear simulations. The resulting storage and loss modulus, and correspondingly relaxation time τ, can be more directly measured through such a simulation. This method was not explored in this thesis due to challenges with long simulation times required for appropriate results. The potentials explored in this thesis are simple Lennard-Jones type interactions, and additional simulations are being conducted to evaluate the effect of changing the shape of the potential, such as by adding additional metastable states. Last, given the mechanistic nature of these simulations, more detailed analysis can be conducted to analyze how the stress in the polymer network is distributed under shear, or how the functionality of the network changes over simulation time. Altogether, this section presents preliminary insights on the role the crosslinker potential plays on the dynamic viscosity of the hydrogel network, and suggests several future directions for study.

2.5.4 Materials and Methods

The ideal network polymer hydrogel was created by tessellating 128 4 arm PEG polymers based on the crystal structure of an Ag2O tetrahedral crystal (https://wiki.aalto.fi/display/SSC/Ag2O). This crystal structure allows a network functionality of 2, which is required to mimic the functionality of the experimental polymer hydrogel. One single polymer is briefly equilibrated under NVT, and then tessellated using the LAMMPS displace commands to create the full polymer network. The polymer is composed of all the same beads, where each polymer bead represents C-O-O, and 233 atoms comprise of one polymer with a molecular weight of 10,000 g mol^{-1}. The beads are described by the potential in Lee et al. who developed force field parameters to model the hydrodynamic properties of PEG in water.[280] The water beads represent 4 water molecules per bead and are also described in Lee et al.[280] Generally, the simulations are implemented in LAMMPS and the bond style is harmonic, angle style is harmonic, dihedral style is Fourier, and pair style is Lennard Jones/cut. The ends of each 4-arm polymer are described by a second type of bead, which is given a different Lennard Jones pair-wise potential. To keep the network fully crosslinked without defining explicit bonds between the polymers, a high interaction potential of a Lennard Jones with an energy of 10,000 kcal mol^{-1}, sigma of 4 Å, and cutoff of 6 Å is used. The network is then equilibrated under NVT for 5 ns, and then further equilibrated under NPT for 10 ns until the pressure, total energy, and radius of gyration of a selection of polymers reaches a stable value. Then, the network is switched to NVT equilibration for ~10 ns at the desired interaction potential of crosslinks before undergoing shear simulation. The temperature of 296 K is used throughout the simulation, and a timestep of 1 fs is used for equilibration.

For the non-equilibrium molecular dynamics shear simulations, varying shear rates are imposed on the equilibrated network under different crosslinker potentials. The NVT integration with the SLLOD equations of motion (fix nvt/sllod) are used with the fix deform command to generate a velocity gradient with the desired strain rate (fix deform xy ${xyrate} erate remap v). The shear is applied in the x-y plane to represent shear stress. A timestep of 2 fs is used for the equilibrations with the desired crosslinker potential and the shear simulations. After a brief run of 0.5 ns after equilibration, the simulation is run for each strain rate until a steady-state viscosity value is reached. The viscosity is calculated by dividing the xy component of the pressure tensor by the strain rate (v_srate) and the length of the box ly (-pxy/v_srate/ly).

CHAPTER 3

ENGINEERING STRENGTH IN METAL-COORDINATED PROTEINS THROUGH COOPERATIVE CLUSTERS OF COORDINATION BONDS

3.1 Significance Statement

In the previous chapter, we characterized the dynamic mechanical properties afforded by metal-coordination bonds, focusing on idealized polymer networks with single metal-coordination bonds. However, natural metal-coordinated systems have several coordination bonds present in various binding arrangements. To more closely replicate the properties of natural metal-coordinated systems, in this chapter, we investigate the role of multiple metal-coordination bonds on the mechanical properties of proteins. Specifically, we test whether proteins coordinated by multiple metal-coordination bonds can exhibit cooperativity and high rupture strength. We also shift our focus from polymers to proteins in this study, as proteins can be designed to have specific secondary structures required to probe our mechanical questions of interest.

Based on this idea, first we conduct a purely computational study to probe the rupture behavior of metal-coordinated peptides. We engineer the cooperative rupture of metal-coordination bonds to increase the rupture strength of metal-coordinated peptide dimers. Utilizing all-atom steered MD simulations on idealized bidentate histidine-Ni^{2+} coordinated peptides, we show that histidine-Ni^{2+} bonds can rupture cooperatively in groups of two to three bonds. We find that there is a strength limit, where adding additional coordination bonds does not contribute to the additional increase in the protein rupture strength, likely due to the highly heterogeneous rupture behavior exhibited by the coordination bonds. Further, we show that this coordination bond limit is also found natural metal-coordinated biological proteins. Using these insights, we quantitatively suggest how other proteins can be rationally designed with dynamic noncovalent interactions to exhibit cooperative bond breaking behavior.

From this understanding, in the second section of this chapter, we experimentally test the strength of the metal-coordinated protein dimers using atomic force microscope (AFM) based-single molecule force spectroscopy (SMFS). We find that the proteins exhibit cooperative rupture, with higher rupture forces than would be expected from a simple addition of multiple coordination bonds. We then use steered MD to lend mechanistic insight into these physical experiments. Altogether, this work provides a quantitative and mechanistic analysis of the cooperativity and intrinsic strength limit for metal-coordination bonds with the aim of advancing clear guiding molecular principles for the mechanical design of strong metal-coordinated materials.

3.2 Introduction

Recent studies have incorporated metal-coordination bonds in materials to engineer strength, self-healing, and energy absorbing properties.[163,284–286] Despite these significant advances, both an understanding of the mechanical role of the metal ions in metal-rich proteins and an extrapolation beyond biological proteins into synthetic materials is missing. This is in large part due to the smaller number of resolved structures of proteins with several metal ion binding sites,[287] unlike the structural resolution for known hydrogen-bonded alpha helix or beta-sheet structures. This lack of structural resolution, coupled with additional challenges around metal-coordination bonds such as polymorphic binding states and speciation,[288] results in an unclear understanding of why natural materials have incorporated metal-coordination bonds or how to optimize their use in biologically-inspired materials towards mechanical function.

The goal of this current chapter is to uncover critical structural-mechanical principles for metal-coordination bonds. We ask whether mechanical cooperativity can be engineered in metal-coordination bonds and to what extent strength and rupture mechanics can be modeled, understood, and predicted from fundamental interatomic and chemical principles. While most of the understanding of the cooperativity of such metal-coordination bonds has been in the context of structure or thermodynamic stability,[289–291] with extensive literature discussing the effects of multivalent binding,[292–294] few systematic studies exist on the cooperative contribution of multiple metal-coordination bonds to the mechanical properties of proteins, and broadly bio-inspired materials engineering.[74,103,295] Developing such a systematic understanding would enable the rational design of mechanically robust metal-coordinated proteins and polymers, complementing the growing experimental work in incorporating multiple metal-coordination interactions in bio-inspired proteins or polymers.[227,296,297]

3.3 Bond clusters control rupture force limit in shear loaded histidine-Ni^{2+} metal-coordinated proteins

To examine the role of multiple metal-coordination bonds on the mechanical properties of proteins, we design *de novo* histidine-rich peptides that coordinate to Ni^{2+} metal ions in an ideal manner and test their mechanical strength using all-atom steered molecular dynamics (SMD) simulations. By analyzing the rupture force of the different metal-coordinated peptides, we find that multiple rupture pathways exist for these coordinated systems and that a maximum of three bonds can rupture at one time. This results in a maximum force limit where adding further metal-coordination bonds does not increase the strength of the peptide. As such, our findings suggest that histidine-Ni^{2+} metal-coordination bonds also exhibit bond-breaking clusters, and that the proposed number of bonds that rupture is the same as the prevailing number of coordination bonds in proteins with excess metal ions. Given this, we extrapolate our findings to demonstrate how proteins might be tuned with dynamic noncovalent bonds more broadly to achieve alternate cooperative bond breaking behavior.

3.3.1 Model Peptide Design

De novo peptides (**Figure 3-1c, Appendix S3-1**) were designed to test whether metal-coordination bonds could behave cooperatively to increase to the mechanical strength of a

metal-coordinated peptide dimer. To probe this question, the transient bond chemistry, the number of coordination bonds, and peptide backbone architecture were specifically engineered. Histidine-Ni^{2+} bonds were selected as the coordination bond because they are a strong binding chemistry and have been widely used and characterized in both biological and non-biological polymer contexts using both experiment and computation.[51,108,298,299]

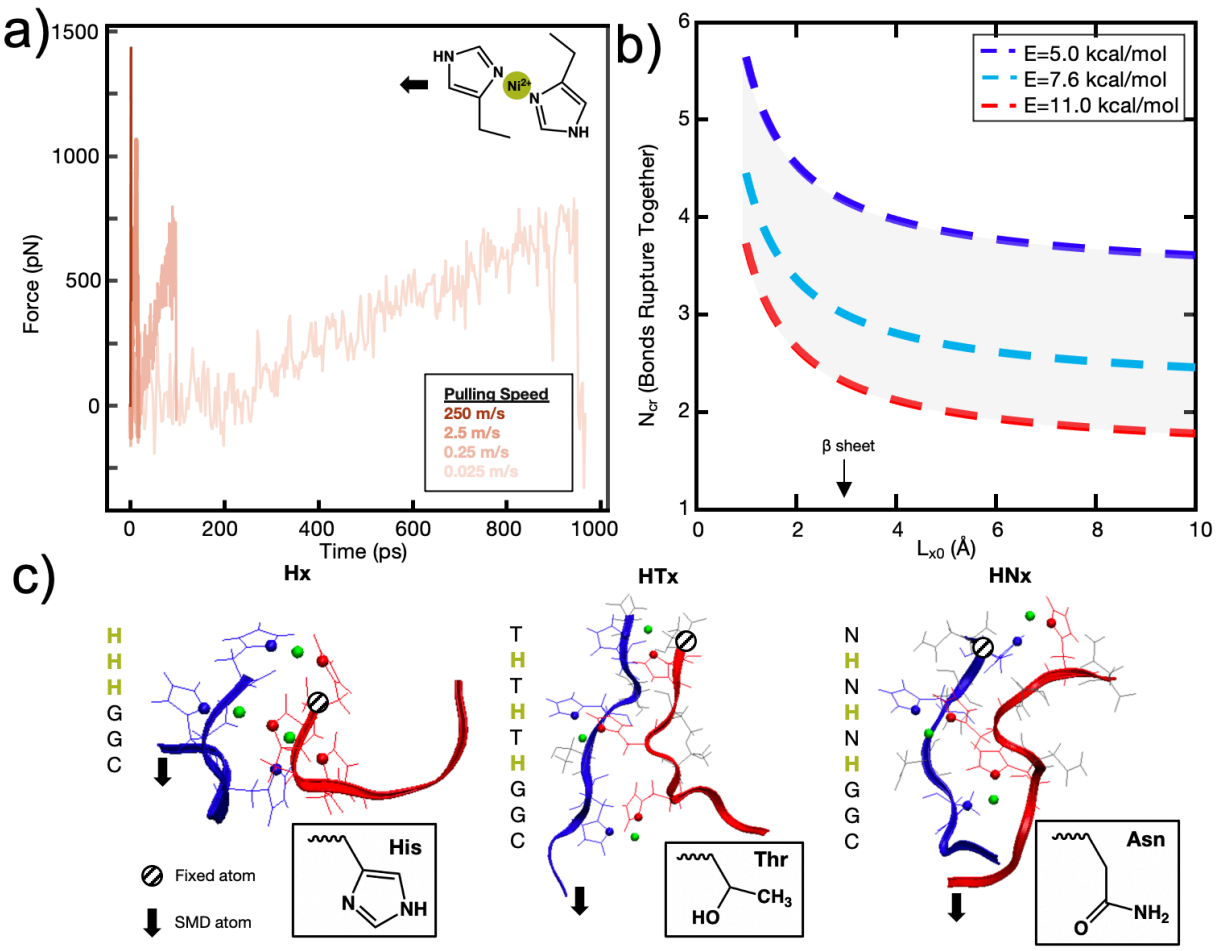


Figure 3-1. A set of de novo model peptides are designed to test metal-coordination cooperative rupture behavior under mechanical loading. a) SMD on a $(histidine)_2$-Ni^{2+} complex (shown in inset) at varying pulling rates shows that the rupture force decreases as the pulling rate decreases. b) The Griffith fracture theory, modified to predict the fracture of hydrogen bonds[300], predicts an N_{cr} for histidine-Ni^{2+} bonds to be 2-3 bonds based on the SMD Bell Model energy of the $(histidine)_2$-Ni^{2+} complex shown in **Figure 3-1a**, and 5-6 bonds for a lower bond energy calculated in literature. c) Three peptide systems are designed for study—a purely histidine system (Hx), and a histidine (H) system with a threonine (T) or asparagine (N) as a spacer residue between the histidine residues. Side chains of each amino acid are shown in the inset, Ni^{2+} atoms are in green, and coordinating nitrogens on the histidines are shown as spheres. Explicit water molecules are not shown for clarity.

To inform the number of metal-coordination bonds that should be placed along the peptide backbone, we applied an earlier model,[300] which predicts the number of hydrogen bonds that break simultaneously (denoted as N_{cr}), to our transient histidine-Ni^{2+} peptides. The equation is rooted in the application of energetic arguments originally proposed in the Griffith fracture theory, which dictates that the release of energy during the fracture process is counterbalanced by the energy penalty required to create new surfaces.[301] The Griffith fracture

theory applies to a crack under tensile loading, which mimics the tensile pulling conditions in our simulations, and the theory has been applied to nanoscale phenomena in earlier work.[302,303]

The following parameters are used to determine the N_{cr} prediction for the histidine-Ni^{2+} bonds: characteristic time scale of τ=0.06 s[216] (for reference, τ~0.1s for a 4PEG-His3 polymer in TRIS buffer which likely forms only 1 coordination bond)[74], applied pulling distance at the moment of rupture x_b of 2 Å, persistence length l_p of 0.4 nm,[304–306] and a bond energy of ~11 kcal/mol based on the SMD studies done in this work on the $(histidine)_2$-Ni^{2+} complex (**Figure 3-1a**). These inputs result in an N_{cr} prediction of 2-3 bonds, given a coordination bond spacing distance (L_{xo}) of 4-7 Å (**Figure 3-1b**). For comparison, the N_{cr} of 5-6 bonds is predicted if other bond energies suggested in literature are also shown.[103,251]

To most closely mimic prior research efforts on the cooperativity in hydrogen bonds in beta-sheets,[300,307] and to isolate the mechanical contributions of the metal-coordination bonds from the unwinding of a protein backbone,[308] we designed peptides with a secondary structure of a beta-sheet-like linear protein. Histidine has a neutral propensity towards a beta-sheet, so to help drive the beta-sheet formation, two molecules were designed with spacer amino acids $(H-)_x$ of threonine (T) and asparagine (N), which have both shown to have a high propensity for producing a flat beta-sheet when in solvent-exposed conditions (**Figure 3-1c**).[309–311] Further, though the polarity of threonine and asparagine are required to aid in solvent-exposed beta-sheet formation, they exhibit a lower propensity for metal-coordination interactions with Ni^{2+} compared to histidine, which indicates that histidine-Ni^{2+} interactions should dominate the coordination binding modes.[312,313] The resulting sequences are CGG-$(HT)_n$ or CGG-$(HN)_x$. In addition, the peptides were designed such that metal-coordination bonds would have a dominating effect on the mechanical properties, as opposed to other secondary structure elements such as interchain hydrogen bonds (**Figure 3-2**).

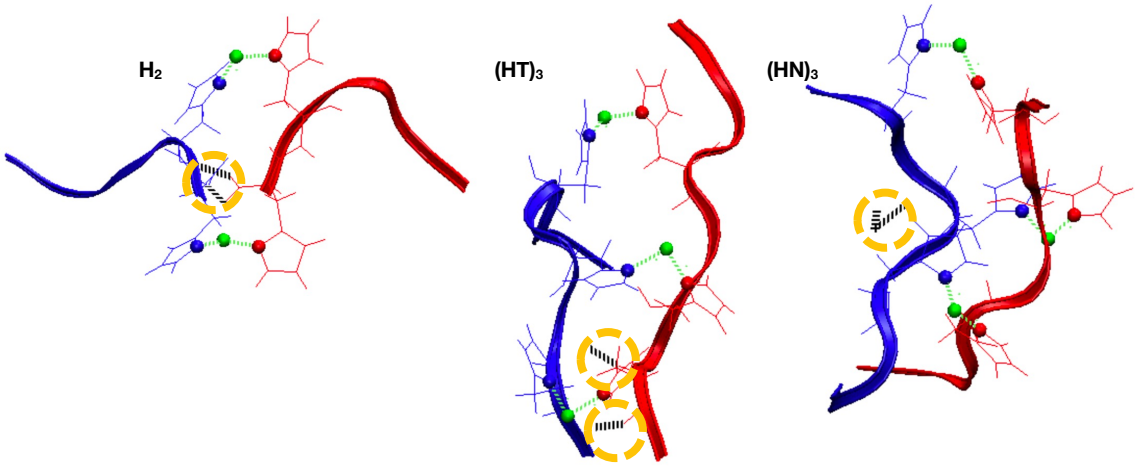


Figure 3-2. Minimized interchain hydrogen bonding between peptide chains. Our peptides were designed to minimize the number of interchain hydrogen bonds, such that the mechanical properties of the peptide dimers could be attributed primarily to metal-coordination bonding. As such, all peptide dimers were analyzed using the Hydrogen Bonds tool in VMD to determine if there was interchain or intrachain hydrogen bonding. The VMD tool uses a 3 Å and 20 degree angle cutoff between hydrogen bond donors and acceptors to designate the presence of a hydrogen bond. ~60% of our peptide dimers were found not to have interchain hydrogen bonding. Only 1 repetition of H2 had two hydrogen bonds, equivalent to the number of metal-coordination bonds in the protein. All other simulations (~97%) had a less than one ratio of number of interchain hydrogen bonds to number of metal-coordination bonds. Because metal-coordination bonds are stronger than hydrogen bonds, and because in all of our peptide simulations, there are more metal-coordination bonds than hydrogen bonds, we attribute the mechanical

properties of the dimers to metal-coordination bonding. A few examples of hydrogen bonding are shown: H2 with 2 interchain backbone hydrogen bonds and 2 interchain metal-coordination bonds; $(HT)_3$ with 2 interchain backbone hydrogen bonds and 3 interchain metal-coordination bonds; $(HN)_3$ with 2 intrachain backbone hydrogen bonds and 3 interchain metal-coordination bonds.

In addition to the *de novo* beta-sheet-like $(H-)_x$ proteins, we also tested derivatives of the canonical his-tag system with regards to their cooperative behavior (**Figure 3-1c**). The his-tag is typically composed of 6 histidines, although the number can vary from two to ten histidines, and it exhibits a high affinity to Ni^{2+} in a Ni-NTA column that has enabled the his-tag to be widely used in protein purification.[314] His-tags have been studied with AFM-SMFS[315] and significant work has also elucidated the structural stability of the his-tag given histidyl mutations.[290] Given these prior research efforts, we also test varying numbers of histidine-only H_x under applied mechanical force.[298]

3.3.2 Rupture Force Behavior and Threshold

Generally, we find that increasing the number of coordination bonds in the peptide seems to increase the observed RF values to a certain extent. For example, $RF_{H3} > RF_{H2}$ and $RF_{HN4} > RF_{HN3}$, which suggests that adding the additional coordination bonds increases peptide strength. An alternative interpretation is that having more coordination bonds along a backbone gives more possible opportunities for cooperative binding. However, this increase in strength of the peptide is not infinite. All three peptide systems exhibit a rupture force limit, indicated by the gray transparent box in **Figure 3-3,** where increasing the number of coordination bonds does not increase the RF of the peptide. This limit occurs around H3/H4 for the Hx system, HT3/HT4 for the HTx system, and HN4/HN5 for the HNx system. The location of the RF limit with respect to the number of coordination bonds on the peptide indicates that roughly three to four bonds work together to rupture in these peptide systems. This is because having the fourth coordination bond in the case of Hx or HTx, as an example, does not increase the RF value of the peptide, implying that three coordination bonds would be enough to reach the strength. The value of three to four bonds aligns with the N_{cr} prediction for **Figure 3-1b**. Note that the canonical his-tag (H6) in a bidentate trans configuration was unstable during equilibration after several attempts to ensure binding and was therefore not shown in subsequent SMD simulations.

The RF data is normalized to the number of coordination bonds in the system and is plotted in **Figure 3-3d-f** to demonstrate the diminishing-returns effect of adding additional coordination bonds to the peptide. This figure is one way to visualize how each coordination bond contributes to the rupture of the system, assuming that all coordination bonds equally contribute to the strength of the dimer. From this normalized figure, we made two observations. First, H2/H3, HT2/HT3, and HN3/HN4 overlap in their normalized RF contribution. These normalized threshold values are 1 bond unit below the RF bond threshold values from **Figure 3-3a-c**. Second, each additional bond contributes less to the strength of the peptide. If each metal-coordination bond in an $(H)_x$ system was to contribute to the mechanics of the peptide in the same strength as $(H)_{x-1}$, we would expect the normalized peptide rupture forces in **Figure 3-3d-f** to be superimposed. This implies that there is an optimum value of MC bonds that achieves maximum strength without wasting additional resources in the form of extra coordination bonds.

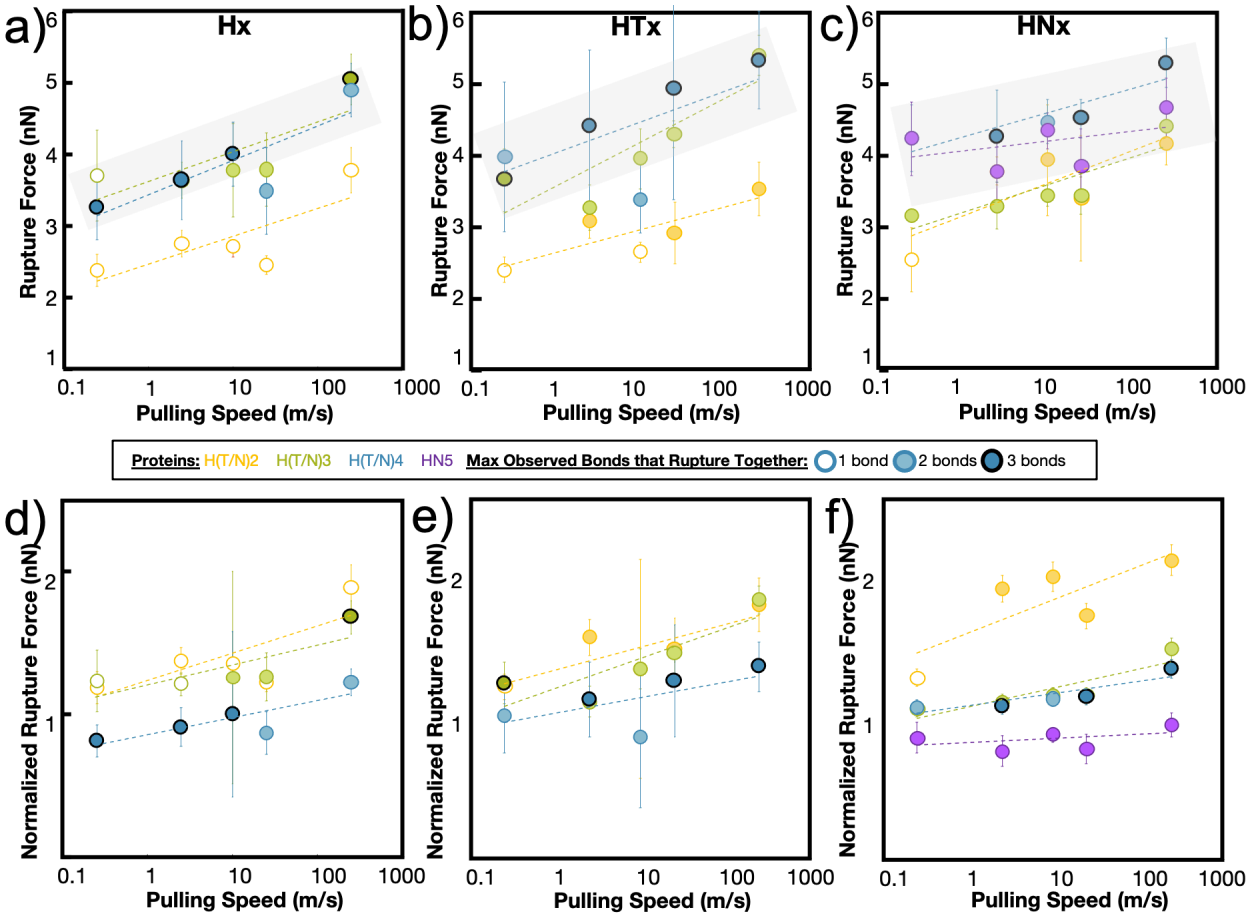


Figure 3-3. Simulation rupture force (RF) vs. pulling speed shows that increasing MC increases RF to a certain extent. Maximum force of bond rupture shows that the RF increases with increasing metal coordination, but only up to a certain extent for the a) Hx, b) HTx, and c) HNx systems. The increase in rupture force approaches a limit (gray transparent box) around H3/H4, HT3/HT4, and HN4/HN5. The colored data points indicate the maximum number of bonds that are observed to rupture together, and only a maximum of 3 bonds are observed to rupture at once even when more metal-coordination bonds are present in the system. H6 was also tested, but is not stable under equilibration and therefore not shown. d-f) When the rupture forces are normalized by the number of histidines in each system, the contribution of each histidine to the total rupture force of the system decreases, indicative of diminishing returns.

To explore this further, and to seek empirical evidence from naturally evolved protein structures, we conducted a search of representative metal-coordination sites of biological metals using the MetalPDB database.[287] We examined representative protein structures with 5-10 metal sites and analyzed how many of those metal ions were consecutively arranged in the protein (**Figure 3-4a**), where consecutive is defined as separated by a distance of 3 Å but not necessarily on the same protein chain (**Figure 3-4b,c**, **Appendix S3-2**). We only analyzed proteins with 5-10 metal ions to characterize how metal ions were arranged when there is an "excess" of metal ions above the ~3-4 cooperative bond threshold discussed above. We found that for this set of 143 representative metal sites, the most likely number of consecutive bonds observed was three for the biological metal ions. This finding seems to support that in most of the cases (63% for biological metal ions) we analyzed, metal ions are organized with 1, 2 or 3 bonds consecutively. This preliminary analysis can be further expanded by analyzing proteins with 2-4 metal ions to understand the spatial distribution of metal ions. Further, we note that MetalPDB mostly includes proteins of catalytic functions, but not structural functions. A more

refined analysis beyond the scope of this study would include such structures once the protein structures have been resolved and would also include dimeric structures such as those analyzed in this investigation. Nonetheless, these biological proteins seem to confirm the existence of an optimum number of a metal-coordination bonds to maximize properties.

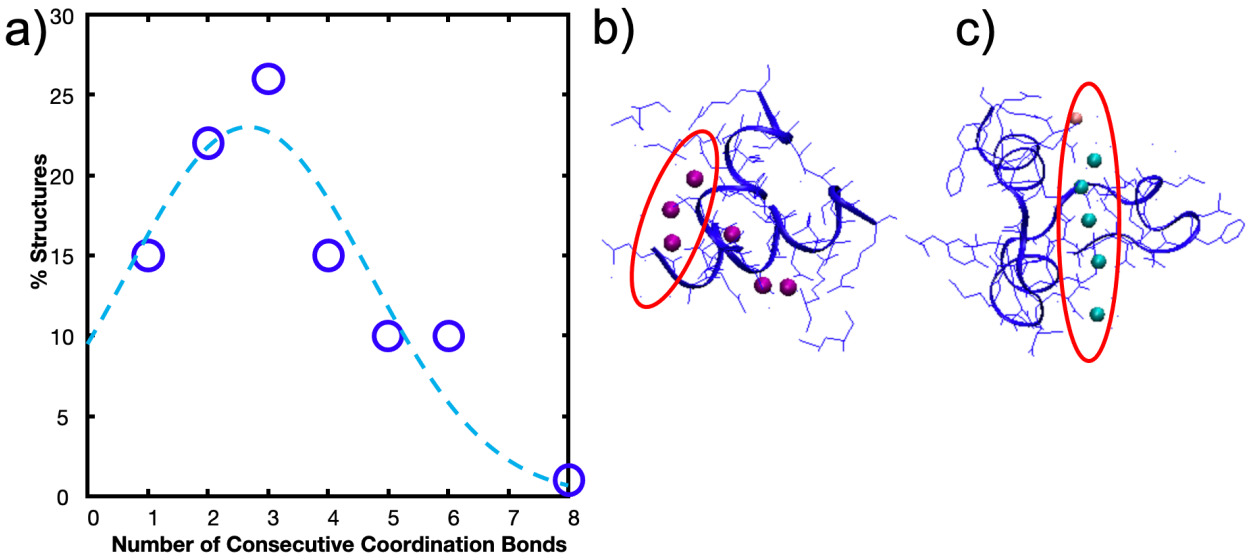


Figure 3-4. Biological metals from MetalPDB[287] follow trend where structures are more likely to have 1-3 metal ions consecutively arranged than 4-8. a) Percent of structures with x number of consecutive metal ions within the 143 structures analyzed. Gaussian fit with a peak at ~2.6 bonds is also drawn. b) Metal Site ID 2eul_6 with 3 of the 6 Zn^{2+} ions consecutively arranged. c) Metal Site ID 3th4_1 with 6 metal ions (5 Ca^{2+}, 1 Mg^{2+}) consecutively arranged.

In our simulations, this optimum number manifests itself as the maximum number of coordination bonds that rupture together. Throughout all peptide systems, only a maximum of three bonds is observed to rupture at once, even when more coordination bonds are present in the system (**Figure 3-3**). The shading of the circle (1, 2, or 3 bonds), is the maximum observed bonds that rupture together across all simulation repetitions in a specific peptide at a specific speed. These rupture events are highly heterogeneous, where depending on the initial conditions or velocities, different combinations of bond breaking clusters occur.

3.3.3 Rupture Force Heterogeneity and Location Dependence

To illustrate this heterogeneity in breaking events, **Figure 3-5** shows the heterogeneous breaking that independent simulations of the H3 peptide undergo at various pulling rates. In **Figure 3-5a,** for example, we observe that three bonds rupture together, but in **Figure 3-5c,** one bond ruptures at a time. Further, even for the same pulling speed, **Figure 3-5b** and **c** depict different breaking pathways, with two bonds breaking in **Figure 3-5b** or sequential rupture in **Figure 3-5c.** This surprising heterogeneity can be rationalized; the metal-coordination bonds here do not have the well-defined binding pocket or geometry that hydrogen bonds have in beta-sheets, enabling several mechanisms of rupture. Additional

contributions to heterogeneity may also arise from other polymorphic binding states, even though only one binding state in a bidentate trans configuration is explored here.

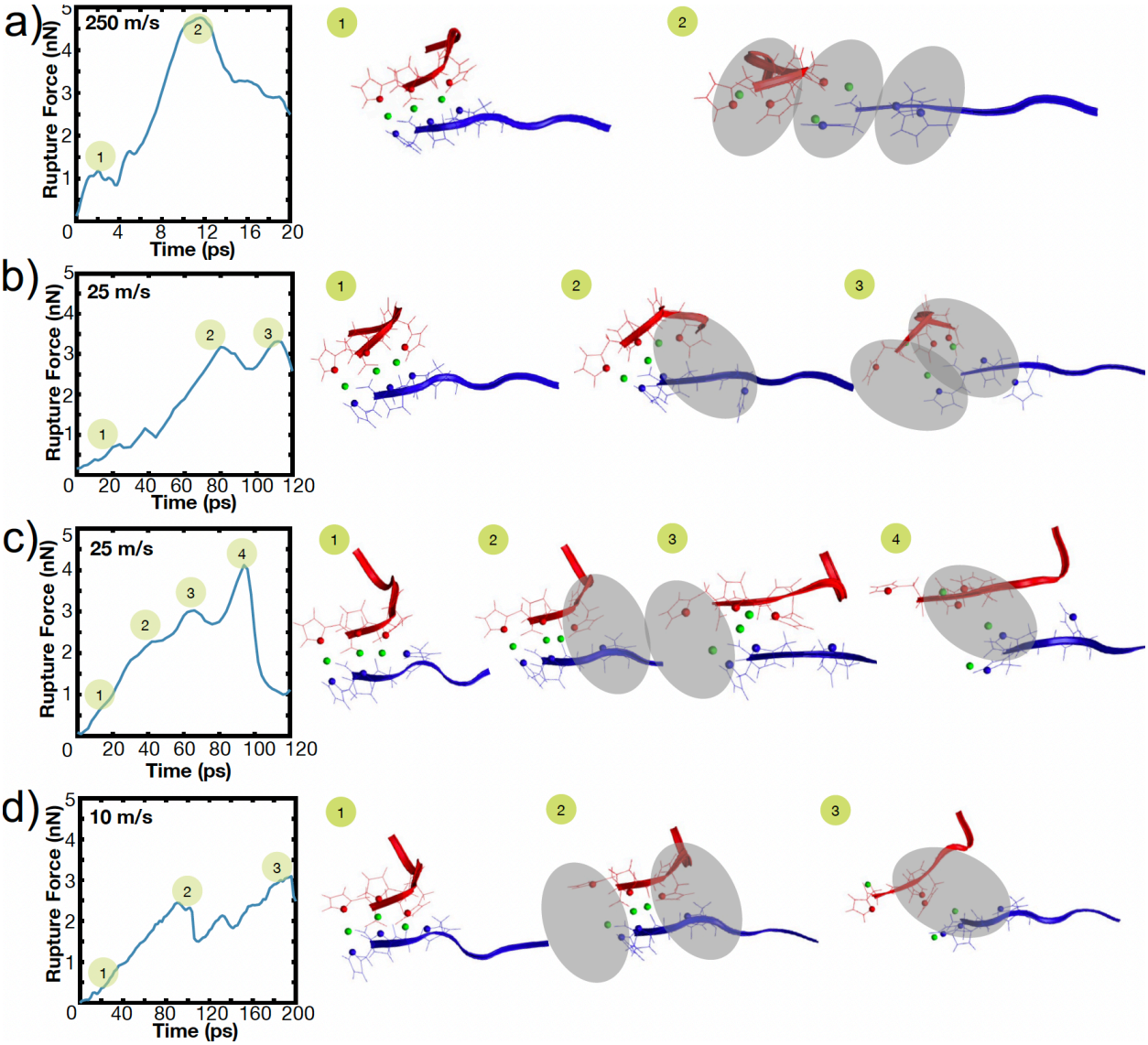


Figure 3-5. Representative deformation mechanisms of independent simulations of H3 breaking at different speeds show heterogeneous rupture behavior. Rupture force diagrams of H3 and the circles indicate the corresponding simulation snapshots with the coordination bond breaking highlighted in the gray oval. Various rupture pathways emerge, even at the same pulling speed (b,c).

The heterogenous breaking patterns exhibit breaking events where the coordination bonds break simultaneously when directly next to each other, spaced one bond apart, or spaced two bonds apart. To better understand how these different positions of bonds may affect the mechanical strength of the peptide, we conducted simulations in the H3 and HT3 peptides where the coordination bonds are placed in different positions (**Figure 3-6**). In almost every simulation (66% for H3, 77% for HT3), the bond closer to the SMD pulling atom ruptures first. However, the resulting rupture forces are only weakly dependent on the position of the bond. The “xMM” systems have moderately higher RF values (p value ~0.15 for H3, ~0.25 for HT3) compared to the other geometric arrangements because the two coordination bonds are

directly next to each other and away from the SMD pulling atom (**Figure 3-6a inset**). These two coordination bonds can directly influence each other's breaking, and as a result, strengthen the peptide system. Surprisingly, the "MxM" system with a coordination bond in the first and third site and the "MMx" system with two coordination bonds closer to the SMD pulling atom have similar RF values. For "MxM," the coordination bonds may be too far to influence each other without the presence of the coordination bond in the second site. For the "MMx" system, the force felt by the first coordination bond near the SMD pulling atom is likely directly felt by the second coordination bond in its proximity. The schematic in **Figure 3-6c** illustrates these molecular learnings. The amino acids that are more closely spaced together have a greater effect on each other's conformations. This mechanism, while at a geometrically larger scale, may have similar energetic underpinnings as the trans/gauche conformations in polymers that are caused by short-range steric interactions.[316]

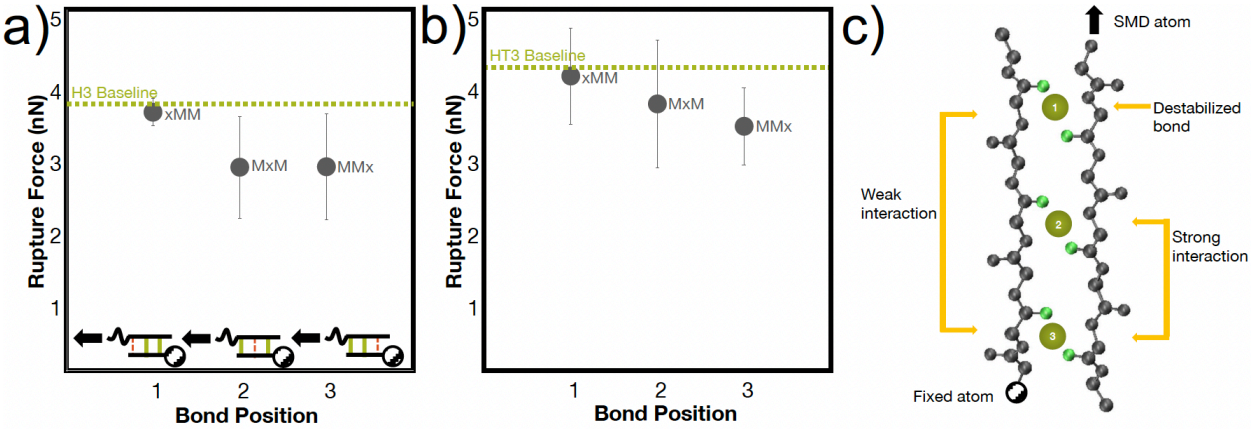


Figure 3-6. Peptide strength is weakly dependent on the location of the coordination bonds. a) H3 and b) HT3 rupture forces vary weakly based on the location of the two metal-coordinate bonds at a speed of 25 m/s (**Figure 3-6a**, inset). Using a two-tailed, unpaired t-test, the p values of "MxM" versus "MMx" are 0.99 for H3 and 0.63 for HT3. The p values for "xMM" versus the other geometric arrangements are ~0.15 for H3 and ~0.25 for HT3. c) The trends from the bond position-dependent rupture force behavior in **Figure 3-6a,b** are conceptualized in the schematic. The rupture force is the lowest when the metal-coordinate bond is closest to the atom that is pulled, likely because the highest localized force is experienced at this bond. The rupture force is the slightly higher when the two coordinate bonds are directly next to each other and away from the atom that is pulled.

3.3.4 Tuning rupture force through metal-coordinated peptide design

These mechanistic breaking pathways, the observation of an increase in RF followed by a plateau, and the comparison to biological metal-coordinated proteins suggest the presence of a critical number of metal-coordination bonds in these peptide systems. Given that this critical number for histidine-Ni^{2+} bonds generally follows the predictions from the N_{cr} in **Figure 3-1b**, we further parameterize the N_{cr} prediction to characterize how changing protein parameters can affect the cooperative strength or number of bonds that simultaneously rupture in other transient systems assuming the N_{cr} predictions from the Griffith fracture theory are also applicable. As in hydrogen bonded beta-sheets,[300] only a cluster of transient bonds break at once along a linear peptide, even if more bonds are present in the system (**Figure 3-7a**). This may be why several biological materials have evolved to feature an array of noncovalent interaction size effects,[1] enabling an optimization of properties with limited mass or weight.

Figure 3-1b shows that the spacing of the bonds along the backbone (L_{xo}) and energy of the bond (E) have the largest effects on the resulting Ncr. Decreasing the spacing helps the bonds more effectively mechanically "communicate" with each other, such that local steric interactions are directly affected by neighboring residues. Further, decreasing the energy of the bond allows the force to be distributed amongst the bonds of the backbone such that high stress or strain is not concentrated at the first bond that the force encounters. Increasing the τ, the characteristic time scale of bond rupture, also increases N_{cr} and various values are plotted in **Figure 3-7c.** This increase is likely due to the idea that increasing τ increases the residence time of the coordination bonds in an area, which would increase their likelihood of rupturing together. The N_{cr} can also be tuned by changing the persistence length (l_p), where an increased rigidity of the protein backbone, or increased l_p, causes a decrease in N_{cr} (**Figure 3-7b**). This may be because the energy required to compensate the change in a rigid backbone (higher l_p) upon additional metal-coordination binding penalty is too high, similar to the binding affinity decreases observed in Hebel et al.[294] Various vertical lines on **Figure 3-7b** show critical l_p values discussed in literature[317–321] to suggest possible backbones that could be used in further experimental or computational studies. Further, the effect of l_p is shown with varying xb, or the applied pulling distance at the moment of rupture.

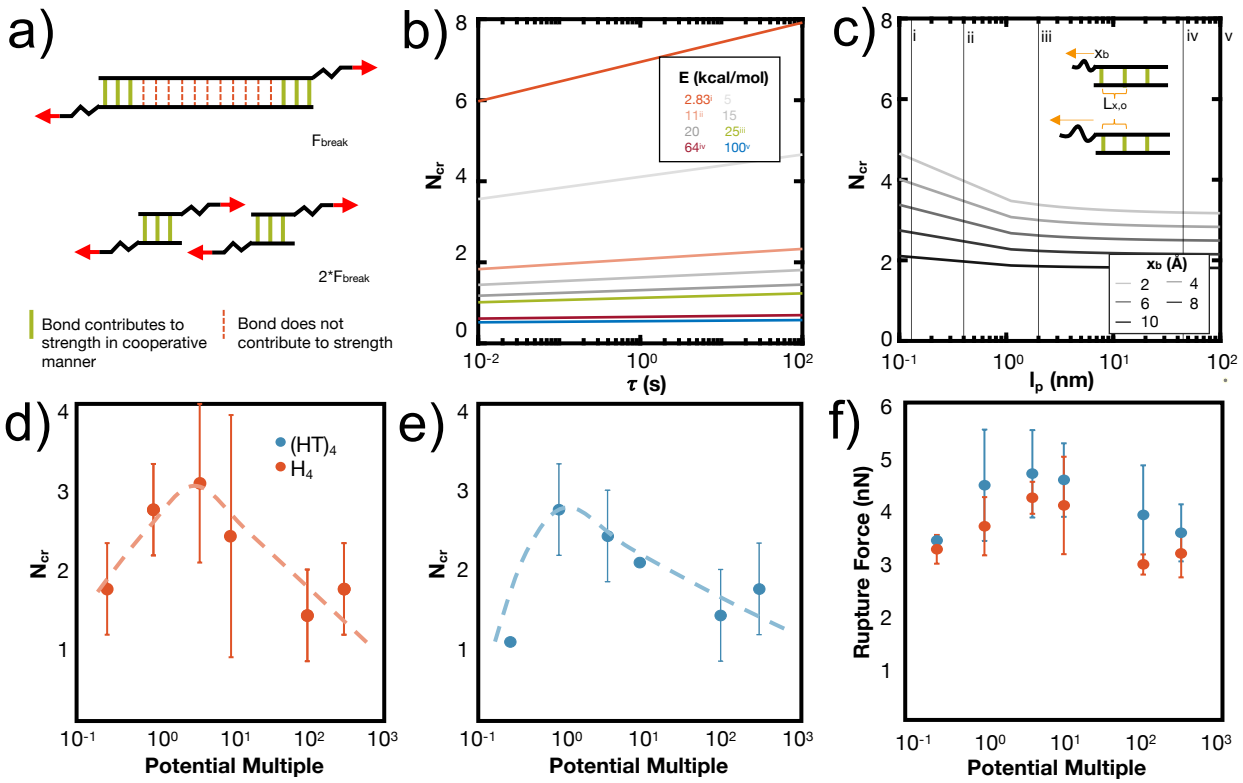


Figure 3-7. N_{cr} can be tuned by changing several design parameters with a major influence of energy of the bond (E) and L_xo, and a minor influence from l_p, x_b, τ. a) Adapted from [300], the schematic illustration depicts that only N_{cr} bonds break in a structure, even if more transient bonds are present. To increase the force at rupture, N_{cr} bonds should be placed in parallel. In **Figure 3-9b**, we show that N_{cr} is most sensitive to the values of E, the energy of the bond and L_{x0}, or the distance between bonds. Here we characterize other variables that affect the N_{cr} to show how design parameters such as polymer backbone or bond relaxation time can be tuned to change the N_{cr}. b) N_{cr} increases as a function of τ, the characteristic time scale of bond rupture. This dependence is plotted across multiple values of bond energy, which specific E values highlighted as: i) 2.83 kcal/mol for hydrogen bonds[300], ii) ~11 kcal/mol for the metal-coordination bonds in this study, iii) 25 kcal/mol for Zn^{2+}(N-methylacetamide)$_4$-(N-methylacetamide)[322], iv) 64 kcal/mol for HS-SH[323], and v) 100 kcal/mol for C-H[324]. c) N_{cr} as a function of persistence length (l_p) demonstrates that as the l_p increases, N_{cr} decreases. We plot this dependence for various x_b, or the

applied pulling distance at the moment of rupture and include the inset to show how $L_{x,o}$ and x_b are conceptually related. Vertical lines are drawn at critical l_p values discussed in literature at i) 0.13 nm for myosin fragments[317], ii) 0.4 nm for elastin-like polypeptides or other proteins[300,304,305], iii) 2 nm for a single disrupted amino acid chain[319], iv) 45 nm for 1000 base pair DNA[318], and v) 100 nm for an alpha helix or worm-like fibrils[320,321]. Changing the strength of the van der Waals potential of the Ni^{2+} metal ion, equivalent to changing the energy of the bond (E), decreases the N_{cr} and rupture force of H_4 (d) and $(HT)_4$ (e,f).

To demonstrate one aspect of this N_{cr} tunability, we changed the van der Waals interaction strength (epsilon) of the Ni^{2+} ions to directly measure how the energy of the bond (E) affected the N_{cr} of the proteins. We tested this on H_4 and $(HT)_4$, as these proteins have the most opportunities for cooperative rupture. Our simulations show that as expected, increasing E decreases the number of bonds that is likely to rupture together in both H_4 (**Figure 3-7d**) and $(HT)_4$ (**Figure 3-7e**). Interestingly, the trend of N_{cr} versus the strength of the interaction potential follows a pseudo-parabolic shape which was not predicted in the initial Griffith fracture theory. Our simulations showed that this is because the low interaction potentials are too weak to keep the structure together, resulting in natural dissociation of the bond. As such, fewer bonds remain to be pulled for cooperative rupture. These N_{cr} trends result in a similar parabolic shape in the rupture force (**Figure 3-7f**) because of the changing number of bonds that cooperatively rupture. The fact that the rupture force decreases even at high interaction potential, which would be expected to require large forces to rupture, shows that the rupture force is highly dependent on the number of bonds that rupture together. Together, these parameters suggest how *de novo* proteins may be designed with other protein backbones or dynamic noncovalent interactions to tune the cooperative rupture of bond clusters.

3.3.5 Discussions, Implications, & Conclusions

In this study, we sought to engineer the cooperative rupture of metal-coordination bonds to increase the strength of coordinated protein structures. Selecting *de novo* linear histidine-Ni^{2+} coordinated peptides as a model system, we have shown simultaneous rupture and how the rupture strength of the peptides can be tuned by adjusting the number of metal-coordination bonds in the protein. We designed idealized linear *de novo* Hx, HTx, and HNx peptide dimers in trans bidentate coordination geometries and showed that by increasing the amount of metal-coordination in the peptides, we could increase the rupture force of the peptides. We found that there was a strength limit in these peptides, at around three to four bonds for all three systems, where increasing the number of coordination bonds beyond this value did not contribute to increased strength of the system. To corroborate this finding, we used mechanistic insight into the metal-coordination breaking mechanisms found in simulation to show that only a maximum three bonds are observed to rupture in each system, even if there are more coordination bonds present in the system. These mechanisms revealed that a heterogenous breaking pattern emerged for the metal-coordinated peptides, where even under the same pulling conditions, different numbers or combinations bonds may break.

These findings indicate important design insights into the use of metal-coordination bonds in natural and synthetic systems. There is a balance between achieving high strength and optimizing the resources, such as the number of coordination bonds used, in a protein structure. Despite the fact that the cooperative rupture limit in our work is dependent on protein structure and microenvironment, similar principles have been found for hydrogen bonded structures, where the most effective use of hydrogen bonds towards mechanical

strength is when the bonds are clustered in groups of three to four to enable cooperative deformation.[300,325,326] Further, this optimization results in several design principles, some of which are demonstrated here through the Griffith fracture theory if applicable to the system of interest, to influence mechanical strength. As such, the mechanistic insights gained here have a much larger relevance to the rational design of metal-coordinated or dynamic noncovalent material mechanics.

Perhaps most importantly, we find that metal-coordination bonds show extremely heterogeneous breaking mechanisms and that designing well-defined binding pockets could force specific rupture pathways. Future studies could design alternative protein backbones to program these rupture pathways, such as alpha helices where cooperativity has been observed in hydrogen bonds.[327] Further, the simulations presented here were an idealized hydrogen-bonded beta-sheet-like configuration, and additional research should be tested on real metal-coordinated proteins systems once structures are characterized. Researchers can also apply methodologies similar to the one shown here to predict the cooperative rupture of other specific protein structures or bond chemistries. Such selectively engineered proteins can later be used to build hydrogels or polymer networks with high strength, toughness, and fast recovery for applications in recyclable polymers, self-healing polymers, artificial muscle actuators or electronic skin, as shown in early examples in literature with such bonds.[11,103,296] Additional improvements in metal-ion force fields can also help improve these predictions.[113]

Altogether, this work contributes clear and fundamental molecular design principles for utilizing multiple metal-coordination bonds for increasing the strength a metal-coordinated protein dimer. These principles help contextualize the structural role of metal ions both within the context of natural systems, as well as in bioinspired synthetic proteins and polymers. Broadly, the systematic understanding from this work contributes to the rational design of cooperativity in metal-coordinated proteins and polymers with mechanical function and expands insights into other dynamic noncovalent interactions.

3.3.6 Materials and Methods

The initial structures of the *de novo* peptide dimers (**Figure 3-1, Appendix S3-1**) in a parallel orientation were first predicted using AlphaFold v2.0.[328,329] Three additional amino acids (CGG) was attached on either side of the peptide to mimic the cysteine that is often used for immobilization in AFM-SMFS experiments and to ensure that the histidine-Ni^{2+} bonds were not directly being pulled. Ni^{2+} ions were added to the system, such that the histidine nitrogen atoms coordinate in a bidentate or tetradentate geometry to the Ni^{2+} ion. Simulations were implemented with Nanoscale Molecular Dynamics (NAMD) and all simulations utilized the CHARMM22 force field[229] with Ni^{2+} parameters from Babu et al.[231] and a 2 fs timestep. Histidine amino acids were modeled in the correct protonation state to match physiological pH and experiments where the histidine-Ni^{2+} dissociation time of τ=0.06 s[216] has been measured. The Ni^{2+} ions were balanced with Cl- ions for charge neutrality and the peptide dimer was solvated with TIP3P water molecules with a 15Å skin. Periodic boundary conditions are used with the Particle Mesh Ewald full system electrostatics method. After careful energy minimization using the conjugate gradient algorithm in NAMD, the simulations are equilibrated for 50 ns under NPT (1 atm, Nose-Hoover Langevin piston pressure control), followed by 50 ns in NVT (Langevin dynamics).

Independent simulations under this procedure were carried out for subsequent SMD tests. The TIP3P water molecule box was extended by 60Å to account for deformation in the pulling direction. The C_α atom on the N-terminus near the cysteine residue was selected as the SMD pulling atom, and the C_α atom on the opposite strand C-terminus was selected as the fixed atom. The structure was energy minimized for 10000 steps. Then SMD data were collected every 1 ps with pulling rates from 0.25 m/s to 250 m/s under an NVT ensemble and 2 fs timestep. The simulation was run until the parallel proteins were fully separated and all intermolecular metal-coordination bonds fully ruptured. The rupture of a metal-coordinate bond was defined as a distance of 3Å or more between the coordinating nitrogen of histidine and the Ni^{2+} ion. The time, distance, force, and type of bond rupture was recorded for each rupture event in each simulation. Simultaneous rupture was defined as multiple bond breaking events within 1 ps of each other in the visual file with the rupture force peaks that were indistinguishable from each other.

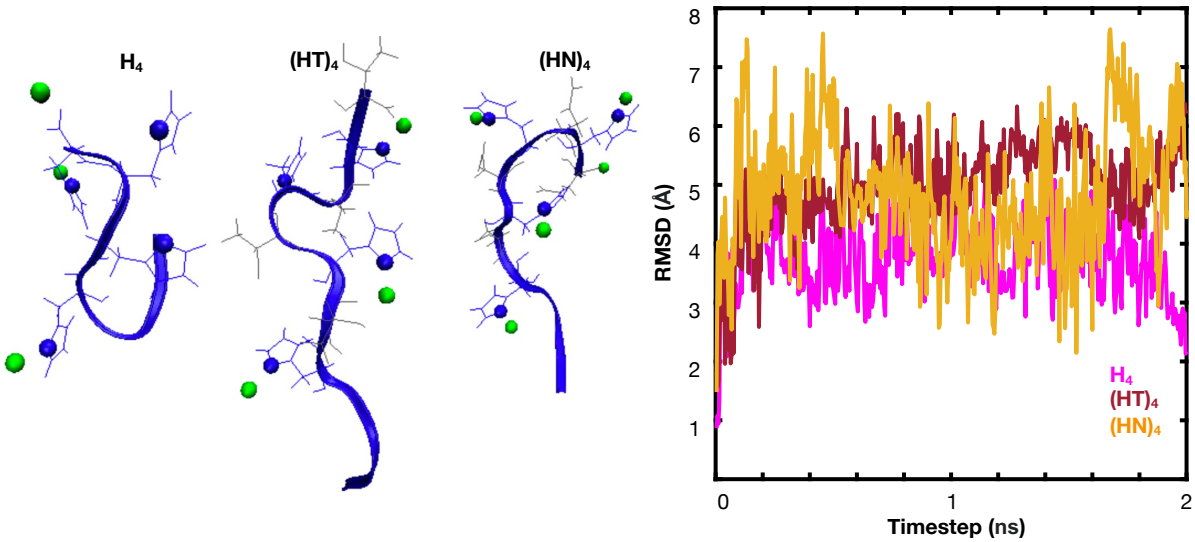


Figure 3-8. Single peptides equilibrated with metal ions. Single peptides (H4, HT4, HN4) which are the longest in their series and have the most likely chance of internal coordination, are equilibrated with explicit water molecules (not shown) for 50 ns NVT and 50 ns NPT. The proteins reach a stable root mean square deviation value but do not show any internal coordination.

We found that tetradentate structures were not stable during the initial equilibration, and the metal-coordination binding sites would quickly dissociate into tridentate, bidentate, or monodentate structures during the equilibration process. As a result, we continued SMD tests with only bidentate coordination. While this lack of tetradentate stability may be due to challenges with the force field modeling for metal ions, speciation models predict a dominance of bidentate coordination stoichiometry in aqueous conditions between histidine and Ni^{2+}.[216,330,331] Further, though the short peptides are equilibrated for a significant amount of time, we note that the peptide structures used in this paper are not necessarily the equilibrium binding state of the peptide with the metal ion, because the peptide may be trapped in a local minima. We also enforce a trans bidentate binding configurations on the metal ions to most closely replicate the hydrogen-bonded beta-sheets. Despite these limitations, we proceed with our simulations because the goal of this paper is to determine whether cooperativity can exist in these bonds under “ideal” conditions, where “ideal” conditions represent the closest geometry to hydrogen bonds where cooperativity has been seen in beta-sheets with trans

coordination along the backbone. Further, preliminary simulations of a single peptide (H_4, $(HT)_4$, and $(HN)_4$) show that internal metal-coordination bonds are unlikely (**Figure 3-8**), likely due to an entropic penalty of loop formation.[332]

3.4 Molecular design of metal-coordinated proteins for heterogeneous and cooperative rupture

In the previous section, we presented an exploration of metal-coordination cooperativity and rupture in *de novo* proteins. In this section, we validate these principles using an experimental system and gain mechanistic insight using corresponding MD simulations. We rationally design a random coil protein template to uncover the heterogeneous and cooperative nature of increasing numbers of intermolecular His-Ni^{2+} bond clusters (**Figure 3-9**). Using isothermal titration calorimetry, atomic force microscope (AFM)-based single-molecule force spectroscopy (SMFS), and steered molecular dynamics (SMD) simulations, we show that the polypeptides dimerized with three histidine residues display surprisingly more-than-additive rupture forces, and we resolve at least two dissociation pathways: the sequential rupture or simultaneous cooperative rupture of two-to-three coordination bonds. Further, the simulations reveal mechanistic insight into several heterogeneous rupture modes of the coordination bond. We posit that biological materials may have incorporated metal-coordination bonds for primarily energy dissipative reasons due to their heterogeneous and cooperative nature. The methodology and insights developed here have important implications for understanding the molecular mechanisms of metal-coordination bond-based stabilization of proteins and the *a priori* design of metal-coordinated materials with desired mechanical properties.

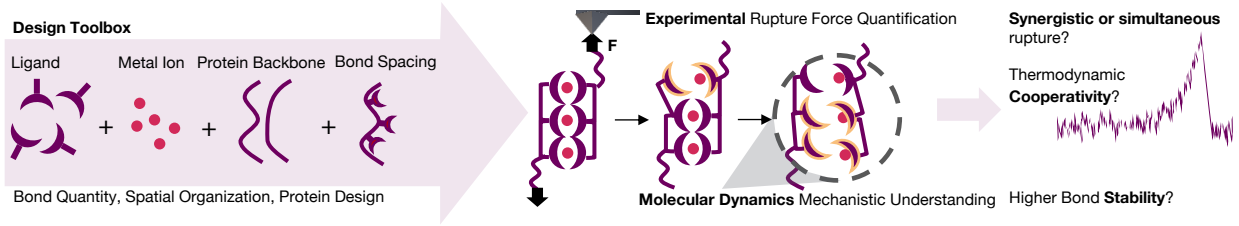


Figure 3-9. Schematic of design contributions and key questions explored this work. The goal of this work is to uncover how multiple metal-coordination bonds can be designed to yield new materials with high rupture forces through higher bond stability, cooperativity, and simultaneous rupture. Experiment and simulation are used to probe the underlying physical phenomena of the metal-coordination bond rupture. The methodology and insights developed here have important implications for understanding the molecular mechanisms of metal-coordination bond-based stabilization of proteins and the *a priori* design of new metal-coordinated materials with desired mechanical properties.

3.4.1 Metal-coordinated polypeptide design for cooperative rupture

Polypeptides with an increasing number of metal-coordinating amino acids were rationally designed to probe how to induce the cooperative rupture of metal-coordination bonds (**Figure 3-10a**). Histidine-Ni^{2+} was selected as the coordination bond, because it is well characterized in biological and non-biological contexts, such as the widely used his tag for protein purification.[51,290,298] An elastin-like polypeptide (ELP) was used as a disordered polypeptide

backbone to provide increased flexibility for the metal-coordination bonds to assemble freely and enable potential cooperative structures as compared to the constrained polyproline II helices used by Sun et al.[103] Using an ELP template also provides a known structure for complementary simulation efforts.

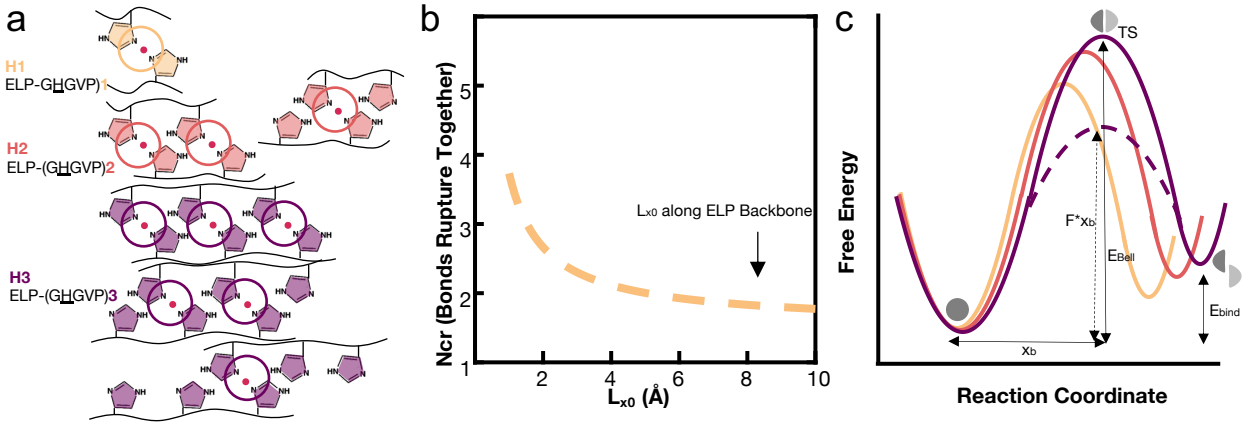


Figure 3-10. Experimental design of metal-coordinated ELPs to probe cooperative binding. a) Three ELPs (ELP-(GHGVP)$_{1\text{-}3}$) are designed with one to three histidine residues (H1 to H3) that define the maximum number of metal-coordination bonds possible. When the number of histidine residues increases, more intermolecularly crosslinked configurations become possible. b) The Griffith nanoscale rupture theory is applied to predict the number of metal-coordination bonds that rupture together (N_{cr}), given different energies required to rupture a single bond (E_{Bell}) or spacing (L_{x0}) of the coordination bonds along the backbone. The following equation is plotted with the relevant parameters listed in the text: $N_{cr} = \frac{k_B T}{E_{Bell}}\left[\frac{x_B}{4l_p}[(1-\alpha_{cr})^{-2} + 4\alpha_{cr} - 1] - ln\left(\frac{1}{\omega\tau}\right)\right]$ where $\frac{E_{Bell}}{L_{xo}} = \frac{k_B T}{4l_p}(\alpha_{cr}(1-\alpha_{cr})^{-2} - (1-\alpha_{cr})^{-1} + 2\alpha_{cr}^2 - 1$. c) Depending on the number of coordination bonds formed, the underlying energy landscape is altered. The energy barrier (E_{Bell}) and the distance between the bound and transition state (x_B), as well as the binding energy (E_{Bind}), is predicted to increase from H1 to H2 to H3. An applied force (F) lowers the transition state (TS) to $F * x_B$, allowing the extraction of energy landscape parameters.

To inform the number of metal-coordination bonds that should be placed along the peptide backbone, we applied an earlier model[300,302,303,333,334] that predicts the cooperativity of hydrogen bonds to our histidine-Ni^{2+} peptides to yield the number of metal-coordination bonds that are likely to rupture together (N_{cr}). The model is rooted in the application of energetic arguments, originally proposed in the Griffith rupture theory,[301] which shows that the stress at rupture is proportional to the internal strength of the chemical bonds of the material.

The following considerations are used to determine an initial estimate of the N_{cr} of histidine-Ni^{2+} bonds. A bond dissociation time scale τ (1/k_{off}) of 0.06 s is used,[74,216] representing the dissociation of histidine-Ni^{2+} in water. The applied pulling distance at rupture x_B is 3 Å, representing the maximum coordination bond distance. The ELP persistence length l_p is 0.4 nm.[304,305] The energy to rupture a single bond (E_{Bell}) of 11 kcal mol^{-1} is based on the SMD of a single (histidine)$_2$-Ni^{2+} complex.[333] Despite the number of parameters that tune N_{cr}, the N_{cr} predictions are most sensitive to E_{Bell} and the spacing of the coordination bonds along the backbone (L_{x0}).[333] These inputs result in an N_{cr} prediction of 2-3 bonds (**Figure 3-10b**), given an L_{x0} of 8-10 Å along the guest residues of the ELP backbone.

Based on the N_{cr} prediction, ELPs with a total of 34 to 36 GXGVP repeats are designed for experimental analysis, so as to have either one (H1), two (H2), or three (H3) histidine residues in the guest positions (X) located at the N-terminus (**Appendix Method 3-S1**). The metal-coordinating repeats can dimerize via a maximum of one to three coordination bonds (MC-ELP,

Figure 3-10a). Our goal is to induce cooperativity in the thermodynamic, kinetic and/or dynamic mechanical stability of the metal-coordination bonds that dimerize these polypeptides. Cooperativity is defined as the simultaneous dissociation or rupture of bonds that manifests in an apparent bond stability that is more than additive. This cooperative rupture should originate from differences in the underlying energy landscape caused by the additional metal-coordination bonds. **Figure 3-10c** schematically illustrates this concept through the phenomenological Bell-Evans Model,[335,336] which is conventionally applied in SMFS experiment to describe force-induced dissociation. We expect that H3 should have the highest energy barrier (E_{Bell}, or consequently lowest k_{off}) and widest potential (x_B), while H2 and then H1 should have lower values. Further, we expect H3 to have a higher thermodynamic stability (E_{Bind}) because of the presence of additional coordination bonds.

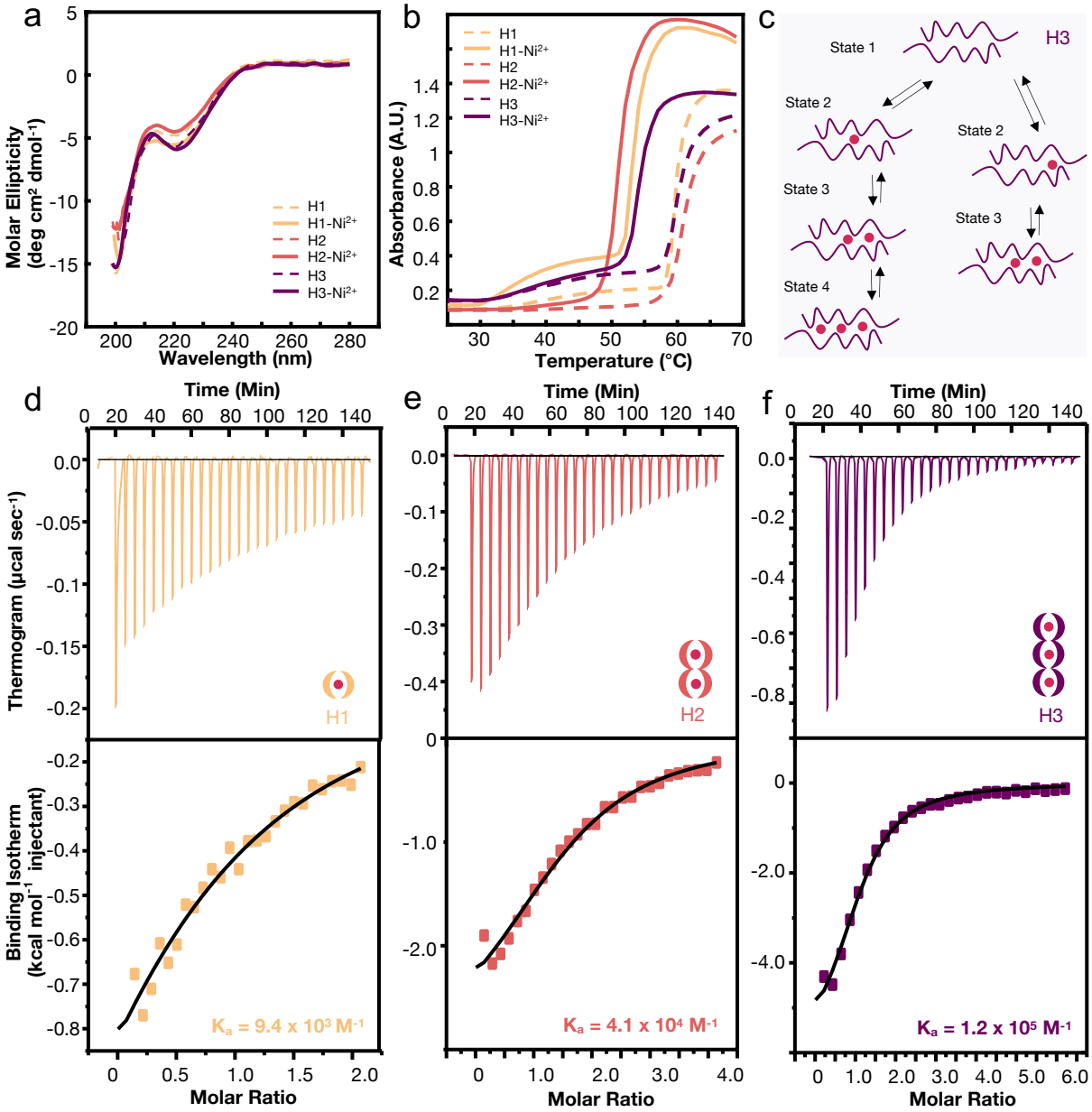


Figure 3-11. Structural and thermodynamic properties of the metal-coordinating ELPs. a) CD spectra show that there is no change in the conformation of the ELPs upon the addition of Ni^{2+}, indicating that the ELPs retain

their intrinsic flexibility. b) UV-Vis spectroscopy indicates that the transition temperature (T_t), even upon the addition $NiCl_2$, is above 25ºC. As such, the ELPs remain soluble at room temperature, ensuring that individual molecules are probed in SMFS experiments. c) Multiple cooperative binding mechanisms of H2 and H3 are possible, where the binding of the first metal ion affects the binding of surrounding metal ions. This is expected to yield a thermodynamically more stable conformation, i.e. a higher association constant. ITC of d) H1, e) H2, and f) H3 shows increasing association constants from H1 to H2 to H3, indicating increasingly stronger binding with Ni^{2+}. A one-site model is demonstrated for ease of comparison. All experimental data is collected in the same measurement buffer used for SMFS, i.e. non-coordinating 10 mM HEPES/ NaOH pH 7.4, 140 mM NaCl.

The rationally designed MC-ELPs were recombinantly produced in *E. coli* and purified using the inverse transition temperature cycling method (**Figure S3-1**). They were structurally characterized in the absence and presence of Ni^{2+} to confirm their disordered nature. Circular dichroism (CD) spectroscopy confirms that the ELPs retain a disordered conformation in both conditions (**Figure 3-11a**). The large negative band at 200 nm and a small one at 220 nm match the spectra of both other ELPs[337] and the random coil secondary structure.[338] UV-Vis spectroscopy demonstrates that all ELPs remain soluble at room temperature (**Figure 3-11b**). These two measurements confirm the flexible, soluble polypeptide structure of the MC-ELPs and, consequently, ensure that the thermodynamic, kinetic, and dynamic mechanical properties represent metal-coordination bonds between individual MC-ELP dimers and not larger clusters.

As a first test to probe the possible cooperative binding of Ni^{2+} ions to the MC-ELPs templates, ITC experiments were performed. The additional coordination bonds increase the association constant K_a by a factor of ~3-4 when moving from H1 to H2 to H3 (**Figure 3-11d-f**), given a one-site model to capture the apparent single-term binding constant of the three polypeptides for ease of comparison. This indicates that the increase in binding affinity is slightly more than additive when increasing the number of metal-coordination sites on the ELP template. As several binding modes and dissociation pathways are possible (**Figure 3-11c**), alternative ITC fits are presented in **Table S3-1.**[339] The two-sites binding model indicates that the second site of H2 and H3 has a higher K_a than the first coordination site, providing first evidence for the cooperativity of binding sites.

Table 3-1. Alternative ITC fits show possible cooperativity based on fit type. Because the exact binding mechanism is unknown, alternative ITC fits for H2 and H3 are shown. Several binding models fit the data well, although moderately better fits are obtained when the two-sites model is used for H2 and H3 instead of the one-site model. The two-sites binding model yields independent binding constants (K) and we refer readers to the manual for the comprehensive equations.[340] The two-sites binding model indicates that the second site has a higher binding constant than the first site. Using a model for independent sites, it is extremely difficult to obtain a unique fit for more than two sites, and as such, the fit with the three independent sites model is not shown.[340] The thermodynamics of metal-coordination bonds have also been probed in other studies.[289–291,341,342] The binding constants in our experiments are ~1 order of magnitude less than those observed through surface plasmon resonance experiments,[298] likely due to the different experimental method. At the same time, they are ~1-2 orders of magnitude higher than those measured for histidine-Zn^{2+} coordination.[103] This is in line with the higher stability of the histidine-Ni^{2+} bond.[111] Significant literature on multivalent binding can also be found in other protein systems for comparison.[292–294]

MC-ELP	K1 (M^{-1})	K2 (M^{-1})	K3 (M^{-1})	ITC Fitting Model
H1	9.39E+03			One-Site
H2	4.11E+04			One-Site
H2	2.89E+04	4.24E+05		Two-Sites
H3	1.25E+05			One-Site
H3	8.81E+04	2.10E+05		Two-Sites

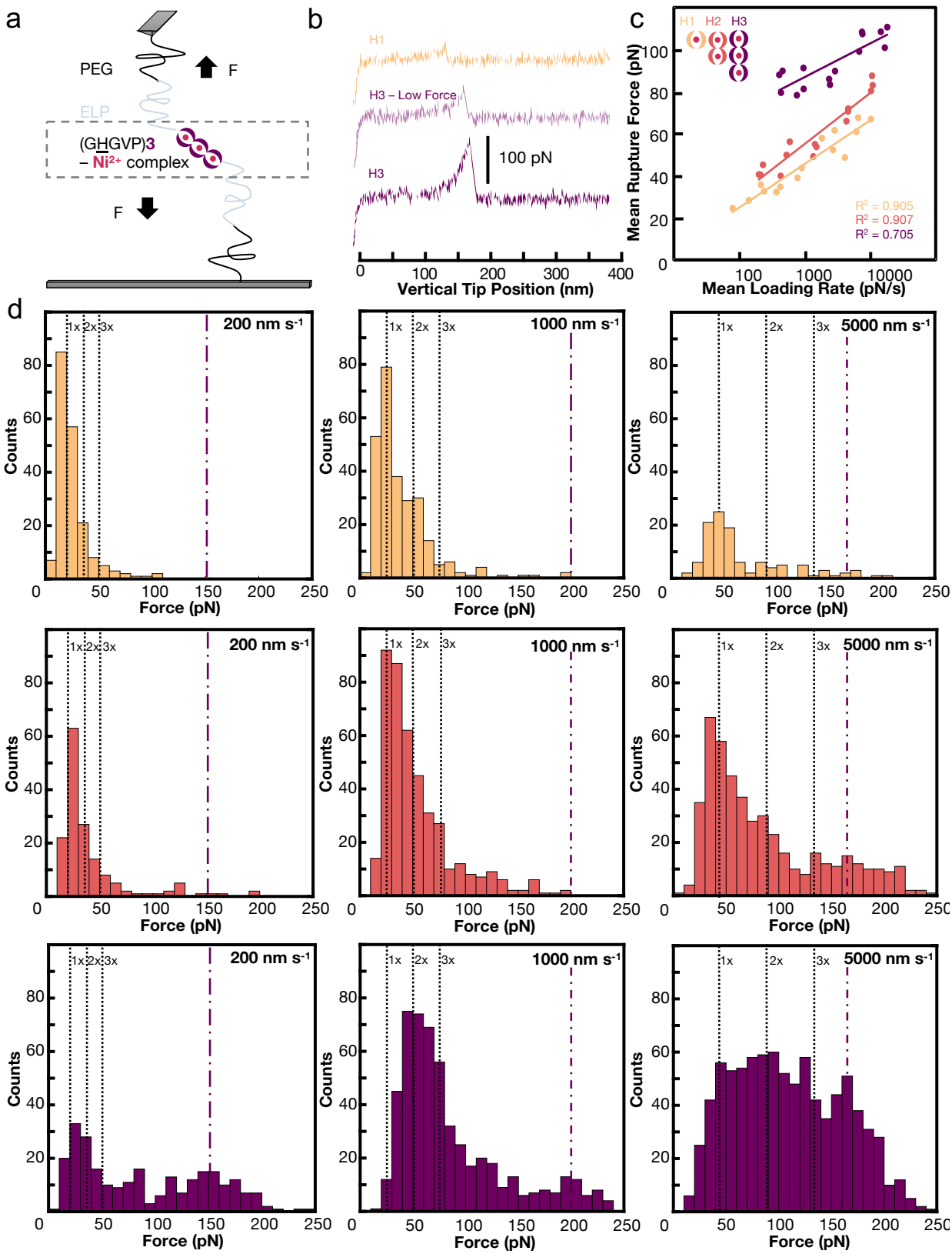


Figure 3-12. SMFS of metal-coordinating ELPs. a) AFM-based SMFS setup. ELPs carry one to three N-terminal histidine residues and a C-terminal cysteine for site-specific immobilization. The cysteine is used to covalently couple the ELPs (180-190 amino acids) to the cantilever and a glass surface, using a 10 kDa (approx. 60-65 nm

long) polyethylene glycol (PEG) linker. b) Representative force-extension curves showing the rupture of H1 (blue), H3 (pink), and H3 with a high rupture force (red). c) Mean loading rate dependence of the mean rupture forces. The Bell-Evans model fit is also shown. d) Representative rupture force histograms obtained at retract speeds of 200 nm s^{-1}, 1000 nm s^{-1}, and 5000 nm s^{-1}. The vertical lines at 1x represent the most probable rupture force of H1 (i.e. one metal coordination bond). The lines for 2x H1, and 3x H1 indicate where the rupture forces for H2 and H3 would be expected if the bond strength would be additive. The red line highlights the high rupture force peak for H3. All experimental data is collected in non-coordinating 10 mM HEPES/ HCl pH 7.4, 140 mM NaCl.

3.4.2 Dual modes of ELP-H_n polypeptide rupture

As ITC averages over all binding configurations and dissociation pathways, the cooperative rupture of the MC-ELP dimers was subsequently investigated with SMFS where such heterogeneities can be resolved (**Figure 3-12a**). The MC-ELPs, which have a C-terminal cysteine, were covalently coupled to cantilevers and glass surfaces, using a maleimide-functionalized polyethylene glycol (PEG) linker (M_w = 10,000 g mol^{-1}). The PEG layer effectively passivates the surface to reduce nonspecific interactions (**Figure 3-13**).[315,343–346] The cantilever approach facilitates ((histidine)$_2$-Ni^{2+})$_n$ complex formation, while retraction results in its complex rupture. **Figure 3-12b** shows representative force-extension curves, each with a clear single-rupture peak. Even though ELP and PEG chains are connected in series, force-extension curves displaying a single-rupture event can be described with the extensible freely jointed chain model as their Kuhn lengths are very similar (ELP: 0.8 nm;[304,305] PEG: 0.7-0.8 nm[347,348]). Considering the combined length of ELP (65-70 nm) and PEG (60-65 nm)[348], only force-extension curves with a contour length >100 nm were used for further analysis.

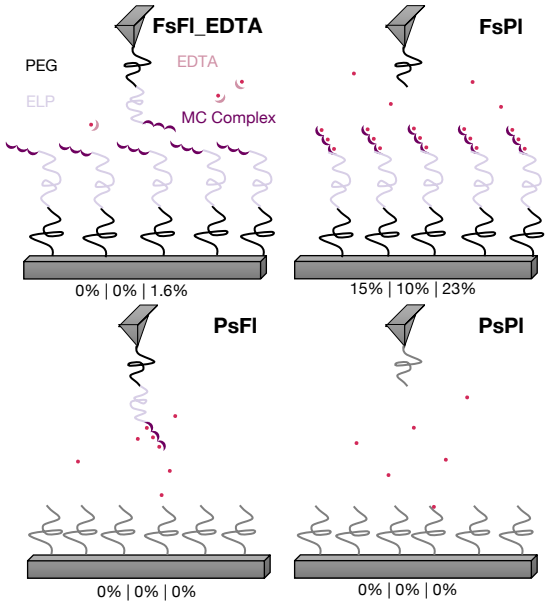


Figure 3-13. SMFS control experiment to quantify non-specific interactions. A series of control experiments was conducted to determine non-specific, metal ion mediated interactions between the maleimide-functionalized PEG and ELP linkers (or any other two components) in the experimental setup. Using H3, two different types of control experiments were performed. The first experiment includes a fully functionalized coverslip and cantilever ("FsFl_EDTA"), where the aminopolycarboxylic acid ethylenediaminetetraacetic acid (EDTA) was added as a chelator during the SMFS experiment. For the second experiment, H3 was omitted from either the coverslip ("PsFl") or the cantilever ("FsPl"), or from both ("PsPl"). Very little interaction is observed in these experiments with exception of the experiment were the H3-functionalized slip was probed against the PEG-only cantilever ("FsPl"). This binding is, however, absent when EDTA is added to the experiment ("FsFl_EDTA"). We hypothesize that this nonspecific binding may arise from the Ni^{2+}-mediated attachment of the ELP to the cantilever. Data values below each subfigure

represent three independent experiments calculating the percent binding in the panel compared to the percent binding observed in a fully functionalized, standard experiment. Raw data for percent binding is presented in **Table S3-1.** In conclusion, this suggests that non-specific binding events constitute a maximum of ~15% of all analyzed force-extension curves, e.g. in the H3 data shown in **Figure 3-12**.

Force-extension curves were measured across several retract speeds, and mean rupture forces increase with increasing mean loading rates (**Figure 3-12c**). H1 demonstrates a narrow force distribution across all retract speeds (**Figure 3-12d**), consistent with the rupture of one metal-coordination bond. The rupture force histograms for H2 and H3 are significantly broader. This again reflects the larger number of possible bound configurations, such as parallel or antiparallel alignment of the polypeptide chains, staggered configurations, and variations in the number of Ni^{2+} ions bound (**Figure 3-10a, 3-10c**). It is notable that the heterogeneity is present despite the use of a site-specific immobilization protocol, minimized nonspecific interactions (**Figure S3**), and use of non-coordinating buffers, which were limitations of previous SMFS experiments with similar systems.[315,344–346] This rupture force histogram heterogeneity is reproduced across multiple independent experiments, indicating that deeper physical principles underlie the heterogeneous distribution. It should be noted that a single narrow peak is also observed when H3 on the cantilever is probed with a H1-functionalized surface. This ultimately confirms that single-rupture events are recorded and that aggregation of coordination bonds is absent in our SMFS experiments (**Figure 3-14**).

The Bell-Evans model[335,336] fitting (**Figure 3-12c**) estimates of the mean rupture force data yields quantitative energy landscape parameters to contextualize the differences between H1, H2, and H3. The parameters for H1 can be accurately determined and are an x_B of 4.5 Å and k_{off} of 7.03 x 10^{-1} s^{-1}. For comparison, for the binding of an analogous system his tag-nickel NTA, Kienberger et. al previously reported an x_b of 1.9 Å and a τ = 15 s (k_{off} 6 x 10^{-2} s^{-1})[315] and Friddle et. al[349] reported an x_B of 0.9 Å and k_{off} of 4.3 s^{-1}. Though the exact quantitative determination of parameters for H2 and H3 is not possible due to their non-Gaussian histograms, Bell-Evans fitting estimates are reported to illustrate the differences between the systems. H2 has an x_B of 3.9 Å and k_{off} of 4.99 x 10^{-1} s^{-1}, and H3 has an x_B of 5.9 Å and k_{off} of 5.29 x 10^{-4} s^{-1}. H2 has similar landscape parameters to H1, but H3 has a k_{off} that is three orders of magnitude smaller than H1, reflecting a higher energy barrier and the lack of deformability of the protein.[350] The H3 energy landscape parameters are consistent with the predicted changes to the energy landscape in **Figure 3-10c.**

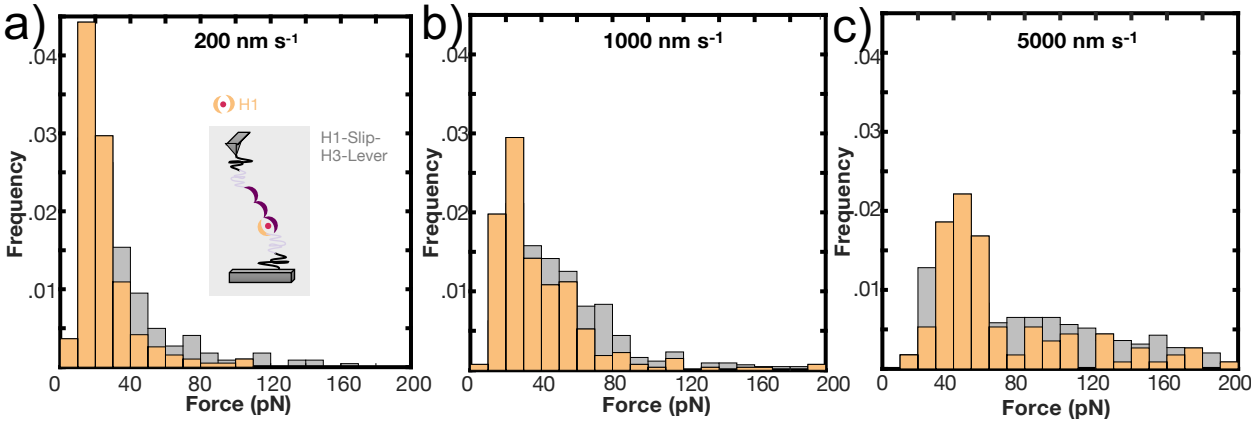


Figure 3-14. SMFS of composite H1-H3 system. Experiments with H1 on the coverslip and H3 on the cantilever (inset, gray histograms) demonstrate similar behavior to H1-only (blue histograms), indicating that each H3 polypeptide on the cantilever only binds to one H1 molecule on the surface. This confirms that the experimental setup is at a low concentration to pull on a single molecule rather than cluster of molecules.

To better understand cooperative effects, we focus on the heterogeneous distributions of H2 and H3. While the rupture force histograms of H1 and H2 are similar, clear differences are seen in the histograms of H3. These histograms show one peak with similar rupture forces as seen for H1 and H2 and a peak with much higher rupture forces (**Figure 3-12d**). Especially at the lower retract speeds of 200 and 1000 nm s^{-1}, the two peaks are clearly separated while they move closer at the highest retract speed of 5000 nm s^{-1}, indicating that the two peaks display a different loading rate dependence. At the slowest retract speed, the first force peak of H1, H2, and H3 is located at ~15 pN for all three MC-ELPs. This suggests that the metal coordination bonds break sequentially at slow retract speeds, such that H2 and H3 effectively behave like H1. As the retract speed increases, the first peak rupture force for H2 remains the same as H1. However, the first peak rupture for H3 deviates from H1. Further, the red dotted line illustrates that another peak for H3 is located at ~150-200 pN (**Figure 3-12d**). The additional peak observed for H3 is clearly higher than 2x or 3x the rupture force observed for H1 (black dotted lines in **Figure 3-12d**). These more-than-additive rupture forces indicate synergistic contributions of the additional coordination bonds. These high forces are surprisingly on the order of titin unfolding[351] and indicate that appropriately spaced coordination bonds can increase the stability of proteins in a cooperative manner.

3.4.3 Uncovering the heterogeneity of metal-coordination rupture

To obtain insights into possible molecular mechanisms of metal-coordination bond rupture, we conducted SMD simulations for shortened analogs of the MC-ELPs. Although the exact binding geometry of the polypeptides with Ni^{2+} ions are unknown, speciation calculations of histidine with Ni^{2+} predict a bidentate coordination stoichiometry in aqueous conditions.[330,331] Further, despite the flexibility of ELPs, intramolecular coordination with higher denticity is unlikely due to an entropic penalty of loop formation,[332] given that the spacing of histidine residues on the ELP backbone is only five amino acids (~1.8 nm). Given these assumptions, we model an “idealized” binding scenario where a bidentate trans configuration is imposed. This idealized coordination geometry is equilibrated for a long time to confirm its stability prior to SMD pulling studies.

For all SMD simulations of MC-ELPs, the rupture forces increase with the pulling speed (**Figure 5a**) with a similar trend as in experiment (**Figure 3-12c**). Just as in the SMFS experiments (**Figure 3-12c**), the difference between H2 and H3 is larger than the difference between H1 and H2. This likely originates from a higher probability of simultaneous coordination bond rupture for H3.

To understand the molecular origins of this behavior, rupture events were separated into pathways where one, two or three bonds ruptured simultaneously (**Figure 3-15b**). SMD provides mechanistic insights into these rupture pathways, which are otherwise inaccessible in the experimental force-extension curves. Considering experimental timescales and thermal fluctuations, the distances between the coordination sites are too short (five amino acids, ~1.8 nm) to distinguish sequential versus simultaneous rupture (**Figure 3-12b**), but are required for cooperative rupture (**Figure 3-10b**).

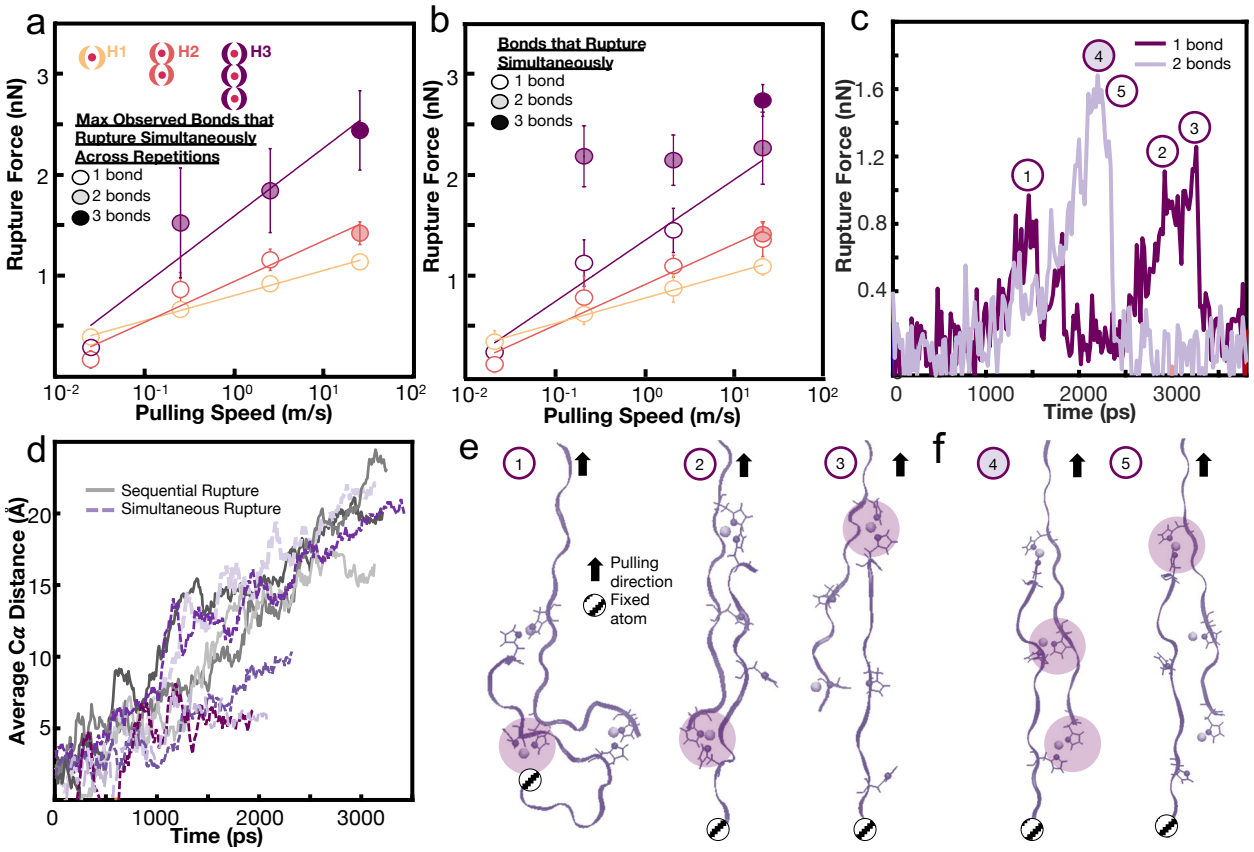


Figure 3-15. SMD of metal-coordinating ELPs. a) Pulling speed dependence averaged over all rupture events. The final rupture force follows a positive linear relationship with pulling speed as expected for all MC-ELPs. H3 has a higher rupture force than H2, followed by H1. The colored data points indicate the maximum number of metal coordination bonds that rupture simultaneously across all repeats. b) Pulling speed dependence is disaggregated by the number of metal-coordination bonds that rupture simultaneously. The cooperative rupture of two bonds has a higher rupture force across all pulling speeds. c) Several rupture pathways are possible and examples of simulation trajectories of H3, pulled at 2.5 m s^{-1}. The comparison of the two examples shows that the simultaneous rupture of two bonds requires more force than what is needed for breaking one bond. d) SMD experiments using different initial structures or randomly initialized atomic velocities. These simulations show a range of deformation pathways, as illustrated by the average distance between the initial configuration and the configuration at a given time of the backbone C_α atoms of the protein chain with the fixed atom. e,f) Simulation snapshots of H3 pulled at 2.5 m s^{-1}. The snapshots illustrate this heterogeneity, where bond breaks sequentially (e) or two bonds break simultaneously (f).

The SMD data corroborate that the low-force peak in the SMFS rupture force histograms is likely due to the rupture of a single coordination bond. When comparing the rupture of single coordination bonds between simulation and experiment, a similar speed-dependent trend is observed (**Figure 3-15b** and **Figure 3-16a**). The simulations further confirm our initial interpretation that the sequential rupture of coordination bonds is more likely at slower pulling speeds. The coordination bonds may relax in response to backbone fluctuations, which have a larger contribution during longer simulation times at slower pulling speeds (**Figure 3-17**). Interestingly, the speed-dependence of H1, H2, and H3 in both **Figure 3-15b** and **Figure 3-16a** does not completely overlap even though only one bond is rupturing. This indicates that the neighboring binding sites affect each other even though not all of them have a metal bound. This observation is also captured in the energy landscapes of only one bond rupturing in H1, H2, and H3 (**Figure 3-16b**). The high-force peak in the SMFS histograms of H3 is thus likely due to the cooperative rupture of multiple coordination bonds. Likely, the cooperative rupture of

two bonds significantly changes the underlying energy landscape to result in the synergistic high rupture forces.

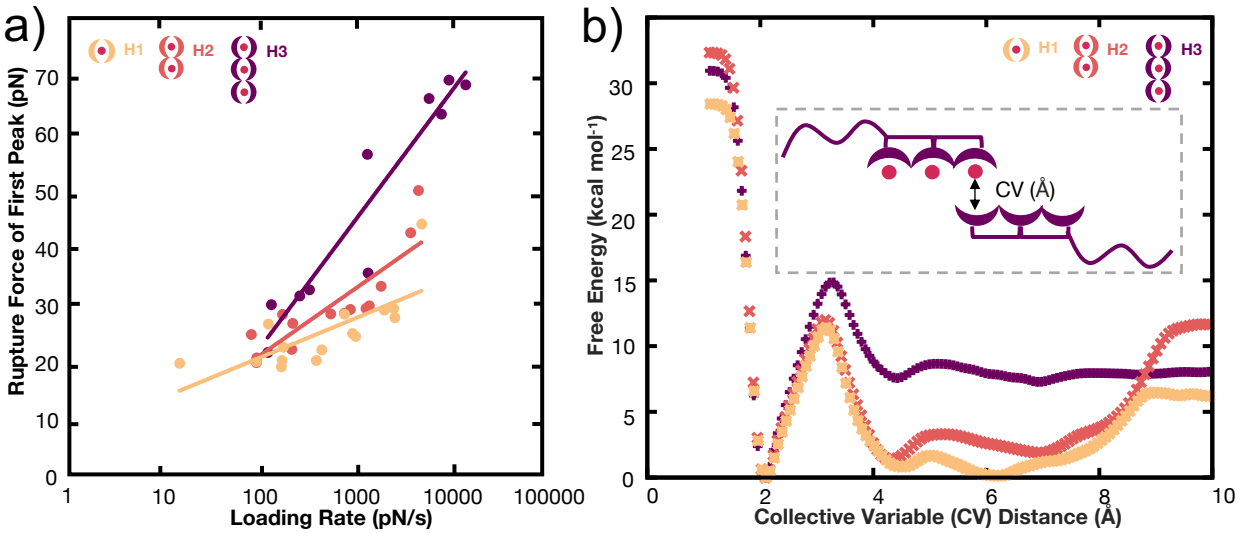


Figure 3-16. Loading rate dependence of single-bond rupture events in SMFS experiments and underlying energy landscape. a) Assuming that the first peak represents rupture events where only one metal-ligand bond breaks, the data was fitted to a Gaussian model with two terms to extract the most probable rupture force value of the "one-bond breaking" rupture events. The loading rate dependence shows a similar trend as the pulling-speed dependence of "single-bond breaking" events in simulation (**Figure 3-15b**). The Gaussian fitting is made in MATLAB using the following equation:

$$y = \sum_{i=1}^{2} a_i e^{\left[-\left(\frac{x-b_i}{c_i}\right)^2\right]}$$

and only data points with a fitting R^2 value of 0.9 minimum are included. b) Free energy landscapes of one bond rupturing in the H1, H2, and H3 show differences due to additional histidine groups present. Using PLUMED, an implementation of metadynamics for accelerated sampling, the free energy of H1, H2, and H3 was calculated for a protein configuration where only one coordination bond is initially bound and then ruptures (see schematic). This set up mimics the one bond rupture events in the AFM. The collective variable, or reaction coordinate, of the free energy landscape of the protein systems is the distance between the coordinating nitrogen on the histidine residue and the Ni^{2+} ion. The energy landscape reveals that even though only one bond is bound and ruptures in H1, H2, and H3, the energy landscape of H3 is different than that of H1 or H2. This indicates that the additional histidine residues affect the coordination bond even though they are not directly participating in a coordination interaction.

Though the two modes of single- and multiple-bond cooperative rupture resolved through SMD explain the majority of the results observed in SMFS, other mechanisms are possible. For example, H2 has minor differences from H1 in simulation but not in experiment. In experiment, it is possible that H2 is only forms one bond, even though two coordination bonds are theoretically possible. It is also possible that increasing the number of coordination sites increases the probability for rebinding along the rupture pathway, as the coordination bonds are highly dynamic. In our simulations however, rebinding did not occur unless a specific metastable artificially constrained protein geometry was used (**Figure 3-18**). Even at the longer timescales probed in the SMFS experiment where rebinding may become more likely as the two chains remain in proximity for a longer period of time, we still consider rebinding as the less likely mechanism. The experimental rupture force histograms that include rebinding events are expected to display a wider distribution of forces.[352] Here, two distinct peaks are observed in the rupture force histograms (**Figure 3-12d**), suggesting that indeed two distinct rupture pathways are probed.

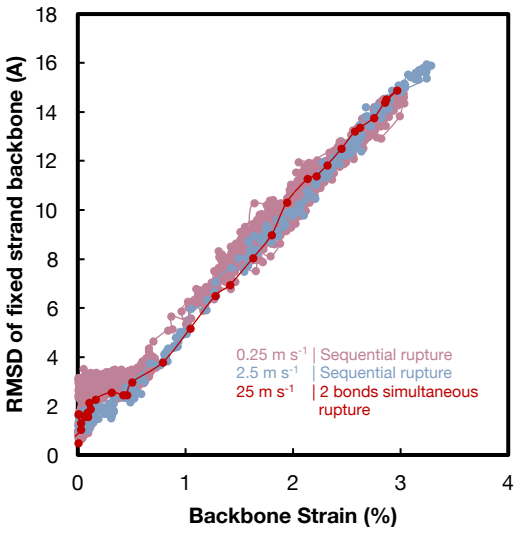


Figure 3-17. Backbone fluctuation based on pulling speed in simulations. Decreasing the pulling speed in simulation results in increasing fluctuation of the protein backbone atoms. This indicates that slower pulling speeds allow greater spatial exploration and relaxation of the backbone. One representative example of H3 at several pulling speeds is demonstrated, and the backbone relaxation dependence on speed is observed across all simulations. Data is plotted every 10 ps.

Despite imposing idealized binding conditions with a bidentate trans coordination, the SMD trajectories demonstrate highly heterogeneous rupture pathways. Several rupture pathways emerge (**Figure 3-15c, 3-15d**), indicating the stochasticity of metal-coordination bond rupture. The same polypeptide at the same pulling speed shows different rupture pathways (**Figure 3-15c, 3-15d**) under different initial conditions including polypeptide structures or initialized atomic velocities. This heterogeneity is also visualized in **Figure 3-15e**, where bonds break sequentially, and in **Figure 3-15f**, which highlights the simultaneous rupture of two bonds.

Our results imply that metal-coordination bonds can exhibit cooperative rupture, but are highly stochastic with significant heterogeneity in rupture pathways even under the same pulling conditions. Further, the two-to-three bonds observed to simultaneously rupture aligns with the N_{cr} prediction in **Figure 3-10b**. Next steps involve investigating the effects additional coordination bonds.

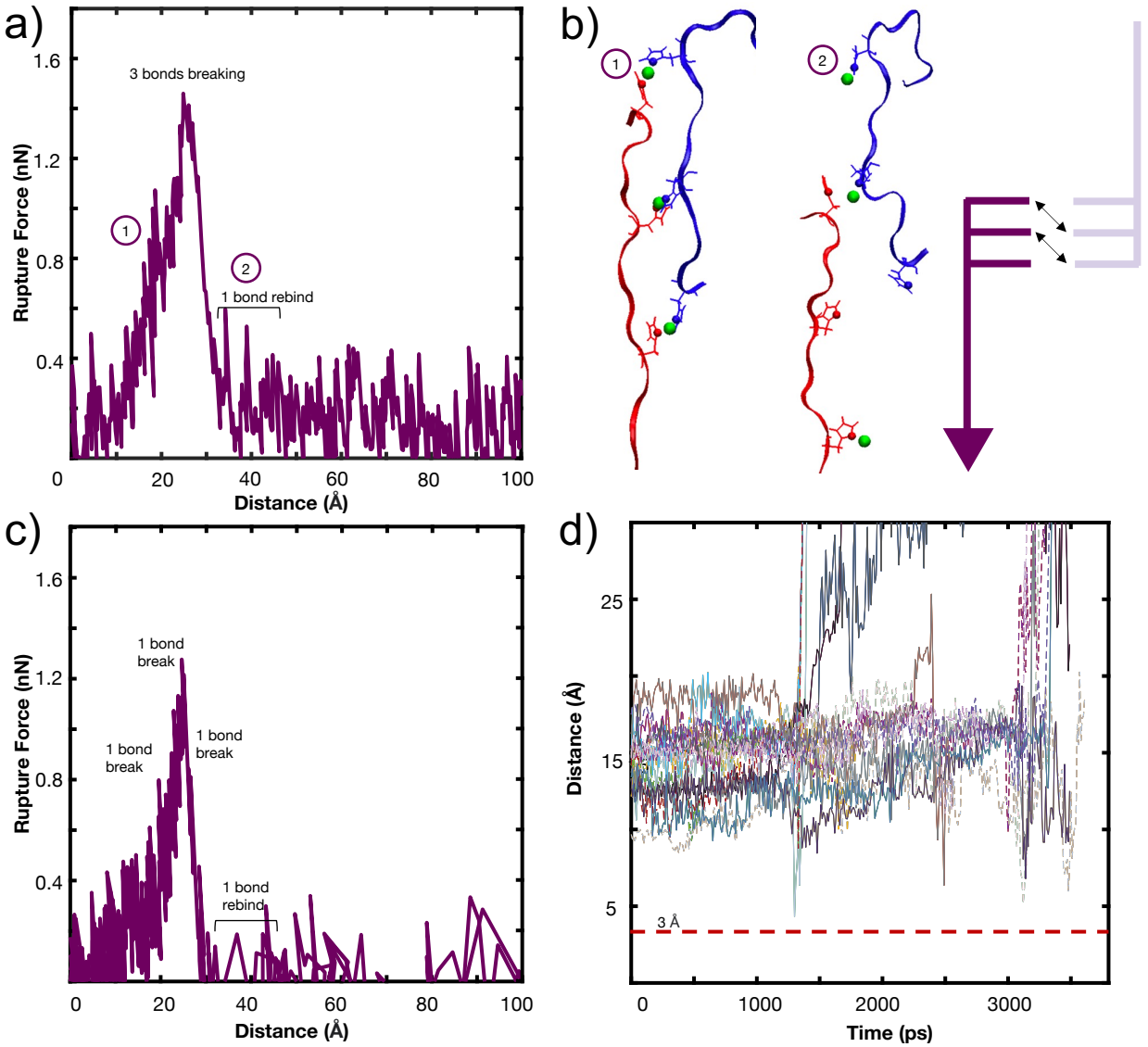


Figure 3-18. Rebinding events in simulations. When the H3 protein is artificially constrained such that the protein backbone is nearly linear, rebinding can be observed under specific pulling conditions. Though the protein is not stably equilibrated under this geometry, SMD simulations are still conducted to explore potential rebinding effects. Under this constrained geometry at a pulling speed of 25 m s^{-1}, in (a), rebinding can be observed in a nondistinctive binding trace at ~600 pN. Notably, rebinding is extremely sensitive, where even changing the randomly initialized atomic velocities for the same protein geometry does not show rebinding. b) Snapshots of this rebinding event and a schematic illustration of rebinding. c) At slower pulling speeds of 0.25 m s^{-1}, the rebinding event occurs at a lower force of ~200 pN. d) Across all simulations tested of the equilibrated H2 and H3 protein structures at varying pulling speeds, rebinding is not observed. Traces of the distance between potential rebinding partners at 0.25 m s^{-1} are shown, and the distances do not reach the ~3 Å required to enable rebinding.

3.4.4 Discussions, Implications, & Conclusions

Using a combination of single-molecule force spectroscopy (**Figure 3-12**) and all-atom steered molecular dynamics (SMD) simulations (**Figure 3-15**), we investigated the contribution of multiple metal-coordination bonds to the stability of metal-coordinated polypeptides (**Figure 3-9**). We found that metal-coordination bonds exhibit cooperative rupture and resolved at least

two dissociation pathways. Along the first pathway, coordination bonds dissociate sequentially. At the same time, synergistic rupture of multiple bonds is possible and characterized by more-than-additive equilibrium binding constants and rupture forces.

Our work has important implications to the characterization and design of proteins with metal-coordination bonds. It is critical that experimental strategies are developed that ensure single-molecule specificity for systems with multiple coordination bonds. Our work contributes explicit methodology, quantitative rupture force values, and a mechanistic understanding of the nature of these bonds—an important insight due to the highly utilized nature of histidine-Ni^{2+} bonds. The principles uncovered in this work likely occur for other metal-coordination systems. In fact, evidence for the sequential and simultaneous rupture of ferric-thiolate bonds in an iron-sulfur rubredoxin protein was recently discovered.[353]

Most importantly, this work contributes to an understanding of how biological materials exploit weak dynamic noncovalent interactions to build strong and tough materials. We reveal that cooperativity for metal-coordination bonds increases protein strength, a concept likely applies for most other dynamic noncovalent interactions. The ability of the bonds to rupture simultaneously makes metal-crosslinked materials strong at high pulling speeds, e.g. when mussel byssal threads are exposed to crashing waves, which can be up to 5 to 15 m s^{-1},[16] or upon crushing mandibles. At the same time, the material can dynamically reorganize at rest through heterogeneous rupture modes, suggesting the strong energy-dissipative role of metal-coordination bonds.[17,26,55,284,354] Altogether, the presented work provides mechanistic insight into the sequential and cooperative rupture modes of increasing numbers of intermolecular metal-coordination bonds on a protein template, allowing for both the rational design of metal-coordinated materials with target mechanical properties (toughness, strength) and the fundamental understanding of biological exploitation of weak noncovalent interactions.

3.4.5 Materials and Methods

Polypeptide design and expression

ELPs are a class of intrinsically disordered, flexible polypeptides composed of repeating GXGVP motifs, where X, the guest residue, stands for any amino acid except proline.[355] The MC-ELP design was based on the work of Callahan et al.[356] and the bacterial expression on Martín-Moldes et al.[357] Bioengineered MC-ELP were designed to contain one to three histidines, located in the N-terminal guest residues. The MC-ELPs have the general structure $(GHGVP)_n(G\text{-}S/T/V\text{-}GVP)_{33}$-C where "n" represents 1, 2, or 3. The ELP backbone is composed of a combination of GXGVP pentapeptides whose guest amino acids are serine, threonine, or valine in a 7:3:23 ratio. A cysteine is added at the C-terminus to enable the immobilization of the polypeptides for SMFS experiments. The full DNA and amino acid sequence are included in **Appendix Method 3-S1.**

The DNA sequences encoding the MC-ELPs were synthesized by GenScript (Piscataway, NJ, USA). The genes, flanked by NheI and SpeI restriction sites, were obtained in a pUC57 vector. Using standard molecular biology protocols, the genes were cloned into the expression vector pET25b+ (MilliporeSigma, Burlington, MA, USA), using the mentioned restriction sites. The correct insertion of the genes was verified with DNA sequencing. All constructs were transformed into the *E. coli* strain BLR(DE3) (Novagen-SigmaAldrich, St. Louis, MO, USA) and grown on LB agar plates supplemented with 25 μg mL^{-1} ampicillin. Using single colonies, seed

cultures were started, using 25 mL of Luria-Bertani medium, supplemented with 25 μg mL^{-1} ampicillin. The seed cultures were grown overnight at 37°C and 250 rpm. The seed cultures were transferred to 1 L of Terrific Broth supplemented with 0.8% (v/v) glycerol and 25 μg mL^{-1} ampicillin. The culture was induced with 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG; (Sigma-Aldrich, St. Louis, MO, USA) at on OD_{600} ~0.8-1. Bacteria were harvested 5 hours after induction by centrifugation at 8000 rpm for 15 min at 20°C.

The MC-ELPs were purified using the inverse transition temperature cycling method.[358] Briefly, the bacterial pellet was resuspended in phosphate-buffered saline (PBS; 210 mg/L KH_2PO_4, 726 mg/L $Na_2H_2PO_4$-7 H_2O, pH 7.4, 9000 mg/L NaCl), using 10 mL of buffer for 1 g of pellet. The cells were lysed using sonication while kept in an ice bath. The cell lysate was cleared by centrifugation at 9000 rpm and 4°C for 30 minutes. NaCl was added to the supernatant to a concentration of 3% (w/v) and the supernatant was incubated at 70°C for 1 h to induce the precipitation of the polypeptides. The sample was then centrifuged for 3 min at 9000 rpm at 30°C. The supernatant was discarded and cold deionized water was added to the remaining pellet. The precipitated target polypeptide was dissolved overnight under strong magnetic stirring at 5°C. The sample was then centrifuged at 9000 rpm for 15 min at 4°C, the pellet was discarded and the supernatant was dialyzed (MWCO 7 kDa, ThermoFisher Scientific, Rockford, IL, USA) against deionized water for 3 days. The yield of the MC-ELPs was approximately ~300 mg L^{-1} of bacterial culture. The purity of the polypeptides was confirmed via non-reducing SDS-PAGE (**Appendix Figure S3-1**).

Isothermal Titration Calorimetry

All titrations were performed using a VP-ITC MicroCalorimeter Microcal LLC at 25°C. Solutions for titrations were prepared in a non-coordinating HEPES buffer (10 mM HEPES/NaOH pH 7.4, 140 mM NaCl). A solution of $NiCl_2$ hexahydrate (720 μM) was titrated as the ligand to 24 μM H3, 36 μM of H2, or 72 μM of H1 polypeptide solution. Reference titrations, performed by adding the $NiCl_2$ solution to HEPES buffer, were subtracted from the corresponding titrations to account for the effect of the dilution. The fitting was performed using the Microcal Origin Software. The equation for the one site model fitting is:

$$Q = \frac{nM_t\Delta HV_0}{2}\left[1 + \frac{X_t}{nM_t} + \frac{1}{nKM_t} - \sqrt{\left(1 + \frac{X_t}{nM_t} + \frac{1}{nKM_t}\right)^2 - \frac{4X_t}{nM_t}}\right.$$

where Q is the total heat content of the solution in V_o, the active cell volume, M_t is the bulk concentration of macromolecule in V_0, n is the number of sites, ΔH is the molar heat of ligand binding, and X_t is the bulk concentration of the ligand, and K_a is the association constant. $n, K_a, \Delta H$ are iterated using standard Marquardt methods until no significant improvement in fit occurs with continued iteration.

CD Spectroscopy

CD spectra were recorded to investigate the secondary structure of the MC-ELPs in absence and presence of excess Ni^{2+} ions above and below their T_t. MC-ELPs were dissolved in HEPES buffer at a final concentration of 15 μM. When present, 300 μM $NiCl_2$ was added in a 10:1 molar ratio of $NiCl_2$ to the MC-ELP. Spectra were acquired using a Jasco J-815 Circular Dichroism Spectrometer (Easton, MD, USA) and a 1 mm cuvette (model 1-Q-1, Spectrecology, St. Petersburg, FL). Spectra are the result of accumulating three scans, acquired at a scanning speed of 20 nm s^{-1} and 4 s of Digital Integration Time (DIT), from 280 to 190 nm. The High-Tension Voltage, HT(V), of the photomultiplier was kept below 600 V.

UV/Vis Spectroscopy

Polypeptides were dissolved HEPES buffer at a final concentration of 30 μM. When present, the $NiCl_2$ solution was added in excess at final concentration of 300 μM, in a 10:1 molar ratio of $NiCl_2$ to MC-ELP. The turbidity of each sample, with and without $NiCl_2$, was measured in a UV/Vis spectrophotometer (Aviv 14DS spectrophotometer equipped with a Peltier temperature controller; Aviv Biomedical, Lakewood, NJ) in a 1 mL quartz cuvette. Absorbance of the sample was monitored at 350 nm as a function of increasing temperature from 25 to 70°C at a heating rate of 2.5ºC min^{-1}.

Preparation of glass coverslips and AFM cantilevers

The cysteine terminated MC-ELPs were immobilized to cover slips and AFM cantilevers via polyethylene glycol (PEG) spacers, using a previously established protocol.[359] Briefly, glass coverslips (Menzel Glaeser, Braunschweig, Germany) and cantilevers (MLCT, Bruker, Camarillo, CA, USA) were cleaned and activated via 10 min UV-ozone treatment, followed by silanization with 3-aminopropyl dimethylethoxy silane (ABCR, Karlsruhe, Germany). Both the coverslips and cantilevers were treated in parallel. They were incubated in 50 mM sodium borate (pH 8.5) for 1 h to increase the fraction of deprotonated amino groups for the subsequent coupling of the heterobifunctional NHS-PEG-maleimide spacer (M_w = 10,000 g mol^{-1}, Rapp Polymere, Tuebingen, Germany). NHS-PEG-maleimide was dissolved in a concentration of 50 mM in sodium borate and incubated on the surfaces for 1 h at room temperature. Following incubation, the surfaces were washed with ultrapure water and dried under nitrogen flow. The MC-ELPs were dissolved to a concentration of ~100 μM in coupling buffer (composition). The polypeptide solutions were added to beads carrying immobilized tris(2-carboxyethyl) phosphine (TCEP, Thermo Scientific) to reduce possible disulfide bonds. The samples were incubated at 4ºC with shaking. The reduced polypeptide solutions were diluted to a final concentration 30 μM and a volume of ~50 μL was pipetted on top of the coverslip or cantilever. After incubation for 1 h at 4 ºC, the cantilever and surfaces were rinsed multiple times with the HEPES buffer to remove non-covalently bound polypeptides. Coverslips and cantilevers were stored in HEPES buffer until use.

AFM-SMFS

All SMFS measurements were performed with a ForceRobot® 300 instrument (Bruker Nano, Berlin, Germany), using MLCT cantilever C with a nominal spring constant of 0.01 N m^{-1}. The measurements were performed at room temperature, using in HEPES buffer with 1 mM $NiCl_2$. To obtain measurements over a broad range of loading rates, data was collected at different retract velocities, ranging from 200 to 5000 nm s^{-1}. For each retract velocity, ~3000 to 5000 approach-retract cycles were carried out on a 10 x 10 μm^2 grid (**Appendix Table S3-1**). Cantilevers were calibrated for each retract velocity as the cantilever properties changed in the $NiCl_2$ solution. The thermal noise method was used,[360] applying a correction factor of 0.817 for the cantilevers.[361] The spring constants determined varied between 0.01 and 0.014 N m^{-1} and the sensitivities varied between from 43.71 and 54.96 nm V^{-1}. Several independent experiments were performed, using independently functionalized cantilevers and coverslips.

The obtained data was converted into force-extension curves, using the JPK data analysis program (version 6.1.41). PEG represents a well-characterized spacer for SMFS measurements and together with the ELP,[304] allows the discrimination of specific versus nonspecific interactions. Further, the spacer has a characteristic force-extension behavior, which can be described by the extensible freely jointed chain (eFJC) model. As PEG and ELP have similar Kuhn lengths (values at around 0.8 to 1 nm.[304,305,347]) a single eFJC fit with one Kuhn length is

possible. Force-curves within a pre-defined range of Kuhn lengths and contour lengths were selected for further analysis. The range of Kuhn lengths was between 0.4 and 1.6 nm and a contour length cut-off of 100 nm was applied. This value exceeds the contour length of one chain (which is?) and thus ensures that the chain ruptured includes the polymer spacers on both the coverslip and the cantilever. For each force-extension curve, the rupture forces and corresponding loading rates were determined using the JPK data analysis program and plotted as histograms (**Appendix S3: 2-4**). Considering the broad rupture force histograms (at least for H3), the mean rupture forces were plotted against mean loading rates to obtain a first overview of the data. To compare the three MC-ELPs the mean rupture forces and loading rates, the data (**Figure 3-11c**) was further fitted to the Bell-Evans model:[335,336]

$$F^{*} = \frac{k_b T}{x_B} \ln\left(\frac{r x_B}{k_{off} k_B T}\right)$$

where $F^{*}$is the most probable rupture force, r is the loading rate, $k_B$is the Boltzmann constant, T is the temperature, k_{off} is the dissociation rate extrapolated to force-free conditions, and x_B is the potential width.

Steered molecular dynamics

The ELP-H_n polypeptides (**Figure 3-9a**) were prepared such that each histidine nitrogen atom is coordinated in a bidentate geometry via the Ni^{2+} ion to the opposite polypeptide chain. Simulations were implemented with Nanoscale Molecular Dynamics (NAMD) and all simulations utilized the CHARMM22 force field[229] with Ni^{2+} parameters from Babu et al.[231] and a 2 fs timestep. The Ni^{2+} ions were balanced with Cl^{-} ions for charge neutrality and the peptide dimer was solvated with a 15 Å mesh of TIP3P water molecules.[362] Periodic boundary conditions were used with the Particle Mesh Ewald full system electrostatics method. After careful energy minimization using the conjugate gradient algorithm in NAMD, the simulations were equilibrated for 60 ns under NPT (1 atm, Nose-Hoover Langevin piston pressure control), followed by 60 ns in NVT (Langevin dynamics).

Independent simulations under this procedure were carried out for subsequent SMD tests. The TIP3P water molecule box was extended by 60 Å to account for deformation in the pulling direction. The larger water box was briefly equilibrated for 2 ns under NPT and 0.5 ns under NVT. The C_α atom on the C-terminus of one polypeptide was selected as the SMD pulling atom, and the C_α atom on the opposite strand C-terminus was selected as the fixed atom. SMD data were collected every 0.2 ps and pulling speeds from 0.25 m s^{-1} to 25 m s^{-1} were used. The slowest pulling speed is still significantly faster than the fastest experimental retract velocity due to computational limitations. The simulation was run until all intermolecular metal-coordination bonds fully ruptured. The rupture of a metal-coordinate bond was defined as a distance of greater than 3Å between the coordinating nitrogen of histidine and the Ni^{2+} ion. The time, distance, force, and type of bond rupture was recorded for each rupture event in each simulation. Simultaneous rupture was defined as multiple bond breaking events within 20 ps of each other in the visual file with the rupture force peaks that were indistinguishable from each other in the force-distance trace.

We found that tetradentate structures were not stable during initial equilibration, as the metal-coordination binding sites would quickly dissociate into tridentate, bidentate, or monodentate structures during the equilibration process. Given this, we continued SMD tests with only bidentate coordination. While this lack of tetradentate stability may be due to challenges with

the complex force fields for metal ions, speciation models predict a dominance of bidentate coordination stoichiometry in aqueous conditions between histidine and Ni^{2+}.[216,330,331] See codes in **Appendix 3.**

CHAPTER 4

NANOMECHANICS OF HIGHLY METAL-COORDINATED PROTEINS

4.1 Significance Statement

In this chapter, we apply the principles learned from metal-coordination bonds in model systems to metal-coordination bonds for mechanical functions in natural biological proteins. As discussed in the first chapter, several biological organisms utilize metal-coordination bonds for mechanical function. Here, we use the *Nereis virens* marine worm jaw to probe the mechanical role of metal-ions in their native biological environment. Though the structure of a major component of the worm jaw, the Nvjp-1 protein, has recently been resolved, a detailed nanostructural understanding of the role of metal ions on the structural and mechanical properties of the protein is missing, especially with respect to the localization of metal ions.

In this chapter, replica exchange MD and steered MD were used to explore how the initial localization of the Zn^{2+} ions impacts the structural folding and mechanical properties of Nvjp-1. We find that the initial distribution of metal ions for Nvjp-1, and likely for other proteins with high amounts of metal-coordination, has important effects on the resulting structure, with larger metal ion quantity resulting in a more compact structure. These structural compactness trends, however, are independent of the mechanical tensile strength of the protein, which increases with greater hydrogen bond content and uniformity of metal ion distribution. Our results indicate that different physical principles underlie the structure or mechanics of the proteins, with broader implications in the development optimized hard bioinspired materials and the modeling of proteins with significant metal ion content.

4.2 Introduction

Nereis virens is a polychaete burrowing marine worm with a hard and stiff jaw used for feeding and defense in abrasive environments. The hardness and stiffness of the marine worm jaw is remarkable reaching up to ~0.8 GPa, paralleling the hardness of human cortical bone.[15,363] Further, *Nereis virens* is able to retain its hardness even in a hydrated state.[14,364] This behavior is especially intriguing, given that its jaw proteins are mostly organic compared to the mineralized calcified teeth or hard tissues found in higher level organisms, which have an inorganic phase of around 75-95% by mass.[15,61,62] Metal ions, and in particular Zn^{2+}, have been found to play a structural role in the mechanical properties observed in these polychaete worms.[14,15,61]

The isolation and structural prediction of Nvjp-1, a major component of the *Nereis virens* distal worm jaw extracts, has started to enable a detailed nanostructural understanding of how Zn^{2+} ions may contribute to the mechanical properties of the protein.[13] Nvjp-1 is a histidine-rich

(over 25 mol%) protein that undergoes significant hydrodynamic changes depending on pH or the presence of metal ions.[13] In agreement with experiment, our group previously computationally predicted the structure of Nvjp-1 and found that the structure becomes more compact as the ratio of Zn^{2+} ions to protein increased.[47] Through additional simulations, Bekele et al. found that pH also affects the protein structure, where metal binding happens with polar residues at low pH and is passed onto carboxylate or imidazole coordination pockets at neutral pH.[365]

While these research efforts have provided important insights into the role of pH and Zn^{2+} quantity on structural binding and mechanical properties, an understanding of how the location or distribution of Zn^{2+} affects these properties is missing, preventing a detailed understanding of how biological organisms use such metal ions for structural function. The larger worm jaw itself exhibits a specific metal ion gradient that directly relates with its stiffness and hardness,[44,61] and a similar distribution could be expected at the nanoscale as well. Further, metal ions are known to play a role in protein folding by changing the underlying protein folding energy landscape.[366,367] For example, in some biological organisms, Zn^{2+} has been found to induce amyloid-like conformations,[368,369] including in some marine organisms.[13,97] Given these observations, it is reasonable that the location of Zn^{2+} may have a strong effect on Nvjp-1 protein structure and resulting mechanical properties. Developing such an understanding computationally would help clarify additional insights into how Zn^{2+} may enable the remarkable mechanical properties in *Nereis virens,* especially because the exact localization of Zn^{2+} ions in Nvjp-1 would be hard to ascertain through experiment. Given that few well-characterized protein structures with several metal-coordination bonds exist, the example with the Nvjp-1 protein provides foundational insights that likely applies to other metal-coordinated protein structures as well. This broader understanding of the location-dependent metal-ion crosslinking effect in proteins would help yield additional design principles for how sclerotized proteins could be synthetically designed to create hard structures like biology.[91]

In this chapter, we focus on understanding how the localization of Zn^{2+} impacts the structural folding and mechanical properties of Nvjp-1, with the goal of contextualizing such work for metal-coordinated proteins largely. Replica exchange MD (REMD) simulations are performed with explicit solvent to investigate the formation of the protein structures in various coordination environments. Metal ions are initiated in different positions of the Nvjp-1 protein, previously determined by REMD in implicit solvent,[47] to understand how the protein folds in the presence of metal ions. Steered MD (SMD) experiments are conducted on the resulting protein structures to reveal how the location of metal-coordination bonds impacts mechanical tensile properties. This combination of structural prediction and mechanical properties gives information about Nvjp-1 binding with metal ions and will enable its broader use in mechanomutable engineering materials.[47,91] With this information, optimized bioinspired materials could be created for a variety of practical applications, especially those requiring hard, sclerotized structures.

4.3 Replica Exchange MD Simulation Convergence

Zn^{2+} ions are initially distributed in three orientations in the solved REMD structure from Chou et al. (**Figure 4-2**).[47] As a baseline case, named “A”, the Zn^{2+} ions are uniformly distributed throughout the protein, a common practice when the exact coordination structures are unknown.[47,370] In the second condition, named “B” for “beta,” Zn^{2+} ions are distributed around

the small clusters of beta-sheets found in Nvjp-1, to determine whether this placement facilitates the formation of amyloid structures as in other proteins.[13,97,368,369] The third condition, named “C” for compact, localizes Zn^{2+} ions in the compact, condensed regions of Nvjp-1. Such localization may have a reinforcing effect[17,371] on the already compact regions of the protein through additional coordination bonds.

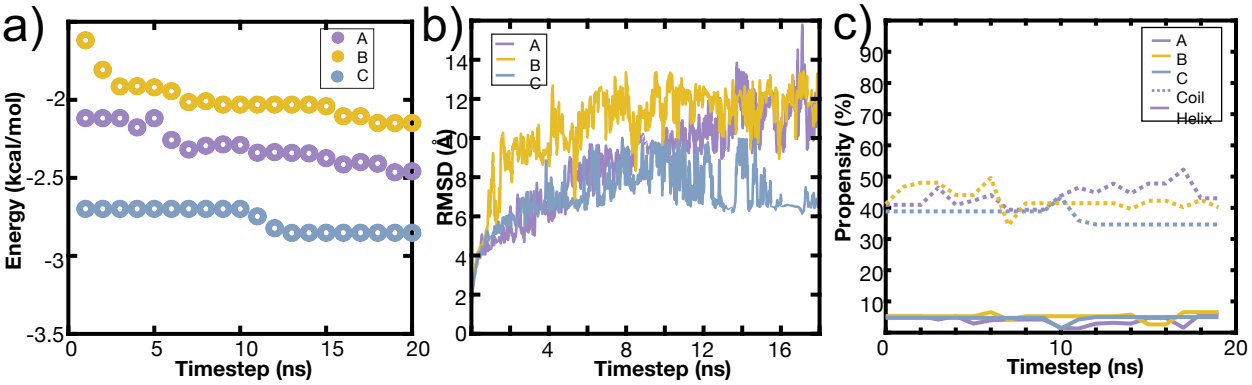


Figure 4-1. REMD simulations are converged. a) The lowest energy cluster for simulations A, B, and C reaches a stable energy value by 20 ns. b) The root mean square deviation (RMSD) of the 300K replica reaches a stable value. c) The coil and helix secondary structure of the 300K replica reaches a stable value by 20 ns.

The REMD simulation convergence was assessed by several criteria. As shown in **Figure 4-1a,** the average energy of the lowest cluster energy across the simulations approached a stable plateau. The root mean square deviation of protein structure at 300K was also analyzed for the simulations and remains relatively constant after 10 ns in **Figure 4-1b**. Last, the coil and helix secondary structure are shown as a function of time in **Figure 4-1c**. This analysis metric also reached a plateau around 12-13 ns per replica. While it is always possible to run the REMD simulation for longer times for more accurate structural predictions, these several criteria together indicate that the REMD simulations have reached a converged state, whereby further characterization would yield reasonable results. The initial starting point of a folded protein structure also accelerates the convergence of the REMD calculation.

4.4 Lowest energy protein structures

To understand how metal-coordination bonds affect the protein’s folding and thereafter mechanical function, three lowest energy representative structures were obtained from the REMD simulations, as shown in **Figure 4-2.** According to our simulations results, the initial conditions of the metal ion distribution affected the resulting structure of the Nvjp-1 proteins. When the metal ions were uniformly distributed throughout the protein, as is the case for simulation A, the protein structure did not deviate significantly from the initial structure. Comparing the smaller magnitude of the structural change in simulation A versus B/C suggests that the metal ions in the middle turn region of the protein contribute to locking the protein in its original structure and providing a higher energy barrier when folding. As these middle turn region metal ions are present in A, but not in B/C, less change in structure A was observed.

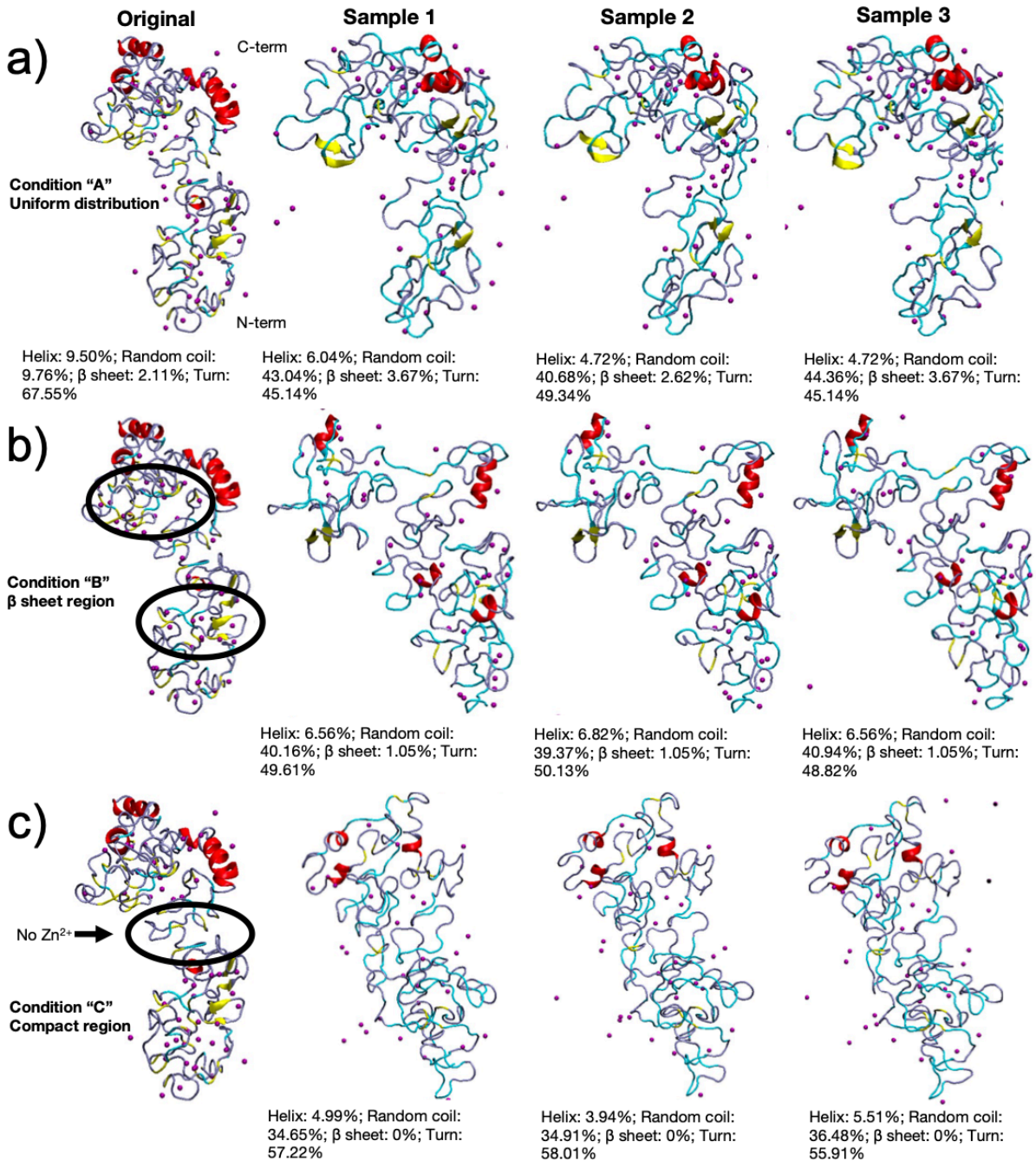


Figure 4-2. Representative protein structures from REMD simulations. Protein structures are initialized with different distributions of metal ions; uniformly distributed (simulation "A"), clustered in the β sheet regions (simulation "B"), or clustered in the compact regions (simulation "C"). Zn^{2+} ions are colored in purple. Secondary structures are colored as: yellow β sheet, red helix, cyan coil, and iceblue turns. Structures 1-3 shows three representative structures from the lowest energy cluster of each converged REMD simulation. a) In simulation A, final protein structures do not deviate significantly from the original structure, indicating that the Zn^{2+} stabilizes the protein structure such that β sheets and helices are mostly conserved. b, c) In simulations B and C, the proteins become more disordered. The difference in the degree of conserved structure in simulation B/C versus A suggests that the uniform distribution of Zn^{2+} and stabilizing effect of Zn^{2+} towards the middle of the protein results in additional energy barriers during protein folding.

Across all three simulation conditions, the converged proteins showed a more dispersed distribution of metal ions, with fewer helixes and turns, and the coil secondary structures dominates the protein structure (**Figure 4-3a, Table 4-1, Figure 4-4**). The amount of random coil increased across all simulations, where A had the most random coil content, followed by B, then C. A had more metal ions clustered towards the C-terminus region of the protein, whereas B has more metal ions clustered towards the N-terminus. The metal ions in C were more uniformly distributed throughout the protein. The proteins also all had similar radius of gyration values within 1 Å of each other, but A and C appeared to have a more spatially uniform distribution of amino acids, whereas B had a 10 amino acid random coil connecting two compact regions (**Figure 4-2**).

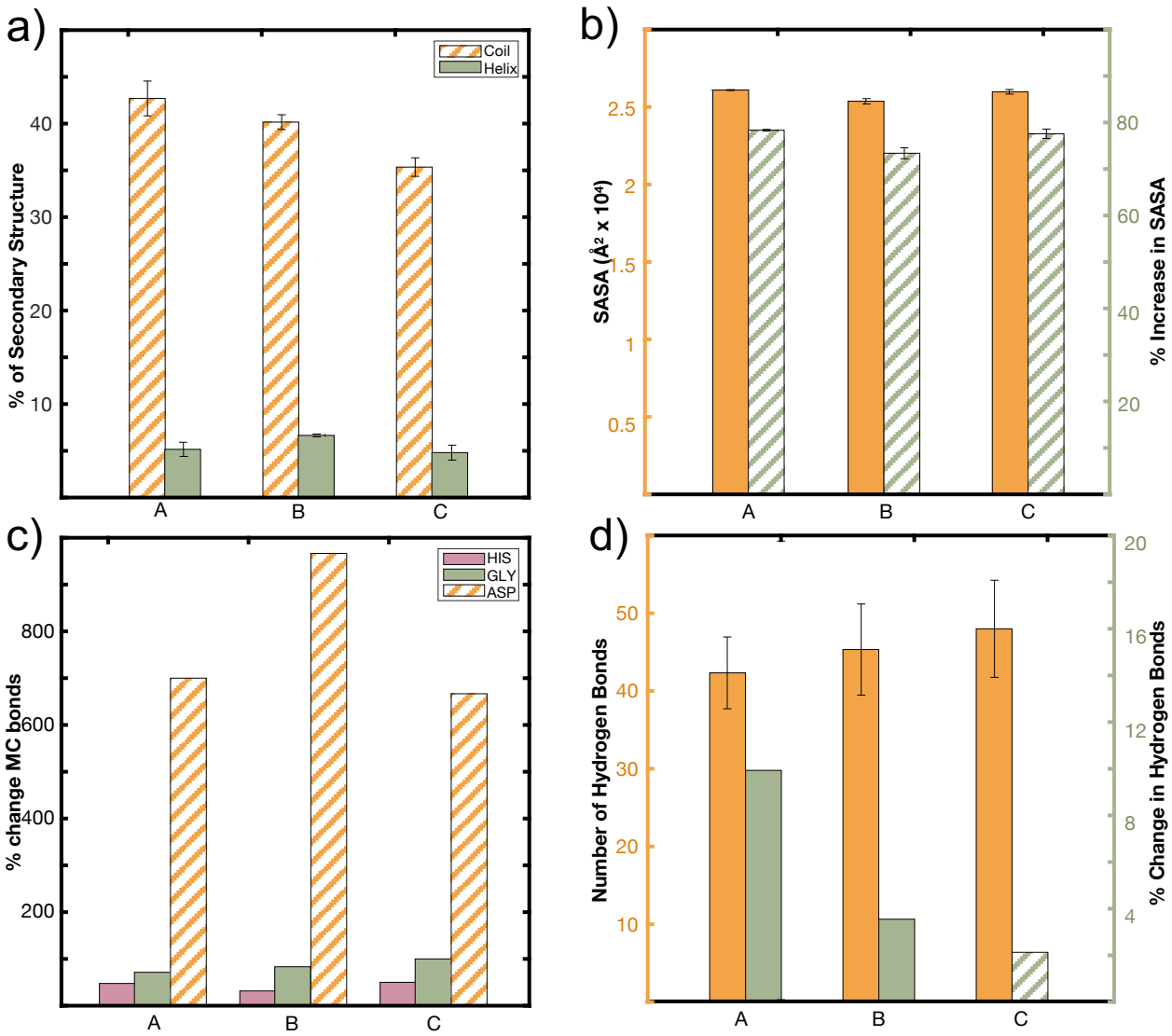


Figure 4-3. Characterization of representative protein structures. To quantify the structure characteristics of Figure 4-2, the STRIDE secondary structure assigner in VMD is used. The diagonal hash filling pattern across all sub-figures represents an increase in the % change, rather than a decrease. All % changes are evaluated from the initial structure "original" in **Figure 4-2** compared to the REMD structures. a) All simulations have more random coil secondary structure from A>B>C, and the amount of helix structure is the same across the 3 sets. b) Structures A,B,C have similar SASA values with an ordering of A~C>B for both the value and % increase from the original structure. c) A reorganization of MC bonds is observed, where His and Gly lose several MC bonds, while Asp gains

MC bonds. Although Gly does not coordinate with metal ions, it is demonstrated here for comparison as the protein is Gly-rich. d) The number of hydrogen bonds and increase in hydrogen bonds in the protein changes from C>B>A.

Table 4-1. Secondary Structure Propensity (%) for Figure 4-3a.

Secondary Structure	Initial	A	B	C
Coil	9.7	42.7	40.2	35.4
Helix	9.4	5.2	6.7	4.8

Further, there was an increase in solvent accessible surface area (SASA) across all structures (**Figure 4-3b, Table 4-2**), indicating that the proteins are unfolding with the addition of the metal ions. While A and C had similar SASA values in terms of both absolute value and percent increase, B had the lowest SASA value and percent increase, indicating that it is the most compact structure.

Table 4-2. SASA (Å^2) for Figure 4-3b.

A	B	C
26095.7 ± 28.1	25371.0 ± 172.5	25984.0 ± 148.1

This increase in the SASA relates directly with the trend observed for metal-coordination bonds present in the system. **Figure 4-3c (Table 4-3)** shows that the aspartate residues gained metal-coordination bonds from an increasing to decreasing order of B then A~C, and that the histidine residues lost metal-coordination bonds where A and C are similar, but B lost the fewest histidine-Zn^{2+} bonds. Further, the absolute number of metal-coordination bonds follows the trend where B has the most coordination bonds, followed by A~C. This, together with the SASA indicates that more metal-coordination bonds result in a more compact structure. Further, the Zn^{2+} ions stabilize the beta-sheet clusters in B, as in other proteins,[13,97,368,369] even though the formation of amyloid-like structures is not observed. The increased coordination with Asp is also in agreement with the carboxylate coordination observed in Bekele et al.[365] These increases in carboxylate coordination were observed in the molecular snapshots in **Figure 4-5b** and **c**, and interestingly, the carboxylate groups also coordinated Zn^{2+} ions on the outer surface of the protein (**Figure 4-5c**). The aspartate coordination is surprising as aspartate is only ~7% of the amino acids in the protein whereas histidine constitutes ~27%.

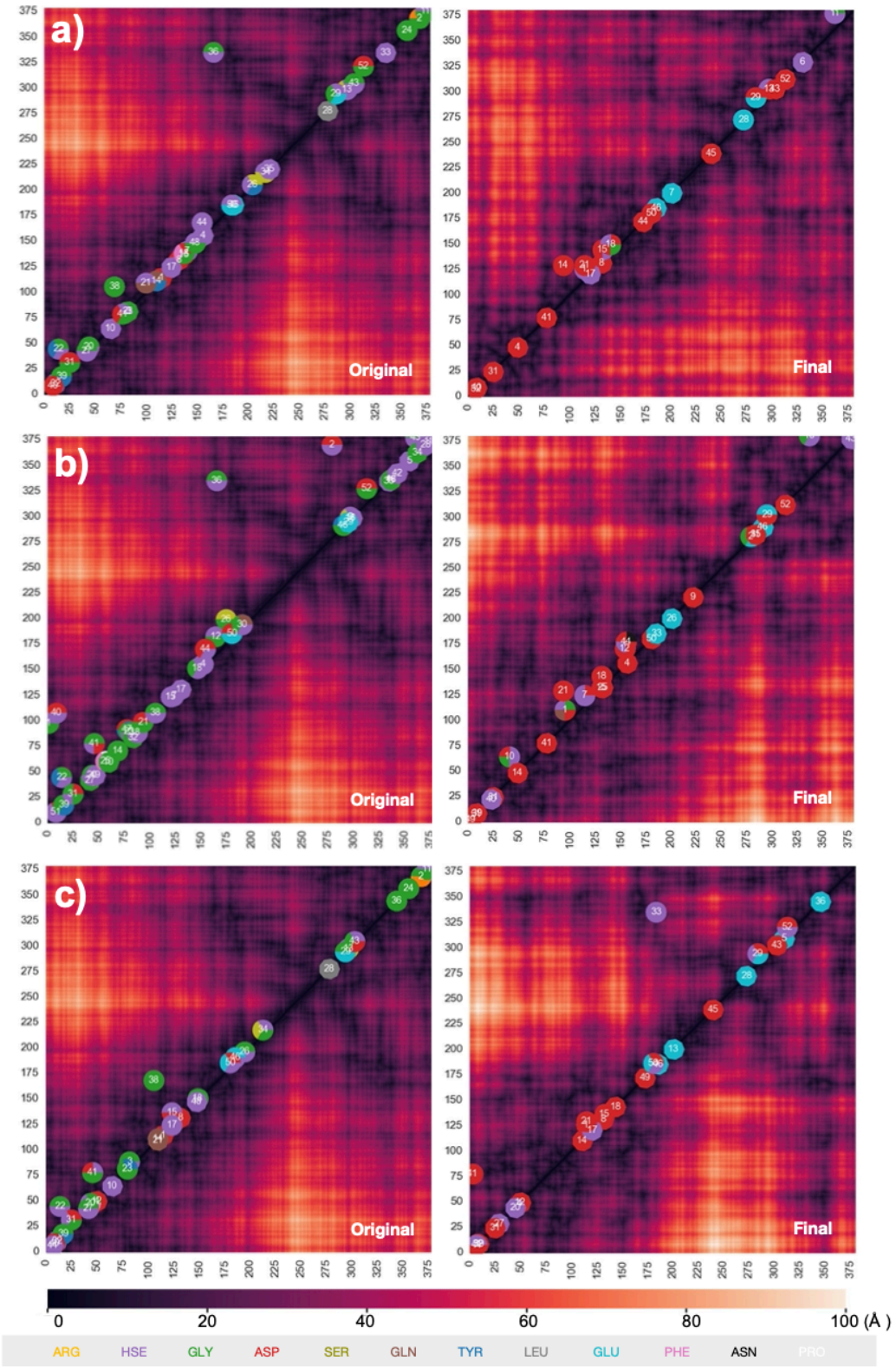

a)
Original
Final
b)
Original
Final
c)
Original
Final
0
20
40
60
80
100 (Å)
ARG
HSE
GLY
ASP
SER
GLN
TYR
LEU
GLU
PHE
ASN
PRO

Figure 4-4. Contact map showing initial coordination and final coordination geometries of simulations A, B, and C. Contact maps for the original starting condition and final representative protein structure after REMD simulations are shown. The circles represent the position of the metal ions, and the indices are used to visualize where the metal ion stays or moves during the simulation. Metal ions are colored by the coordinating residues in the binding pocket.

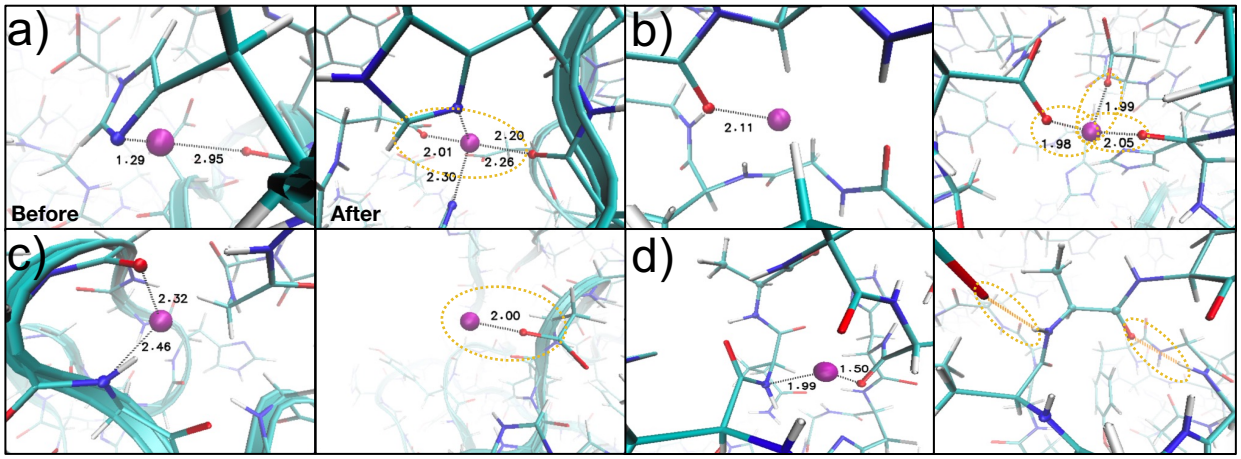


Figure 4-5. Metal binding pockets observed in REMD folding. Simulation snapshots for structure B showing the changing coordination environment before and after the REMD simulations have run. Several changes to the coordination environment of the Zn^{2+} ion (purple) with the protein nitrogen groups (blue) or oxygen groups (red) are shown. The coordination site in a) gains coordination partners to convert from a bidentate to tetradentate binding arrangement. Aspartate residues are significantly involved in coordination through b) increased coordination bonds with the carboxylate groups of the aspartate or c) aspartate involved in bonds that keep the monodentate metal-ion coordination without being loadbearing. d) Some coordination bonds convert to hydrogen bonds.

Across all simulations, the loss of histidine-Zn^{2+} bonds is likely because the protein was initially oversaturated with Zn^{2+} as an initial concentration of 8 wt% was selected to replicate the metal ion concentration found in the tip of the *Nereis virens* worm jaw.[13,61,372] Instead, the resulting wt% was roughly half of the initial concentration, at 4.1±0.08, 4.7±0.08, 4.8±0.17 for A, B, and C respectively (**Figure 4-4**). Further simulations beyond the scope of the current study could experiment with changing the initial concentration of metal ions present, as this is likely has a strong effect on the structure and mechanics of the protein.

Table S3. Final Number of MC bonds for Figure 3c.

Amino Acid	A-initial	A	B-initial	B	C-initial	C
Asp	3	24±0	3	32±0	3	23±0
Gly	7	2±0	18	3±0	7	0±0
His	21	11±0	19	13±0	20	10±0
Total Coordinated Amino Acids	40	42±0	53	56±0	40	40±0

To understand the role of hydrogen bonding in the structure, the absolute and change in the number of hydrogen bonds is plotted across the simulations (**Figure 4-3d, Table 4-4**). C had an increase in the number of hydrogen bonds and A and B had a decrease. An example of the conversion of a metal-coordination bond to a hydrogen bond in the protein structure is seen in **Figure 4-5d.** The trend in hydrogen bonding does not correspond to the SASA or number of metal-coordination bonds in the system. This further leads us to believe that that metal-coordination bonds are more structurally important than hydrogen bonds in the protein folding and compactness of the *Nereis* proteins.

Table 4-4. Hydrogen bonds for Figure 4-3d.

Initial	A	B	C
47	42.3±4.6	45.3±5.8	48±6.2

4.5 Mechanical tensile properties of protein

We find that the structural compactness of the various Nvjp-1 proteins (B>A~C) is independent from its properties under tensile SMD simulations (C>A~B) due to different underlying mechanistic principles. In the SMD simulations in **Figure 4-6a**, we found that C has a higher linear elastic modulus and yield strength, followed by B, then A. We attribute this trend to the nanostructural features of the protein (location of metal-coordination bonds, hydrogen bonds, secondary structure), rather than the measure of global compactness (SASA, number of metal-coordination bonds) discussed above. Though the exact quantitative influence of each nanostructural element cannot be clearly differentiated, three contributing mechanisms explain these mechanical differences. First, the trend of strength from C to B to A follows the trend in hydrogen bonds in **Figure 4-5-3d** rather than the metal-coordination bonds. Hydrogen bonds are also important for mechanical strength, and there are roughly twice the number of hydrogen bonds in the protein as there are metal-coordination bonds, resulting in a large influence of the hydrogen bonds on the mechanical properties of the protein. Second, the trend of SMD mechanical properties directly follows the trend in the amount of turn structure in each of the proteins, which decreases from C to B to A, and indirectly follows the amount of coil structure, which increases from A to B to C. Thus, the random coil likely contributes minimal mechanical resistance to the pulling, whereas the turn structure stabilizes the Nvjp-1 protein against mechanical disruption. Last, and of most novel to this work, C has an even distribution of load-bearing metal-coordination ions throughout its structure (**Figure 4-6b, c, S2**). This uniform distribution of metal ions plays an important role in providing resistance to rupture, especially because the detailed rupture mechanisms of the proteins are heterogeneous. As further evidence to this phenomena, **Figure 4-6a** shows that A and B have bonds breaking together towards the end, but C has a more uniform breaking of bonds through simulation time. Not only does this uniform distribution result in increased stiffness, but it also results in an increased toughness ~25-30% compared to simulation A or B. These results together indicate the importance of hydrogen bonding, secondary structure, and the distribution of metal ions on the mechanical properties of the Nvjp-1 protein in tension.

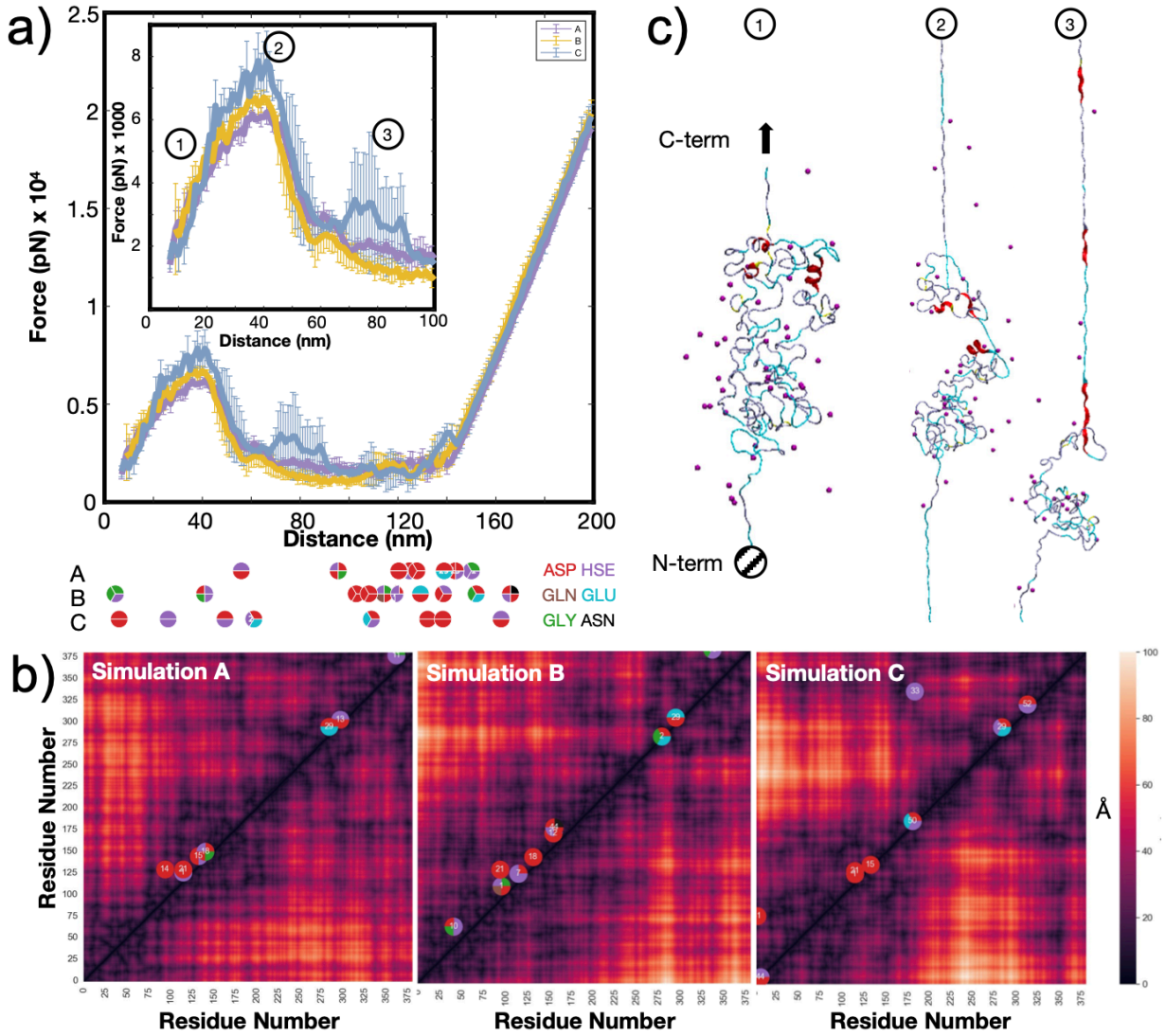


Figure 4-6. Steered molecular dynamics of Nvjp-1 proteins. a) Force-displacement behavior of Nvjp-1 proteins from SMD simulations at 1 m/s shows differences between the various Zn^{2+} conditions. Inset is an expanded view of the force-displacement data. Circles below show when each metal ion breaks coordination bonds to become non-loading bearing, color-coded by the initial coordinating residue types. b) Contact map of load-bearing bonds in simulations A, B, and C shows that C has metal ions that are more uniformly distributed through the protein. c) Screenshots of Nvjp-1 proteins show how the protein deforms for simulation C, when the N-terminus is fixed and the C-terminus is pulled.

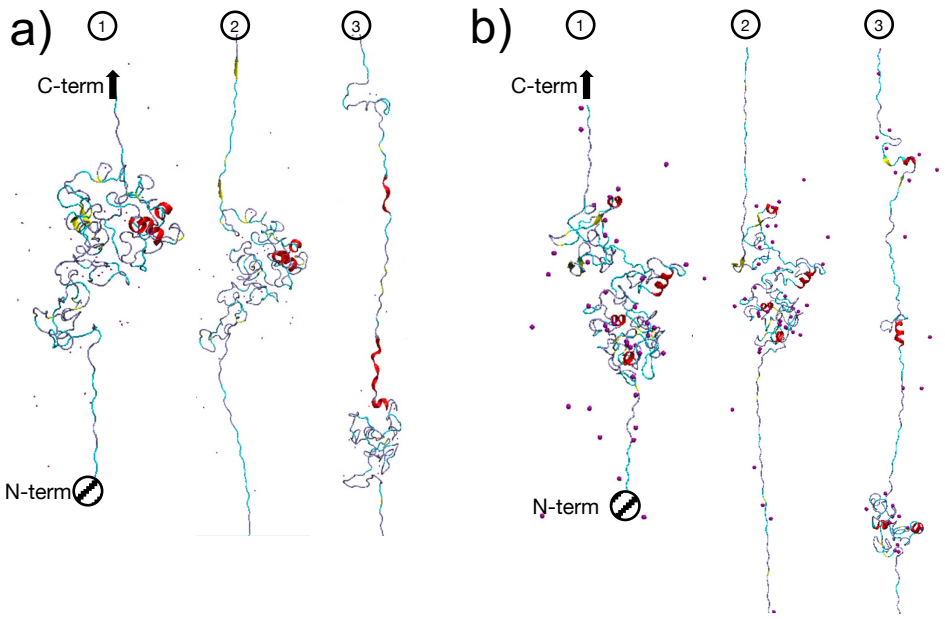


Figure 4-7. Steered molecular dynamics of Nvjp-1 proteins. Screenshots of Nvjp-1 proteins show how the protein deforms for a) simulation A and b) simulation B. Numbered circles correspond to the deformation snapshots in **Figure 4-6a**.

4.6 Discussions, Implications, & Conclusions

The objective of this work was to understand whether and how the initial spatial distribution of Zn^{2+} ions affects the structure and mechanics of the Nvjp-1 protein from *Nereis virens*, to elucidate why the *Nereis virens* exhibits high hardness. We found that while a uniform distribution of metal ions preserves much of the original structure during REMD simulations, Zn^{2+} ions in the beta regions help make the structure more compact by preserving or increasing the number of metal-coordination bonds present in the system in with metal ions clustered in the β sheet regions. However, these structural compactness trends are independent of the mechanical trends found for these proteins due to different underlying mechanisms contributing to each phenomena. While the number of metal-coordination bonds helps determine the compactness of the structures, the amount of hydrogen bonding, turn secondary structure, and uniform distribution of the metal ions result in the highest elastic modulus and toughness.

Our findings indicate that for proteins with high amounts of metal-coordination, the initial distribution of metal ions affects the resulting structure, which in turn affects the mechanics of proteins by changing the secondary structures and ion distributions present in the protein. These metal-ions greatly influence the protein-folding energy landscape, by altering the energy barrier the proteins must overcome during the folding. Though the REMD method should theoretically be independent of initial conditions, the pivotal role of this large quantity of metal ions and the practical computational resource constraints that would require an extremely long simulation time to overcome all the energy barriers imposed by the metal ions indicates that further simulations of metal ions in highly coordinated proteins must consider initial

localization. Further, the independence between the structural compactness and the tensile mechanical properties of the proteins in this work indicates that the protein folding process and mechanical unfolding process have different energy landscapes affected by different structural constraints. For the protein folding, the quantity of metal ions makes the structure more compact. For the tensile mechanical unfolding, the hydrogen bonding secondary structures that dominate the protein structures have a large mechanical effect, with additional contributing effects from the distribution of metal ions. This insight was especially surprising, but contributes to growing literature suggesting the dissipative as opposed to load-bearing nature of the metal-coordination bonds. Understanding the difference between these underlying energy landscapes could further present important insight into how metal ions can be added in different quantities or in different locations to tune specific folding or rupture behavior.

The work presented here likely has broader implications on the role of metal-ion crosslinks on the structure and mechanics of proteins. If the exact relationship between the location of the metal ions and the structure or mechanics can be derived, we can not only gain insight into the biological mechanisms but also design new synthetic materials. While this work has revealed a number of initial parameters, such as the role of Zn^{2+} ions in helping the Nereis worm jaw become compact, or the role of the turn secondary structure in increasing the strength of Nereis, several future directions emerge to further building our understanding. Specific protein structures can be designed to compare the contribution of hydrogen bonds to metal-coordination bonds, because the large, complex nature of this protein makes it difficult to parse the exact contribution of hydrogen bond versus coordination bonds to the strength of the protein. Improvements in the force field parameters for metal ions, especially for structures where there are a significant number of metal ions such as in Nvjp-1, could enable more accurate mechanistic insight. Simulations doing compression tests such as nanoindentation on bio-nano composites[373–375] may also help reveal insights into the compactness and hardness of the structure.

4.7 Materials and Methods

REMD is used to explore the structural effects of metal ion localization on the Nvjp-1 protein. REMD enables an efficient computational search for likely protein structures by overcoming kinetic trapping in local energy minima during protein folding.[376] The starting protein structure is derived from Chou et al.,[47] who used REMD to fold the primary protein sequence in implicit solvent. To simulate the system at pH 8, the protonation states of Nδ1H for histidine (His), Oδ2(–) for aspartic acid (Asp), and Oε2(–) for glutamic acid (Glu) were adjusted accordingly. Simulations were implemented with Nanoscale Molecular Dynamics (NAMD)[377] and all simulations utilized the CHARMM22 force field[229] with a 2 fs timestep under NVT conditions with Langevin dynamics. Zn^{2+} ions were added to a concentration of 8 wt% to replicate the metal ion concentration found in the tip of the *Nereis virens* worm jaw.[13,61,372] Counterions of Cl^- were added to balance the charge. A TIP3P explicit water box with a 16 Å skin was used, and periodic boundary conditions and the Particle Mesh Ewald full system electrostatics method were applied.

After a brief energy minimization using NAMD's conjugate gradient method to avoid bad contacts and a 30 ns NPT equilibration to achieve correct pressure of 1 atm, 96 replicas were

used for the REMD simulation with a temperature range of 300K – 480K and an exchange time of 200 fs to allow for the system relaxation under an NVT ensemble. The simulations were collectively run for ~2 μs across all replicas. The REMD simulation in the trajectory at 300K was analyzed using the K-means clustering algorithm in the MMTSB toolset.[378] This algorithm clustered the structures based on conformational similarity based on a root-mean-square deviation (RMSD) within 2 Å. Three representative structures within the lowest energy cluster were identified for subsequent analysis. The Visual Molecular Dynamics[379] STRIDE algorithm[380] was used to quantify the structural characteristics of the representative structures. The protein contact map was calculated using the Protein Contact Maps tool, and the metal ions were added to the contact map based on visual analysis.[381]

Representative structures from the lowest energy clusters of the different simulation conditions were solvated in an TIP3P explicit water box extended by ~60 Å for SMD simulations to account for deformation. After preliminary simulations suggested minor differences between the different pulling orientations, the terminal Cα atom on the N-terminus was fixed and the terminal Cα atom on the C-terminus was selected as the SMD atom. SMD data were collected every 10 ps at a pulling rate of 1 m/s. The simulation was run until the protein was fully extended. The coordination bond was defined as broken when the distance between the metal ion and coordinating polar atom exceeded 3 Å.

CHAPTER 5

MACHINE LEARNING APPROACHES TO DESIGNING MECHANICAL PROPERTIES IN COLLAGEN BIOMATERIALS

5.1 Significance Statement

In the previous chapters, we characterized the mechanical properties of metal-coordination bonds in polymers and proteins using experiments and simulations. We used multiscale simulation methods including density functional theory at the quantum scale, molecular dynamics at the atomistic scale, and coarse-graining of monomers at the macroscopic scale. These multiscale models allowed us to connect several time and length scales of materials chemistry with the resulting macroscopic materials properties. A natural extension of and upcoming frontier in these multiscale models is the use of big data approaches to characterize, design, and predict the properties of materials. Specifically, machine learning has emerged as a frontier across several disciplines to analyze large datasets and develop design principles without requiring an underlying knowledge of detailed chemical interactions.[107,334,382] In this chapter, we explore how machine learning-based computational methods can be applied to predict the mechanical properties of biomaterials from basic chemistry.

Machine learning methods require large datasets for model training. As such, in this chapter, we will focus on collagen for mechanical function, because there is more available data on collagen mechanical properties than on metal-coordination bonds mechanical properties. Collagen is the most abundant structural protein in humans and is therefore often used in biomedical applications for tissue repair and regeneration. Designing *de novo* collagen to maintain its structural integrity *in vivo*, important for its mechanical performance and subsequent utility, remains a challenge today. In this chapter, we develop a deep learning framework to generate new collagen sequences with desired thermal stability and validate our deep learning framework using both simulation and experiment. Given this validation, we discover key insights into the prevalence of amino acids in collagen triple helices and find a mechanistic relationship between our simulations and experiment. We then show that alternative machine learning models such as transformers can also be applied to make similar predictions.

The frameworks developed here enable researchers to create new collagen sequences with desired mechanical stability for biomedical applications using high-throughput machine learning methods. Further, these frameworks can be used for other biomaterials fields, such as metal-coordination for mechanical function, once large datasets are developed to enable these methods.

5.2 Introduction

Collagen is the most abundant protein in animals and is found in the extracellular matrix of skin, tendons, bone, and vasculature, as well as other tissues. The term “collagen” encompasses a family of at least 29 glycoproteins with common features responsible for the outstanding properties. Repeat units of glycine-X-Y (GXY) dominate the sequences, where the X and Y amino acids are usually occupied by proline (about 28% of the time) and hydroxyproline (about 38%).[383] These GXY sequences adopt a left-handed polyproline type-II helical conformation which, after forming trimers, folds into a triple α-helical structure called tropocollagen, the basic structural unit of collagen.[384] Tropocollagen is typically 300 nm long and 1.5 nm in diameter and assembles into hierarchical collagen structures including fibrils and fibers (**Figure 5-1a**).[385–391] This hierarchical structure enables collagen to provide significant mechanical capacity under physiological conditions, exhibiting a tensile modulus of 0.2 to 0.86 GPa, while maintaining elasticity in the human body.[392–399]

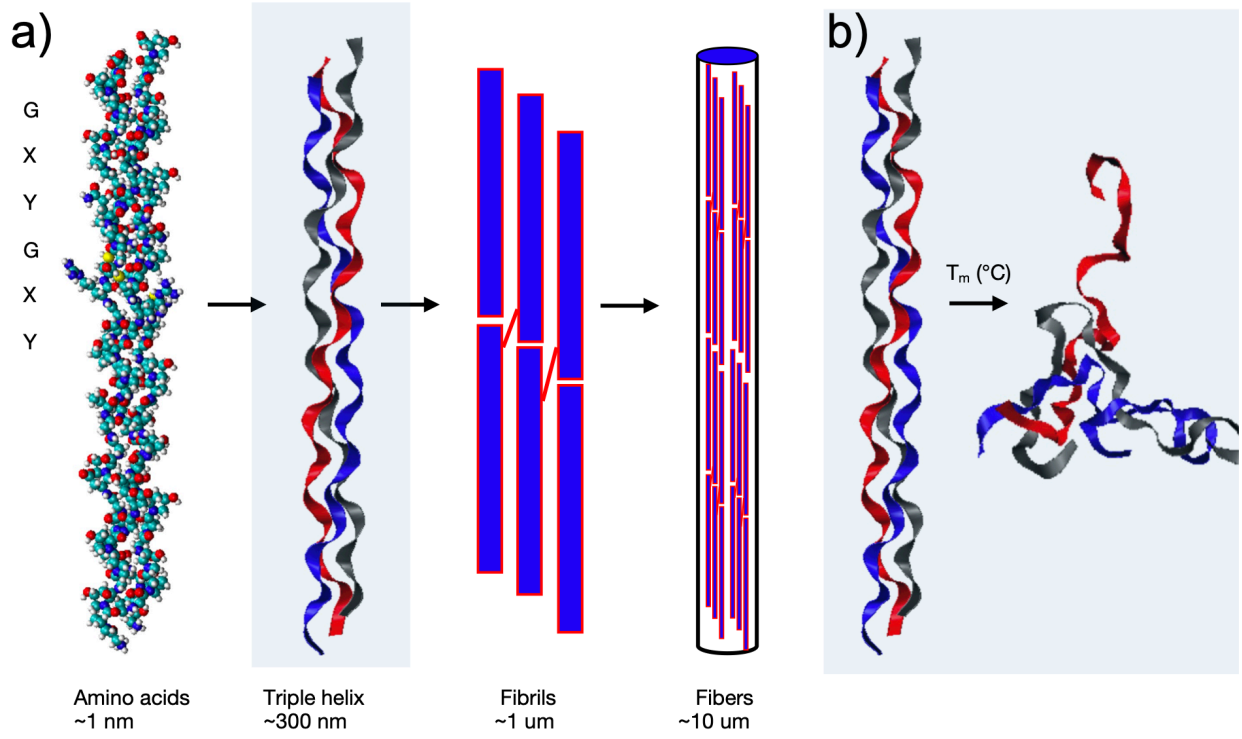


Figure 5-1. Hierarchy of collagen helps maintain its structural integrity. a) Collagen amino acid primary sequence, often in the form of G-X-Y repeat triplets, form a larger chain. The three chains come together to form a triple helix, characteristic of collagen, which is also known as tropocollagen. The tropocollagen assembles into larger fibril and fiber units. b) This work focuses on thermal stability of tropocollagen. Thermal stability is characterized by the T_m value, which is midpoint temperature of the denaturation process of the triple helix of tropocollagen to a disordered state. Once collagen is not in a triple helix, it no longer contributes to mechanical stability of the larger fiber.

Given this remarkable self-assembly process, the resulting mechanical properties, the inherent biocompatibility, collagen-based biomaterials are routinely sought for *in vivo* tissue repairs, drug delivery systems, and other biomedical applications.[400,401] However, designing collagen to assemble *in vitro* to emulate the structural hierarchy and thermal stability of collagen *in vivo*

remains challenging and limits the widespread use of collagen as biomaterial constructs. Therefore, most of the collagen used today has a reduced triple-helix content, and thus reduced thermal stability and mechanical properties, which results in rapid degradation *in vivo*. To overcome these challenges, synthetic collagen-based biomaterials are often stabilized via chemical crosslinking and related methods,[402] which while effective in extending longevity *in vivo* can negatively impact biological responses to collagen and alter the mechanical properties of the materials.

Given the importance of collagen's structural integrity for its mechanical function and thermal stability, one useful metric is the melting point (T_m), defined as the midpoint during the temperature window in which the collagen triple helix unfolds (**Figure 5-1b**).[403] The thermal stability of collagen from different biological species or *de novo* collagen designs has been characterized experimentally.[404–409] Computational studies have used molecular dynamics or coarse-graining to determine sequence-structure-function thermal and mechanical properties in collagen across its different length scales.[388,410,411] These prior research efforts have made significant progress in understanding how mutations in the primary sequence affect the thermal stability of collagen. However, these approaches are computationally expensive and limited in the possibility to explore vast variations of sequences and mutations.

5.3 Discovering design principles of collagen molecular stability using a genetic algorithm, deep learning, and experimental validation

A predictive framework that facilitates the *a priori* design of collagen sequences with specific T_m values without prior knowledge of chemical interactions would enable the efficient design and subsequent synthesis of thermally stable collagens for specific applications. Such a framework for discovery could significantly propel the field of collagen-based biomaterials forward. Towards this goal, equations were developed to predict T_m values of collagen triple helices based on local interactions between different amino acid chemistries in collagen tripeptides following a GXY triplet ordering.[403,412,413] In addition, an algorithm (Scoring function for Collagen-Emulating-Peptides' Temperature of Transition SCEPTTr) was developed to predict the registry and T_m values of synthetic collagen-based triple helices.[414] Other recent approaches have been based on machine learning, which has emerged as a useful tool in the analysis of large datasets to help develop design principles for biological materials without knowledge of underlying biological interactions.[415,416]

Here we report the development of a machine learning model to design *de novo* collagen sequences with desired T_m values. We apply a self-evolutionary algorithm, 1D convolution, bidirectional long short-term memory (LSTM), and dropout features to predict T_m values of existing collagen sequences.[417] To demonstrate the predictive power of our approach, we use molecular dynamics (MD), circular dichroism (CD) spectroscopy, and differential scanning calorimetry (DSC) to verify the T_m values of a few of our *de novo* collagen sequences. From this approach, we are able to derive two new insights. First, our model has the highest predictive accuracy for *de novo* collagen sequences with strong triple helix folding as measured through the triple helical quality (ratio of positive to negative peak intensity (RPN)) value extracted from CD spectroscopy, and we demonstrate a correlation between hydrogen bonding in the triple

helix found through MD and RPN value.[418] Second, given the high-throughput nature of our work, we identify key collagen triplet amino acid sequences that especially contribute to the thermal stability of collagen. These GXY triplet sequences should inform the design of next generation thermally stable collagen sequences. The goal of this work is to demonstrate the use of this generative algorithm in suggesting *de novo* collagen sequences with desired T_m values, thus contributing to a more efficient method of designing new collagen-based materials with tailored properties.

5.3.1 Collagen Dataset

We collected 566 collagen sequences with reported T_m values from a survey of literature (see **Appendix S5-1**). These available melting temperatures were collected from PubMed, Web of Science, Scopus, Directory of Open Access Journals (DOAJ) Google Scholar.[385,411,413,419–438] Experimental thermal stability data sets for the observed T_m values (in °C) for the Gly-X-Y tripeptide units in triple helical collagens-like peptides are based on host-guest peptides and are integrated to produce an algorithm for predicting global melting temperatures. These experimental results are expanded further using predictions of Tm values of host-guest sequences from Persikov et. al.[413] The distribution of the dataset is presented in **Figure 5-2** where sequences have experimentally measured melting temperatures ranging from a few degrees C to 70°C with the mean at ~30°C. The data shows a normal distribution which is used in the machine learning model (**Figure 5-2c**). Outliers in the data, specifically the negative T_m value sequences, are generally from extrapolated experimental data.[439]

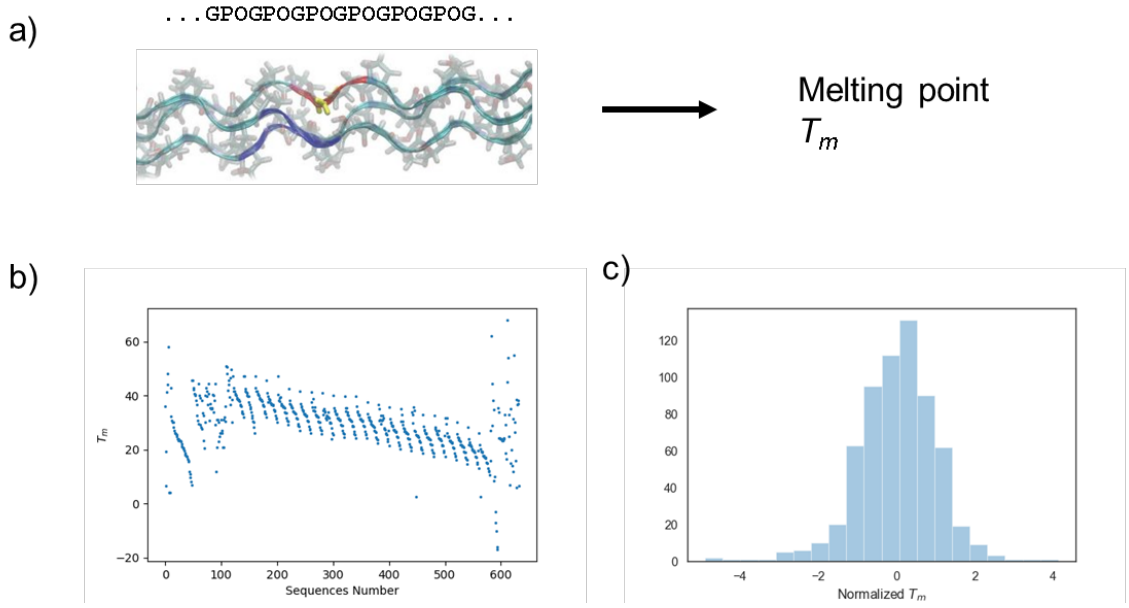


Figure 5-2. Distribution of data from literature, based on experimental results. a) Overview of the problem studied here, to predict the melting point Tm from the sequence of collagen molecules. b) Experimental melting temperatures collected. c) Normalized T_m value distribution. Thermal stability data sets for observed Tm values for Gly-X-Y tripeptide units in triple helical collagen-like peptides are integrated here to produce an algorithm for predicting global melting temperatures. Data from [385,411,413,419–438].

5.3.2 NLP LSTM end-to-end neural network and model training

We provide a summary of the deep learning model reported here, as shown in **Figure 5-3**. This model is named ColGen to represent a collagen sequences generator which is capable of generating new type I collagen sequences with known T_m values. A natural language processing (NLP) model was adopted.[440,441]

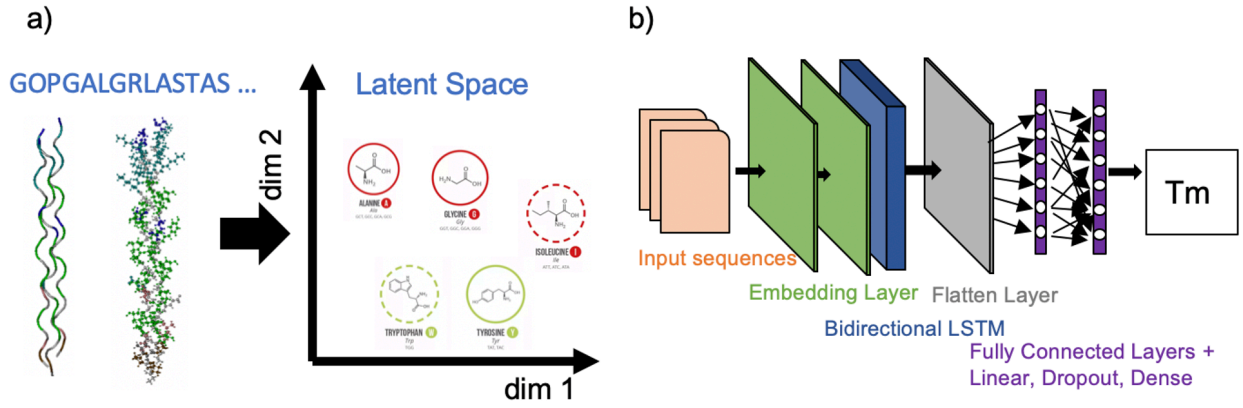


Figure 5-3. Overview of machine learning model. a) We design a deep learning network to discover hidden features of collagen sequences by introducing embedding layer. b) The structure of our deep learning model starts at an embedding layer, followed by two 1D convolution layers, then we flatten all the features and send them into a fully connected layer for regression to determine T_m value.

Tokenization is used, reflecting a common approach when processing the raw texts or sequences of symbols. Tokens can be considered as the fundamental building blocks of our data which was represented in single letter code for each amino acid.[442,443] Tokenization is first applied to the collagen sequences before passing them through the deep learning model. Every collagen sequence was first encoded into a series of digital tokens, that is, every amino acid is treated as a unique number from 1 to 21 for all the 20 essential amino acids and one nonessential amino acid, hydroxyproline, found in collagen.

The tokenized sequences were passed through an embedding layer which is able to recognize the relationship between tokens during the training process. After the embedding layer, data flows through a 1D convolutional layer to harness the internal features from the input. The data is then routed into bidirectional LSTM layers to learn all hidden features from each collagen sequences. Finally, a fully connected neural network - composed by two dense layers – is used to ultimately output the predicted T_m value as a scalar value. This information provides an end-to-end model that relates a sequence of amino acids of varying lengths to its T_m. The neural network features a total of 49,041 trainable parameters.

The dataset was randomly split into training dataset testing dataset, where 80% of the data is used for training and the other 20% is used for testing to examine the ability of prediction of our model. To ensure reproducible output of the exact results, the random seed was set where possible (such as using the tf.random.set_seed, train_test_split(X, y, test_size=int, random_state=int command).

The model is trained on a Xeon workstation with a GTX-3090 GPU, for 200 epochs. It is worth noting here, the well-trained model can be deployed in a laptop or desktop computer without further requirement of GPUs.

We begin the analysis by training the machine learning model. We find that ColGen demonstrates good predictive accuracy of T_m values in the testing set (**Figure 5-4a**). The data shows that testing data is generally well predicted, and a large range of temperatures can be predicted by the model. In addition, the training and validation error remain consistent at a mean squared error of ~0.2 after 120 epochs (**Figure 5-4b**).

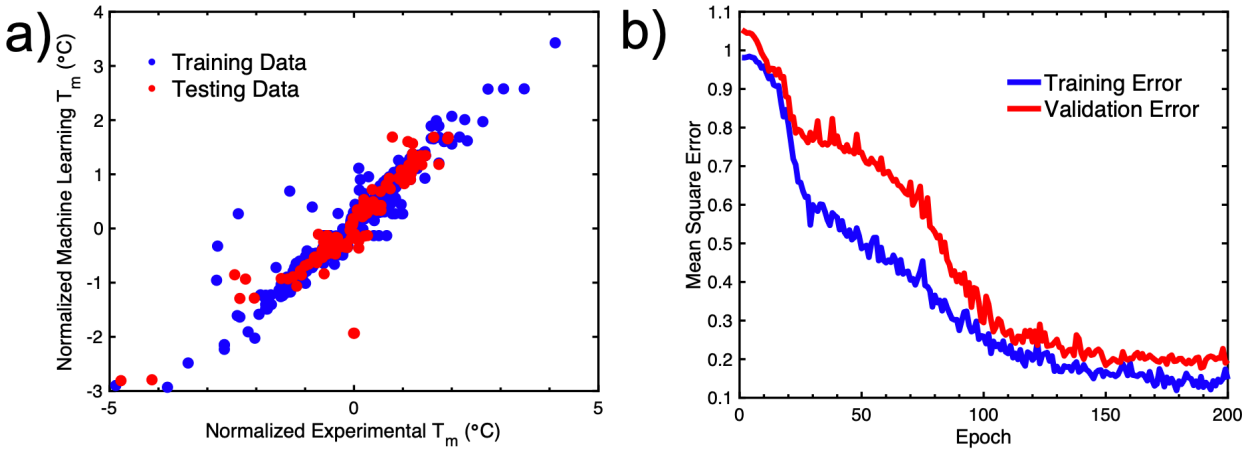


Figure 5-4. Predictive accuracy of ColGen, and training performance. a) Data comparing training with test set demonstrates a 95% confidence interval. Plotting R^2 of training / testing / generation. b) Training and validation error over epochs demonstrate a well fit model. The validation and training errors reach a plateau around 150 epochs.

5.3.3 Incorporating a genetic algorithm to generate new collagen sequences

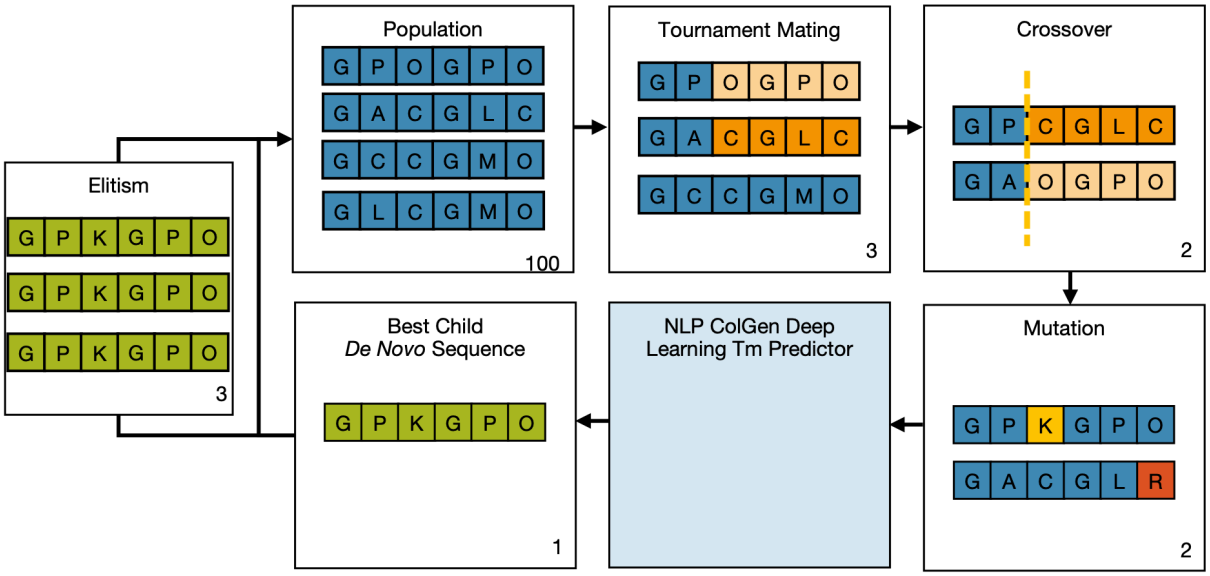


Figure 5-5. Machine learning based genetic algorithm used. Three sequences are randomly selected from a randomly generated population based on the dataset collagen sequences. The three sequences undergo

tournament mating to identify the two best parents, which undergo further crossover and mutations within their sequences to produce children offspring. The resulting children are then evaluated with the ColGen deep learning T_m predictor and the best child matching the desired T_m value objective function is output. If elitism is implemented in the model, the child is overrepresented in the initial population to help preserve its general sequence features. Numbers in bottom left represent number of sequences in each stage.

We next developed generative model, implemented as a genetic algorithm, to generate new collagen sequences (**Figure 5-5, Figure 5-6)**.[440,444] This model is named ColGen-GA to represent a collagen sequence generator, which is capable of generating new homotrimeric type I collagen sequences with specific T_m values. ColGen-GA builds on the predictor model ColGen discussed above.[417]

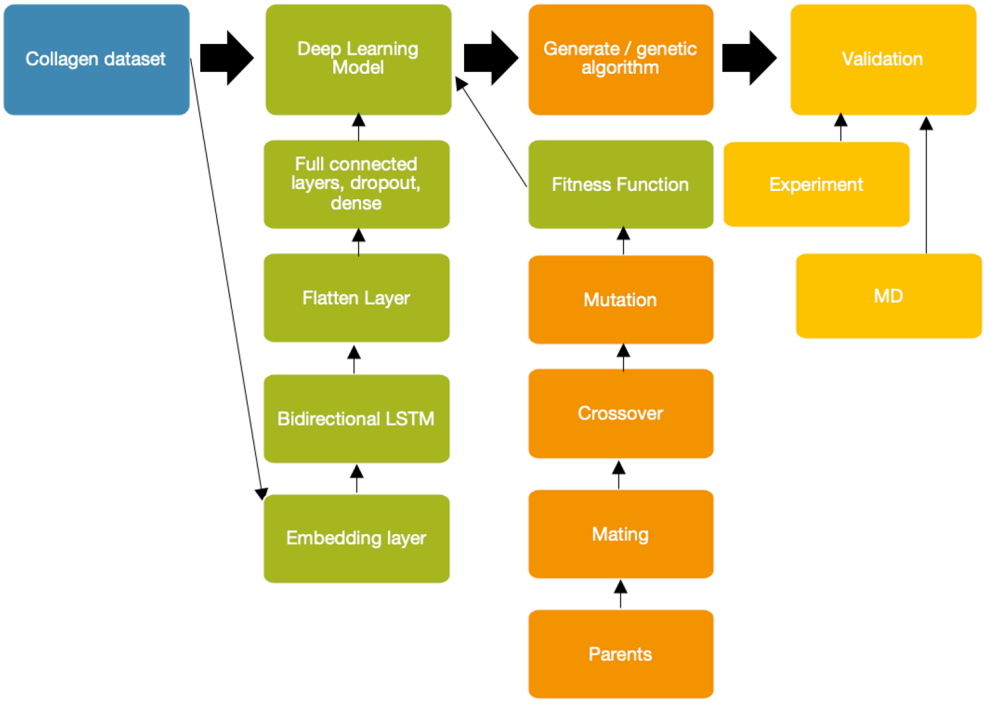


Figure 5-6. **Model work flow.** Model work flow from data collection, to generation, to prediction, to validation.

Each generated collagen sequence is optimized to meet the objective function of the algorithm, which is a T_m value of choice. In this work, the T_m values are selected as 22 ºC or room temperature, and 37 ºC or body temperature, as these are the two most relevant temperatures for bioengineering applications.

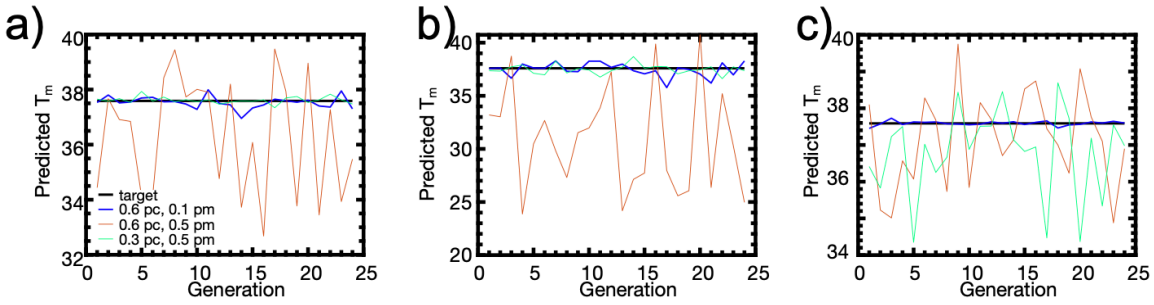


Figure 5-7: Effect of mutation rate, crossover rate for different population numbers on solutions to GA. For population sizes a) 100, b) 25, c) 200, increased mutation rate and crossover rate increases the range of the T_m

values that is generated via the generative algorithm, while resulting in fewer solutions at our desired T_m. In contrast, increasing population number increases the number of solutions at our desired T_m value, with a significant downside of a longer resulting runtime (4 mins for 100 population, 2 mins for 25 population, 15 mins for 200 population). Based on our goals of balancing diversity of solutions, with run time and number of solutions at the target T_m, we selected the red line drawn in figure a.

In the genetic algorithm, an initial population is randomly selected from the existing collagen dataset. Three parents are further randomly selected from the initial population to undergo tournament mating, where the two parents with the closest T_m value to the desired T_m value are selected. The T_m value is calculated from the ColGen model above.[417] These parents then undergo crossover and mutation to produce "children" sequences. The crossover and mutation rate are optimized to ensure that there is sufficient sampling of solutions, which prevents genetic drift while not leading to a loss of good solutions. This optimization is a balance between the number of generations required to reach convergence versus the number of unique sequences generated (**Figure 5-7**). The child with the closest fit to the desired T_m value is then selected as the final output. This whole process is repeated over several iterations or "generations", until we reach a converged state around the desired T_m value (**Figure 5-8**). Further, we tested generation methods with "elitism," which is where the best children are overrepresented in the initial population, such that the better traits stay in the genetic pool for longer, and "randomness," which is where the initial population in the next generation is unrelated to the children from the previous generation. While elitism helps ensure that the quality of the generative algorithm does not decrease over time, it has a disadvantage of converging on a local minimum rather than finding the best solution. The selection of population sizes, mutations and crossover frequencies, and elitism implementation are critical in generating a wide number of sequences without losing the best features found in the generations. The specific parameters used in this model are listed in **Table 5-1.** Due to computational modesty, the genetic algorithm model can also easily be deployed on a laptop or desktop computer without further requirement of GPUs.

Table 5-1: Parameters in the genetic algorithm model.

Parameter	Value
Seed value	55
Crossover Percent	0.6
Mutation Percent	0.1
Population Size	100
Contestants for tournament mating	3
Number of parents	2

The ColGen-GA model can quickly reach the desired T_m value, and maintain that value over several generations. In **Figure 5-8**, the desired normalized T_m value of 0.9, corresponding to 37 ^oC, is reached almost immediately. The convergence is even faster when elitism is implemented.

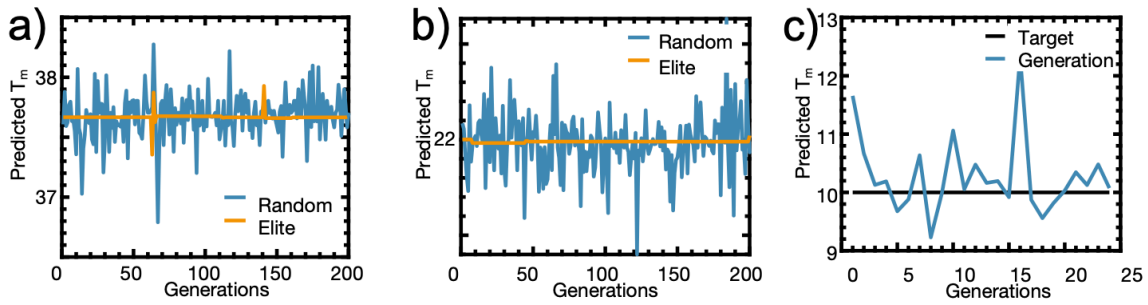


Figure 5-8: Convergence of genetic algorithm around designed Tm values for a) 37 °C and b) 22 °C target T_m values. c) Using the same parameters as for 37 °C and 22 °C, we are also able to generate sequences at a target temp of 10 °C. However, fewer sequences are generated given that 10 °C is two standard deviations outside of our training set. As such, it is difficult for the ColGen algorithm to generate several sequences in this range. The standard deviation for the generated sequences at 37 °C is 0.2 °C, 37 °C is 0.15 °C, and 37 °C is 0.6 °C.

5.3.4 Prediction of the effect of collagen mutations on T_m values

Because the ColGen model enables fast characterization of the T_m values of different collagen sequences, the model was used to understand how mutations and chain length affect T_m value. ColGen enables a rapid search of these effects. $(GPO)_{10}$ was used as a standard comparison, referred to here as the "pristine sequence," as it has the highest known thermal stability value in literature [404,430,445,446] due to stereoelectronic effects from hydroxyproline [447–449]. Mutations were made in either the G, P, or O position, where either the G, P, or O positions were replaced with another amino acid, and this process is repeated over all amino acid substitutions. The resulting T_m values from each of these calculations are then averaged.

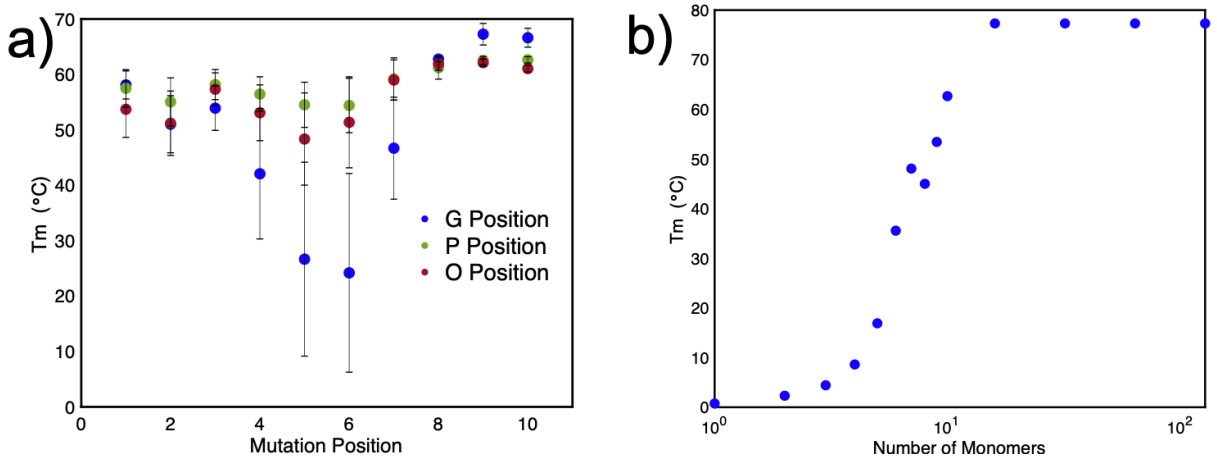


Figure 5-9. Characterization of the effects of various types of mutations, predicted by ColGen. a) T_m values of mutations in G, P, or O position demonstrates that mutations in the middle of the sequence are the most destabilizing for T_m values. Mutations in the G position are the most destabilizing to the peptide. Error bars indicate standard deviation of all amino acids that were mutated. b) Thermal stability as a function of collagen sequence length, where length is number of repeat units (GOP) demonstrates that there is a critical length at which the T_m can no longer be increased significantly. This critical length is consistent with other studies.[450]

Based on the GolGen model, mutations in the middle of the pristine collagen sequence have the greatest destabilizing effect on collagen T_m values (**Figure 5-9a**). This suggests the

presence of a critical transition location along the length of the sequence, potentially near the midpoint of the sequence or at least away from the chain ends depending on overall length, that is critical in holding the full chain length together. Further, mutations made along the first several residues at the N terminus of the collagen sequence are more destabilized and have a lower T_m value than mutations made along the last several residues at the C terminus. This indicates a directionality along the sequence and is validated by experimental data which shows that the N-terminal regime is required for the trimerization of other triple helical collagens.[451,452] This is in contrast to other work on fibril-forming procollagens which suggests that type 1 collagen molecules in vivo have a C-terminal that is responsible for chain selection and trimerization.[453–457] This difference between our model and procollagen fibrillar formation results is likely because several peptides included in this training data may have been folded in N-to-C direction with a nucleation domain at the N terminus.[458,459]

The effect of chain length on T_m was also measured. As shown in **Figure 5-9b,** the model faithfully captures that T_m values increase upon increasing number of amino acids (triplets) due to increasing hydrogen bonding between triplets. However, there is a limit to the increase in thermal stability that is also captured by the model. This leveling off is achieved at about 14 triplet repeats at around 80C and is consistent with experiments and a thermal stability prediction algorithm developed by Persikov et. al.[413] Similar intrinsic strength limits have been found in hydrogen bonded alpha helix and beta sheet structures.[460–462]

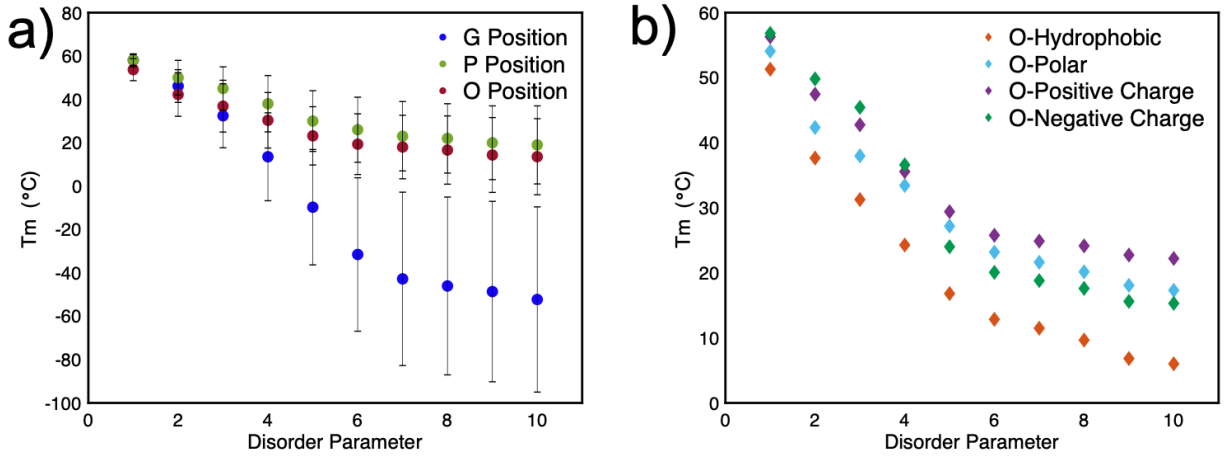


Figure 1-10. Characterization of effect of disorder on Tm, as predicted by the model. A) T_m values of disorder arranged by G, P, or O position confirm that increasing mutations along the chain decreases thermal stability of the triple helix. Error bars indicate standard deviation of all amino acids that were mutated. b) T_m values of disorder in the O position demonstrates that initial mutations to polar, positive charged, and negative charged amino acids confer the same degree of stability in the molecule. However, upon increasing mutations, polar amino acids are the least destabilizing to the triple helix, suggesting that they should be used for bacterial expression of collagen where expression of O is not possible.

To quantify how the increasing number of mutations affects the thermal stability of collagen, we define the term “disorder parameter” which means the number of repeating triplets with mutations in the G, P, or O position compared to the pristine sequence. Thus, increasing disorder parameter increases deviation from the pristine sample. As expected, increasing disorder decreases the thermal stability of the pristine sequence consistent with experimental data (**Figure 5-10a**).[413]. Further, disruptions in glycines are the most destabilizing. This is consistent with experimental findings that disruptions in glycine severely impact stability and

often constitute disease states, though we cannot directly correlate the position of our glycine mutations to the position of naturally observed mutations due to the shorter sequence length we employ in this study.[439,463–465]

This model could be useful in informing sequence design of bacterially-produced collagen, which enable larger production of tailored collagen sequences.[407,424,466] Given that bacteria are unable to express hydroxyproline for bacterially-produced collagen, the machine learning model developed here could also serve as a tool to predict which amino acids may help as a replacement to hydroxyproline. **Figure 5-10b** demonstrates these results, showing that positively charged, negatively charged and polar amino acids are only slightly destabilizing compared to $(GPO)_{10}$ up to a certain extent, but positively charged amino acids (R, H, K) are the least destabilizing compared to other types of amino acids if significant mutations are introduced in the O position.

5.3.5 De novo sequences with desired T_m values

From the ColGen-GA, we are able to produce several *de novo* sequences within our desired temperature range (T_m values of 22 ºC and 37 ºC). We selected two sequences in each temperature for further validation (CP1 and CP3 for 22 ºC, CP2 and CP4 for 37 ºC), as well as another *de novo* sequence generated from a different generative algorithm where the collagen primary sequence is not required to be in a G-X-Y order (CP5). **Table 5-2** shows strong agreement between our initial prediction of the T_m value and the T_m value found using experiment within a few degrees centigrade. Temperature sweep experiments revealed that the T_m values for the CPs and Type I collagen control as measured by CD and DSC were in good agreement with those predicted by ColGen-GA (**Table 5-2, Figure 5-11, Figure 5-12**). The slight difference between CD and DSC is attributed to the higher heating rate in DSC experiments.

Table 5-2. De Novo Collagen Sequences. Summary of names, amino acid sequence, and T_m values of samples studied listed in order of increasing to decreasing T_m value.

Name	**Sequence**	**Method**	**ColGen T_m** (°C) **MODEL**	**CD T_m** (°C) **EXP**	**DSC T_m** (°C) **EXP**	**T_m calculator**[2] (°C) **MODEL**
Col. Type 1	Bovine Collagen	Control	-	40.6	40.9	-
Std.	GPOGPOGPOGPOGPOGPOGPOGPOGPOGPO	Reference	62.0	-	-	63.8
CP5	GPOGPOGPOGPOGPOGPPAGPOGROGRO	Previous algorithm[1]	46.6	21.5, 40.4, 60.9	26.5, 44.9, 62.8	22.2[3]
CP4	GYOGPOGPOGKOGPOGKOGPOGPOGPHGPM	Random	37.7	41.2	42.8	40.6
CP2	GPOGPOGPRGMOGPOGPOGPOGPO	Elitism	37.3	35.4	36.4	38.5
CP3	GPOGPOGDOGATGPOGRCGPQGPOGPOGPO	Elitism	22.0	20.8	22.6	21.1
CP1	GIAGPAGPOGDAGPOGPOGPOGPO	Random	22.2	18.6	20.4	25.0

CP6	GVMGWGGALGYHGERG MNGHTGND	Previous algorithm[1]	-3.3	Does not form stable helix	-76.2
CP7	GEIGEVGSHGVNGHEGGF GYGGMGGG	Previous algorithm[1]	-26.6	Does not form stable helix	-83.0

[1] These *de novo* collagen peptides were generated from a previous genetic algorithm not discussed in the section. Their T_m values however were predicted from ColGen, and as such, are useful for understanding the validation of the ColGen model.
[2] T_m calculator prediction from the work of Persikov et al.
[3] T_m calculator prediction from the work of Persikov et al. is unable to calculate T_m values for sequences that do not follow $(GXY)_n$ formatting.

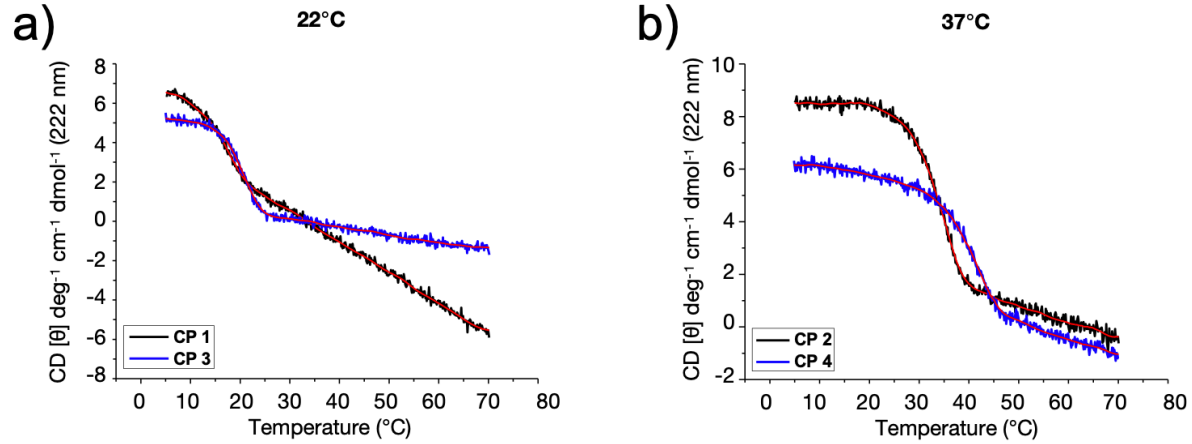


Figure 5-11. CD temperature scan at 222 nm for collagen peptides demonstrating triple helix structure. a) 22 °C peptides CP1, CP3, and b) 37 °C peptides CP1, CP3. Scans at 1 °C/min with sampling every 0.1 °C indicate that de novo peptides have T_m values within a couple of degrees of the target T_m.

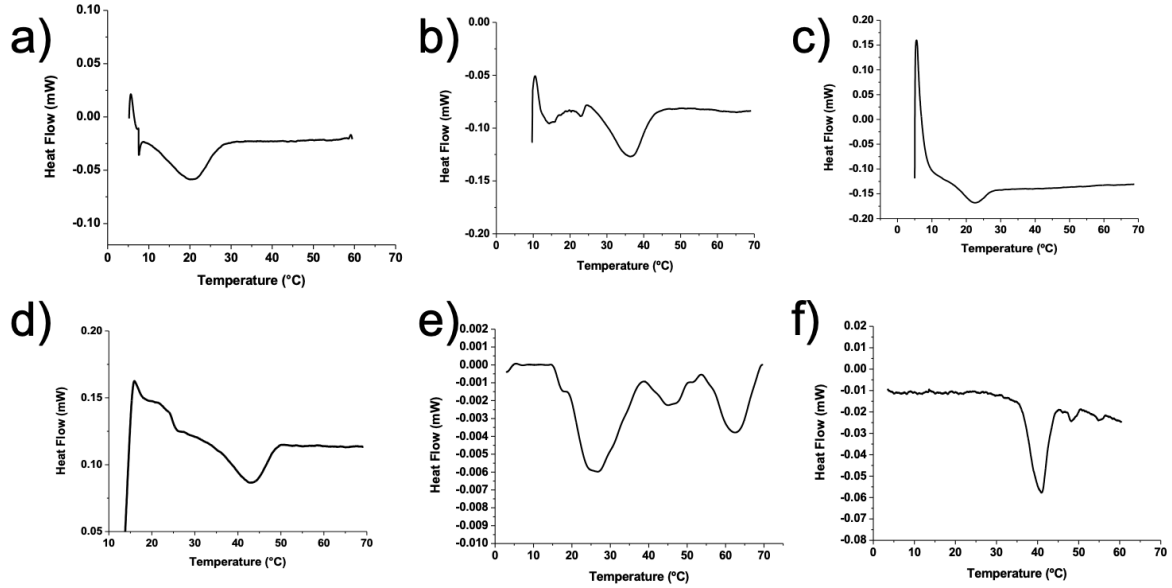


Figure 5-12. DSC thermograms of Type I collagen and de novo peptides. DSC temperature scans for a) CP1, b) CP2, c) CP3, d) CP4, e) CP5, and f) Type I collagen for reference. DSC temperature scans are consistent with the T_m values predicted for the *de novo* peptides.

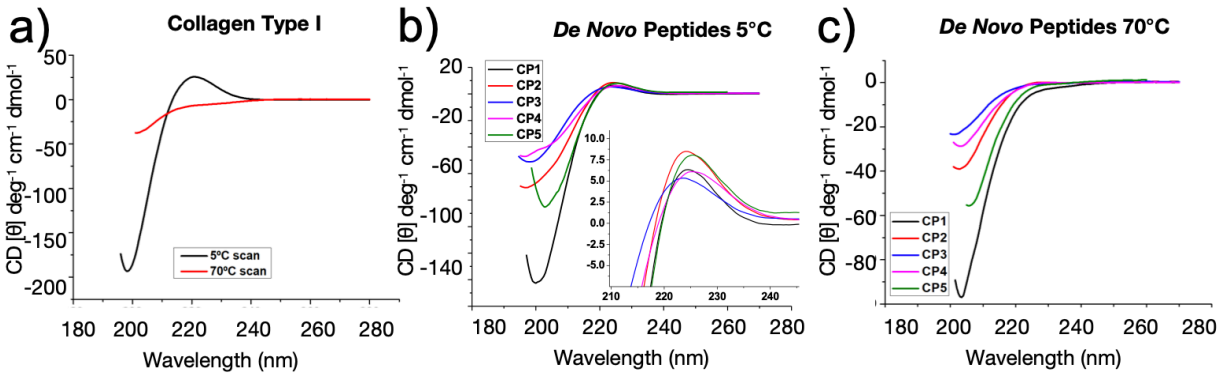


Figure 5-13. CD wavelength scan at 222nm for collagen peptides demonstrating triple helix structure. a) type I collagen as a control at 5 °C and 70 °C, and *de novo* peptides and b) 5 °C and c) 70 °C. *De novo* peptides demonstrate same characteristic behavior as type I collagen sequence, indicating that they have a triple helical structure. Both the type I collagen and *de novo* peptides denature at 70 °C.

Table 5-3. Ellipticity values of de novo collagen sequences for the maximum and minimum in the 5 ºC scan, the absolute value of the RPN and the location of the minimum in the 70 ºC scan for all tested samples.

SAMPLE	Max @5 ºC (222nm)	Min @5 ºC (196-8 nm)	RPN (abs. value)	Min @70 ºC (nm)
Col. Type I	25.79	-193.2	0.133	200.5
CP1	6.82	-133.70	0.051	201.3
CP2	8.52	-70.64	0.121	200.4
CP3	5.31	-51.14	0.104	200.7
CP4	6.09	-47.2	0.129	201.2

The CD spectra of the *de novo* CPs show that the CPs are able to form triple helical structures (**Figure 5-13a,b**). The CPs and the control follow a standard CD triple-helix forming collagen spectrum. There is a clear positive signal at 222 nm in the 5 ºC wavelength scan (**Figure 5-13a,b**), related to the presence of triple-helix, which disappears in the 70 ºC scan at which collagen denatures (**Figure 5-13a,c)**. The ratio of positive signal at 222 nm to negative at 196-198 nm (RPN) serves as a concentration-independent measurement of the quality and quantity of triple-helix formation. An RPN value of 0.133 for Type I collagen is the highest compared to all the CPs (**Table 5-3**), indicating that it forms the best triple helix, as expected CP4 exhibited the highest RPN ratio (0.129) of all four CPs, followed by CP2 (0.121) and CP3 (0.104), values similar to those of Type I collagen. In contrast, CP1 and CP5 exhibited RPN values 61.6% and 35.3% lower value than the control, reaching values of 0.086 and 0.055, respectively. These results indicate that CP2-4 were able to interact cooperatively, developing stable triple-helixes in a similar way to Type I collagen. Such interactions were less favorable for CP1 and CP5, as seen by the lower RPN ratio. The RPN value follows the order of: Type I collagen control ≈ CP4 > CP2 > CP3 > CP5 > CP1. This is also consistent with the intended T_m values of the CPs, where CP2 and CP4 were designed to have higher T_m values (around 37 ºC) and thus maintain a more stable triple helical configuration. Interestingly, CP5 showed a multi-step denaturation process with temperature, which was related to the interrupted GXY sequence (**Figure 5-14**). This multi-step behavior hinders the assignment of a single T_m value to CP5.

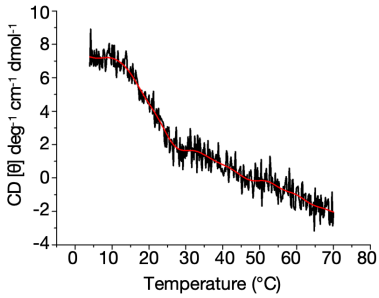


Figure 5-14. CD temperature scan at 222 nm for CP5 has a multistep denaturation process with several T_m values.

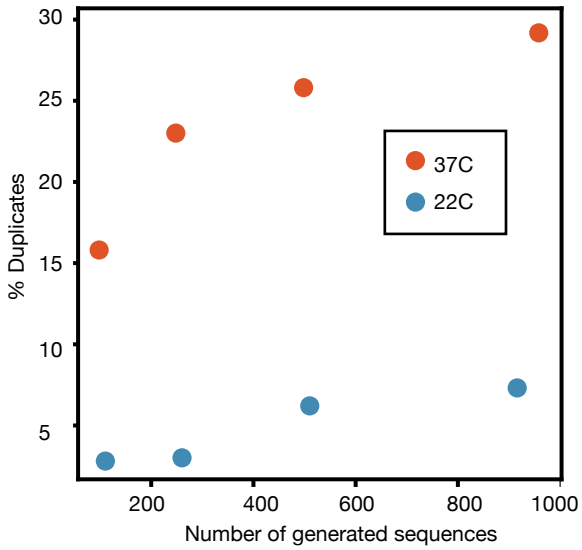


Figure 5-15. Percent of duplicated sequences given the number of generated sequences. No constraint is applied to ColGen-GA to make sure each sequence is unique. The 37 ºC sequences have more duplicates. This is likely because fewer triplets have the thermal stability necessary to reach the 37 ºC target compared to the 22 ºC target. We select the 1000 generated sequences and remove duplicates for subsequent analysis in the section.

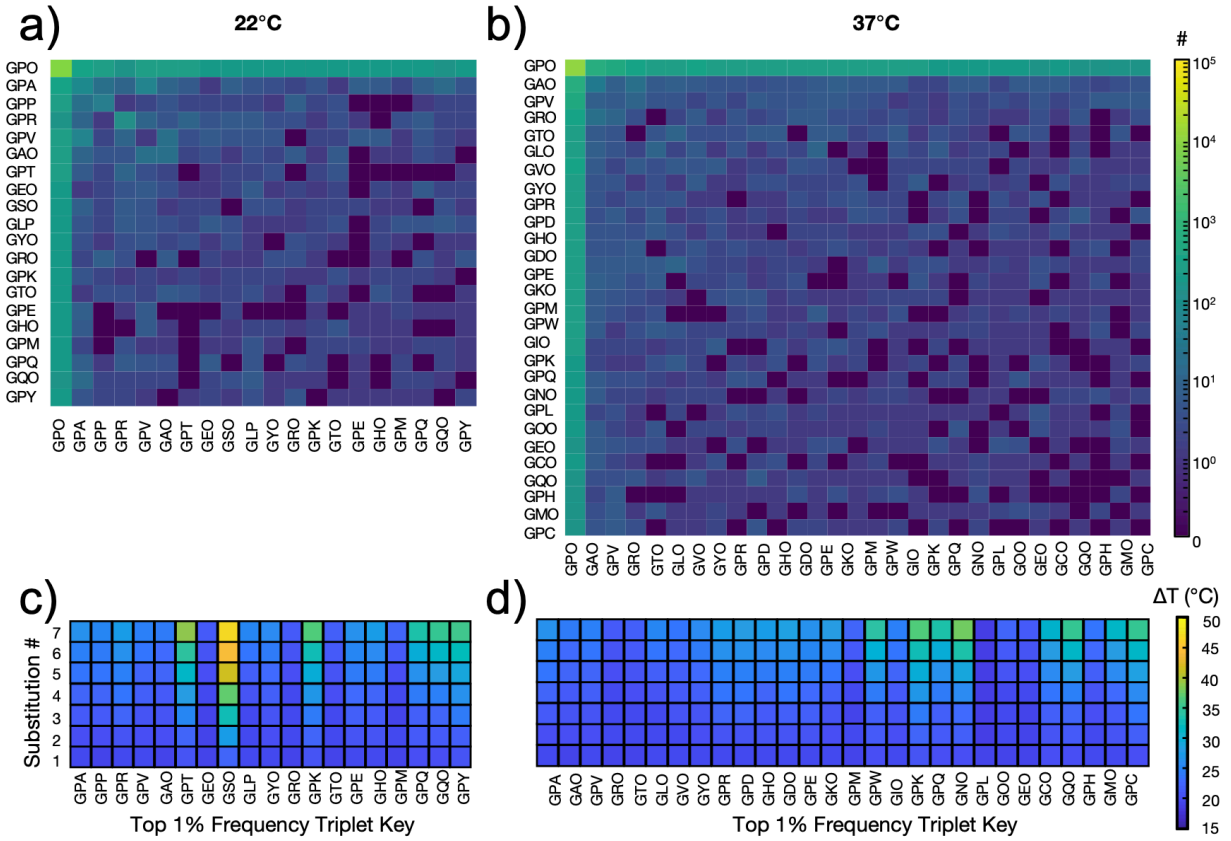


Figure 5-16. High-throughput identification of most frequent sequences in de novo collagen peptides. Co-occurrence matrix of the 1000 generated *de novo* collagen sequences for 22 °C (a) and 37 °C (b) when sorted by most frequent triplets shows which triplets occur together in the same sequence. These most frequent triplets from 22 °C (c) and 37 °C (d) are substituted n times into a $(GPO)_{14}$ ideal standard peptide and their destabilizing effect on T_m is evaluated, where $\Delta T_m = T_m(GPO)_{14} - T_m(\text{sequence})$.

Given the validation of the model with experiment and MD, we conduct high-throughput processing to derive insights into GXY triplets of collagen that are most suitable in achieving desired T_m values. After generating 1,000 *de novo* collagen sequences with T_m values of 22 °C and 37 °C (**Figure 5-15, Appendix S5-2, S5-3**), we found the top 1.3% most commonly occurring triplets within our generated sequences and determined their co-occurrence matrices in **Figure 5-16a,b**. The co-occurrence matrix helps show which GXY triplets occur with other GXY triplets to provide a graphical insight for how to build a larger sequence from a combination of triplets. GPO emerges as the most commonly present triplet in the generated sequences. This is in agreement with literature because GPO is the canonical triplet in increasing the strength of CPs and $(GPO)_x$ is often used as a gold standard in collagen mimetic peptides for thermal stability.[467] Beyond the presence of GPO, we also find a number of other triplets that emerge as useful motifs in achieving the desired T_m values. In alignment with others who have noted the stabilizing effect of KGE/KGD[414,467] lysine, glutamic acid, and aspartic acid contribute stability to the collagen peptide, as the residues GPK, GEO, and GDO have a minimal decrease in thermal stability compared to other frequently occurring triplets (**Figure 5-16b**). Interestingly, all of these triplets are in a GPY or GXO configuration where either P or O are present with the other guest amino acid replaced. Further, these guest amino acids do not follow a consistent physicochemical trend and there is a range of hydrophobic, polar, or charged residues that contribute to the mechanical stability of the CP. This lack of physicochemical consistency is not something we could have inferred from analytical

outcomes given that there is no trend in the data. We also note that the demonstrated guest amino acids do not follow the same occurrence as amino acids naturally occurring in collagen, where basic residues are in the Y position and Glu or hydrophobic residues are in the X position.[467] Instead, the guest amino acids demonstrated here follow a different occurrence given that this is a synthetic set. This provides further justification for using large data generation means through machine learning to derive design principles. Further, comparing the highest frequency triplets in the 22 °C (**Figure 5-16a**) to 37 °C (**Figure 5-16b**) data, we find that ~50% of the most frequent triplets in 22 °C are also found in the list of most frequent triplets at 37 °C. These are likely the triplets that contribute most to the thermal stability of the collagen sequence, while the other triplets help achieve the lower target temperature of 22 °C.

Given these most frequent triplets, we sought to understand which triplets had the greatest effect on the T_m values (**Figure 5-16c,d**). When substituted into the highest stability $(GPO)_{14}$ sequence, the top triplets in the 37 °C sequences induce a lower amount of destabilization of T_m value, where the destabilization is measured as $\Delta T_m = T_m(GPO)_{14} - T_m$ (*de novo* sequence), compared to the most frequent triplets in the 22 °C *de novo* sequences.

5.3.6 Relationship between triple helical quality and T_m values

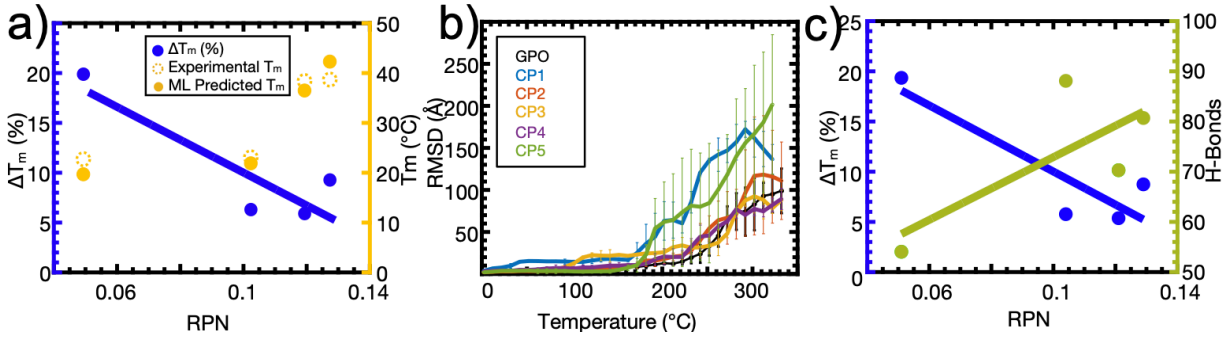


Figure 5-17. Relationship between collagen triple helix quality and Tm values using experiment and MD simulation. a) There is an inverse relationship between the RPN ratio and the difference in T_m value between the experimental CD T_m and ColGen predicted T_m. $\Delta T_m = \frac{(T_m,experiment\ CD - T_m,ColGen-GA)}{T_m,experiment\ CD}$ This indicates that ColGen algorithm is able to more robustly predict the thermal stability of higher quality triple helices. The RPN also follows a direct relationship with T_m value, indicating that more stable triple helices have a higher T_m. b) MD simulations show that the CPs maintain roughly the expected stability, measured by RMSD of the triple helix, as predicted by ColGen. c) Hydrogen bonding analysis at 50 °C in the MD simulation shows a similar correlation as the RPN ratio in figure part a. Peptides with more hydrogen bonding generally have a lower deviation from ColGen-predicted T_m values compared to experimental T_m values. Further, RPN has a direct relationship with the number of hydrogen bonds in the CP, indicating that a higher quality triple helix has more hydrogen bonding.

Upon experimentally measuring the RPN and the T_m, we found that the higher the RPN value, the lower the differences between ColGen-GA predicted and measured T_m values (**Figure 5-17a**). This is likely because a higher RPN ratio corresponds to a higher quality triple helix, which is more likely present for the high T_m value sequences as discussed.

To further validate the ColGen-GA model and provide support to the CD and DSC experiments, MD simulations were also used to simulate experimental heating of the triple helical peptides. While exact T_m cannot be extracted from MD due to the faster heating rate used in simulation

compared to experiment, MD simulations confirm that the stability ordering of the CPs is $(GPO)_{10}$ > CP4 ~ CP2 > CP3 > CP5 > CP1 (**Figure 5-17b**), as observed in CD and DSC measurements. In MD, $(GPO)_{10}$ rather than bovine collagen is used as a model collagen peptide mimetic with the highest T_m value. The MD results demonstrate that all of the peptides are correctly ordered in terms of their thermal stabilities, except for CP5, whose thermal stability in MD simulations is predicted to be much less than experimentally measured. This discrepancy is likely due to a poorer prediction of triple helical structure for CP5, as it does not follow the GXY pattern consistently.

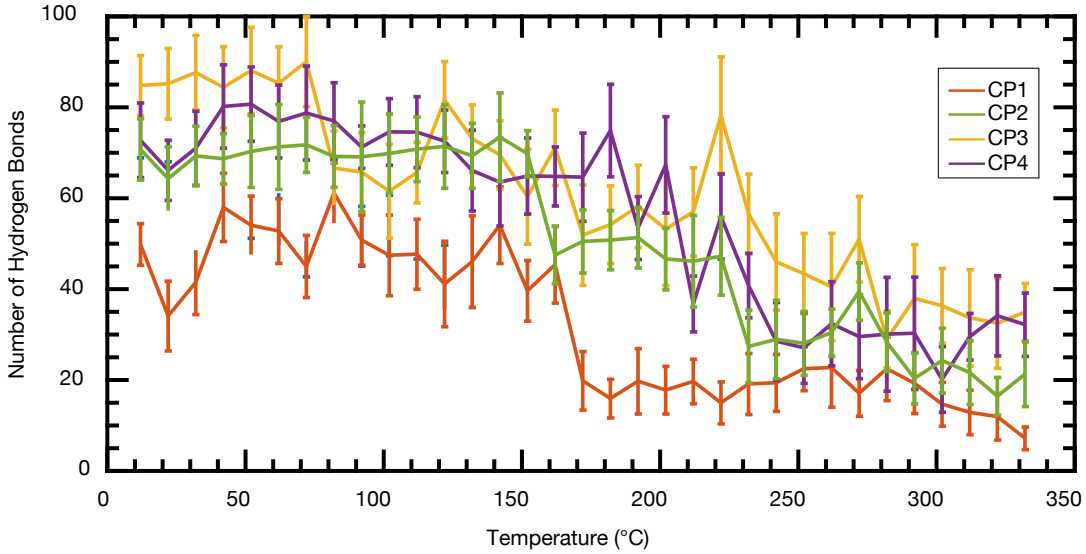


Figure 5-18. Number of hydrogen bonds in CPs as a function of temperature. MD simulations show how the number of hydrogen bonds in the CPs decreases as the CPs denature with increasing temperature.

Table 5-4. ML predicted and experimental Tm values. Calculated, literature values, experimental T_m and difference between theoretical and experimental values of T_m, expressed as a percentage ($\%\Delta T_m$), of all collagen samples tested in this work.

	T_m (°C)		
SAMPLE	ML Calculated	CD	$\%\Delta T_m$ (CD)
Col. Type I	≈ 40	40.6	1.48
CP1	22.16	18.6	-19.35
CP2	37.3	35.4	-5.37
CP3	22.0	20.8	-5.77
CP4	37.6	41.2	8.74

MD simulations also enabled us to further validate the relationship between RPN and T_m values and the accuracy of our predictions by developing a mechanistic understanding of the different CPs. We evaluated the triple helix quality of the different CPs by measuring the amount of hydrogen bonds between the strands as a proxy for triple helix strength and related it with their RPN value (**Figure 5-18**). We found that the CPs with higher RPN present more hydrogen bonds in their triple helix structure compared to CPs with lower RPN values (**Figure 5-17c, Table 5-4**). An image of representative hydrogen bonds is provided in **Figure 5-19** and we note that in our sequences, glycine qualitatively demonstrates the most hydrogen bonding. To our best knowledge, this is the first demonstration of the direct relationship between the number of hydrogen bonds computed in MD and RPN values experimentally measured with CD.

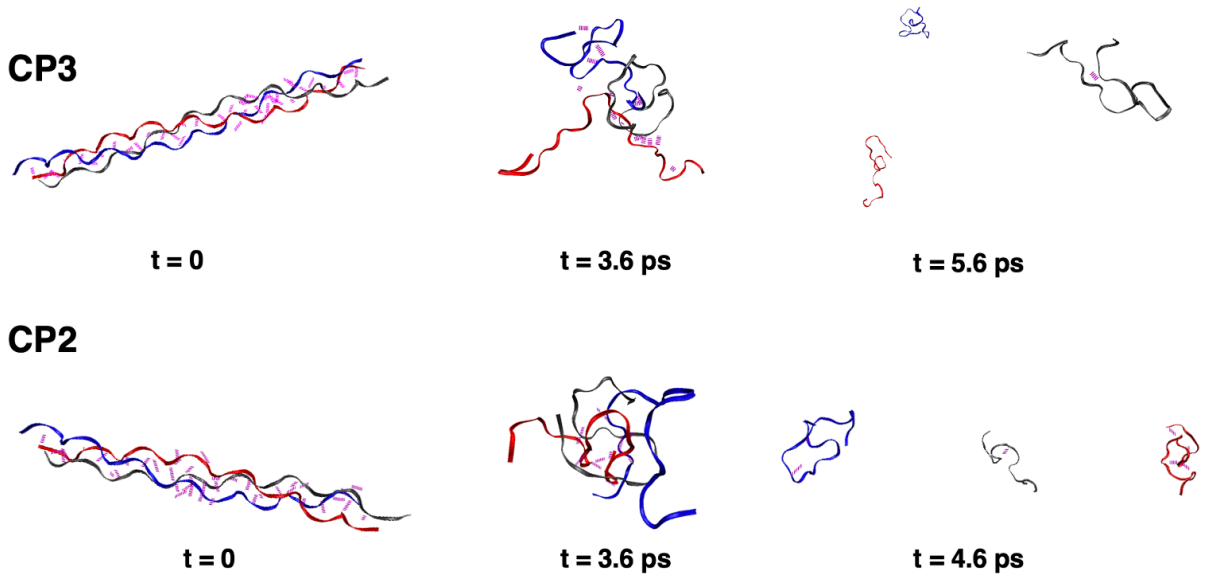


Figure 5-19. Evolution of the collagen structures of CP2 and CP3 over time during thermal treatment. In both cases, we observe how the H-bonds present in the middle of the triple helix are those that break later during the process. The ends of the triple helix unravel at first, becoming the initiators of the unfolding activity. As also reported in Figure 5-17, and clearly noticeable in this Figure, the H-bonds between chains progressively break and those that remain at last are only those present between the single chains because of their spatial configuration. Gly demonstrates the most hydrogen bonding.

5.3.7 Discussions, Implications, & Conclusions

We developed a platform that uses a deep learning model trained with input sequences from literature to generate *de novo* collagen sequences with desired T_m values. The model, ColGen-GA, incorporates ColGen, an LSTM-based T_m predictor of collagen sequences and a generative genetic algorithm to produce new sequences with specified thermal stability behavior. Our model shows good prediction power in extrapolating T_m values of the testing set for collagen sequences not included in our training dataset. The trained deep learning model is able to predict T_m values within an acceptable error range considering the amount of experimental data.

The machine learning algorithm also allows us to quickly determine how specific amino acids mutations, the amount of disorder in the sequence, and the sequence length affect the thermal stability of collagen. We determined that mutations in the middle of the sequence greatly affect stability and that the maximum achievable temperature is already reached at a sequence length of 14 repeat units. We then validated our model by selecting some of the generated sequences and testing them experimentally with CD and DSC and computationally with MD simulations. The CD experiments confirmed that these new collagen mimetic peptides had triple helical structure, and together with DSC, the experiments confirmed the predicted T_m values of the *de novo* sequences within a few degrees centigrade. By studying the quality of the triple helical formation of the CPs, as measured through its RPN value, we determined that higher RPN value CPs have less deviation between the experimental and ColGen-GA predicted T_m values. This means that CPs that have a higher quality triple helix, and thus a higher T_m, are

likely to have a better T_m prediction than CPs which are less able to form triple helices. These experimental results were also further validated by MD simulations, which showed that the *de novo* peptides followed the stability order as predicted. We also showed for the first time, that more hydrogen bonds correlate directly with higher RPN values and higher T_m values. Such a result is particularly relevant since a mechanistic understanding of the relationship between the RPN values and the number of H-bonds promotes a deeper understanding and rationalization of the thermal stability of collagen and, more broadly, protein sequences.

Given the validation of the model, we used the large dataset and computational power of the machine learning model to discover important triplets in the collagen sequences. ColGen-GA enables the fast generation and prediction of T_m values of 1000 new sequences in just 8 hours using a laptop, compared to the 10 day per collagen sequence simulation required for MD on 2 nodes and 32 CPUs. Because of this computational power and speed, we are able to derive new insights into important collagen sequences than what our previous modeling or experimental capacities enabled. We identified the highest frequency GXY triplets from the *de novo* sequences generated for target T_m values of 22 °C and 37 °C. The triplets identified can be used by other researchers when designing new collagen sequences with specific T_m values. To assess these triplets in the context of a longer sequence, we also provided a co-occurrence map to understand how these important triplets work together in longer sequences. These triplets would not have been discovered through analytical means, as we find that a consistent physicochemical principle, such as hydrophobicity or charge to explain the triplet behavior, is not present in the most frequent triplets. We envision that the triplets identified here could be used in creating useful collagen sequences.

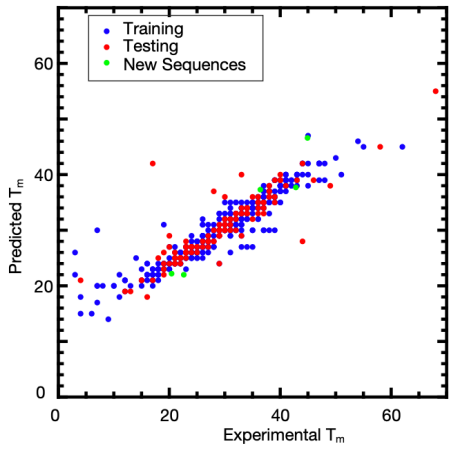


Figure 5-20. Benchmarking performance of ColGen model. *ColGen* model has good predictive capacity for training and testing data, as expected, and even for validation data for our new experimentally synthesized and measured collagen sequences.

While the model enables a quick prediction of T_m values and generation of new sequences (**Figure 5-20, 5-2b,c**), there are some limitations that should be expanded upon in future work. These limitations primarily arise from the data set used to train the model, which could be further expanded to include a wider range of sequences and sequence lengths. For example, the current dataset has a maximum T_m range of 70 °C. Further, the model is only trained on the standard amino acids and hydroxyproline. As such, it would be unable to predict the T_m values of other non-native amino acids and this problem too should be resolved by an expanded data set. Most of the collagen sequences used in our present dataset were collagen mimetic peptides rather than complete collagen sequences. Incorporating longer sequences would help

in the design of collagen proteins from bacteria, which often produce longer protein sequences.[468,469] Further, most of the sequences in the dataset incorporate hydroxyproline (O) or a specific subset of guest residues. As such, building new sequences that do not incorporate O, which is an especially important limitation to bacterially-produced collagens which have no means of producing O, would be challenging with the current data set used. Expanding the compositional diversity of the dataset would also help further improve the prediction. Another limitation of exploiting GA could be the efficiency. Compared with other optimization tools, such as the gradient method, GA has a slow computational speed when processing large amounts of initial populations or local minima. The convergence rate for the same size of initial population can be modulated by simulated annealing to change the probability of crossover and mutation on the fly.[470] Finally, manipulating the model itself would help with the prediction of heterotrimer sequences beyond the homotrimers presented here such that the model is more robust with respect to initial conditions. One way to do so may be to incorporate existing fundamental knowledge of collagen directly into the model.

Despite these limitations, the reported approach represents a powerful and efficient tool in the design of collagen sequences with specific T_m values. Our approach should lead to the design of new collagen biomaterials and tunable properties with *a priori* desired T_m values. Further, our presentation of triplets will help inform how to build mechanically robust collagen sequences at desired temperatures into the future, especially given the vast design space of 10^{21} combinations of $(GXY)_{10}$ sequences.

Beyond the generation of new sequences as an engineering tool, our approach contributes to an understanding of collagen denaturation rates and how these T_m values correlate to structure. Such information is important, for example, in understanding the mechanical behavior of specific tissues with impacts on denaturation or biological function in scenarios such as thermal treatments for cancer. Further, many collagen-based diseases such as *Osteogenesis imperfecta*, are based on mutations in the primary sequence of collagen. This method would help offer insight and perspectives on these disease states in context of thermal stability, with implications for future repair routes. Another aspect is collagen degradation by matrix metalloproteinases (MMPs), crucial in many physiological processes, such as wound healing, tissue remodeling and organ morphogenesis. It is well known that stable triple-helices are far more resistant to MMP degradation than denatured collagen, reflective of the structural stability of the matrices.[471,472] Thus, CPs capable of forming better triple helices are, thus, more resistant to MMP degradation. Similarly, a higher mechanical integrity and structural order of collagen results in a more robust collagen matrix.[471,473] Considering that human mesenchymal stem cells proliferate, propagate and differentiate in response to the mechanical properties of the matrix they develop in,[474] we envision that the ability of designing collagen sequences with tailored thermal stability with this deep learning method would allow us to create new biomaterials with on-demand MMP degradation rates, mechanical properties, and tailored influence on cell behavior. Finally, the role of collagen sequences in the context of mineralization *in vivo*, such as with hydroxyapatite and bone formation, can benefit from these new methods related to engineering approaches to modulate organic (collagen) and inorganic (e.g., hydroxyapatite) interfaces related to mechanics and bone structure-function.

5.3.8 Materials and Methods

Collagen samples

Several *de novo* sequences from the ColGen-GA model were selected for synthesis and experimental validation. These collagen-like peptides (CP) were synthetized by GeneScript Biotech (Piscataway, New Jersey, 95% purity and trifluoroacetic acid removal). As the triple-helix forming control, commercially available bovine Type I collagen was used (PureCol TypeI Collagen, AdvancedBiomatrix, Catalog #5005).

Circular Dichroism

Spectra were acquired using a Jasco J-815 Circular Dichroism Spectrometer (Easton, MD, USA). CPs and bovine collagen Type I were dissolved in PBS at 0.3 mg/mL (final pH 7.1-7.3). Samples were kept at 5 ºC for 72 hours before scanning in the far UV (180 to 260 nm) at 5 ºC. Ellipticity at 222 nm was monitored as function of temperature, while heating the samples from 5 to 70 ºC at 1 ºC/min with data collection every 0.1 ºC. For derivatization of the temperature scans and calculation of the minimum of the first derivative, the data was smoothed using a fast Fourier transform (FTT) filter with a cutoff frequency of 0.342 Hz. T_m values were calculated as the minimum of the first derivative of the temperature scans. After reaching 70ºC, the temperature was maintained, and samples were scanned from 180 to 260 nm. CD spectra included accumulating three scans at a scanning speed of 20 nm/s and 4 seconds of Digital Integration Time (DIT). For all plotted data the High-Tension Voltage, HT(V), of the photomultiplier was kept below 600V.

Differential Scanning Calorimetry (DSC)

Thermograms were acquired using a TA Instruments DSC (TA Instruments Q100 series, New Castle, DE, USA). CPs and collagen Type I were dissolved in PBS at 50 mg/mL (final pH 7.1-7.4) and kept at 5 ºC for 72 hours before measurement. A total of 20 µL of each sample was hermetically sealed in an aluminum pan (Hermetic Tzero pans model 901684.901, TA Instruments, New Castle, DE, USA) and scanned from 5 ºC to 65 ºC at a rate of 2.5 ºC/min using as a reference 20 µL of PBS. The melting temperature was considered as the minimum of the endotherm (See Supporting Information).[413]

Molecular Dynamics (MD)

MD simulations were performed using the NAMD code with CHARMM force field, which also includes parameters for the hydroxyproline residue. We prepared each peptide topology using the triple-helical collagen building script (TheBuScr) based on the primary amino acid composition, including the hydroxyproline residue. The protein was solvated with a 2.4 nm boundary water box using TIP3P water molecules as the solvent. The total number of atoms in the solvated system was approximately 90,000. A 1 fs timestep was used and rigid bonds were applied to constrain the bonds of the water molecules. Van der Waals interactions were computed using a cutoff for a neighbor list at 1.4 nm, with a switching function from 1.0 to 1.2 nm. For electrostatic interactions, the particle-mesh Ewald sums (PME) method was used with periodic boundary conditions. A preliminary energy minimization was performed using a steepest descent algorithm. The systems were then equilibrated at 275 K for 5 ns each under a constant atom, volume, and temperature (NVT), then constant pressure (NPT) ensemble. The resulting systems were further equilibrated under NVT for 2 ns before beginning the heating process to mimic the CD and DSC experiments. The temperature of the simulation was increased by 10 K every 10 ns from 275 to 600 K.[475] Root mean square deviation of the protein

backbone and hydrogen bonding number were determined from the last 6 ns of each temperature by using visual molecular dynamics (VMD) plugins on DCD trajectory files that were output every 50 ps. Each simulation was repeated 3 times.

5.4 End-to-end transformer model to predict thermal stability of collagen triple helices using an NLP approach

The key constituent of ColGen is a combination of convolutional layers with long short-term memory (LSTM), creating an artificial recurrent neural network. This LSTM model offers one manner in which ML methods can learn the underlying physics of collagen molecules, but has the downside of being difficult to train due to its requirement for sequential input. The recent development and application of transformer ML deep learning models[40] presents a new solution to this training problem,[43] offering faster and more efficient training of the data and the model. Briefly, the transformer model adopts self-attention to process data out of order and learns the context of each element via positional encoding. This non-sequential method of training could be relevant to collagen, where short-range (sequential) and long-range (non-sequential) interactions play a role in the structure.[44,45] The transformer framework has increasingly become the model of choice for NLP-type of problems in language and science applications and has most recently been used in AlphaFold 2 to predict protein structures.[46,47] While transformer models are powerful, since they can be generalized to a variety of applications and modalities (sequence regression problems, sequence to sequence translation such as secondary structure prediction, and other needs including field predictions[48,49]), they can also be difficult to train and often require large amounts of data. This has been exemplified in recent developments of very large language models based on these architectures, sometimes reaching hundreds of billions of parameters.[50–52] Further, to our best knowledge, while a few very recent examples exist of the application of these transformer models to predict the structure or binding properties of some other protein systems,[53–57] they have thus far not been used to directly predict biophysical properties of proteins.

In this section, we probe whether transformer ML models can be applied to predict the T_m values of collagen sequences, given the small dataset size, how this approach compares to our previous ColGen NLP model, and whether the predictions agree with available experimental data.[58–61] We develop two models, a larger ProtBERT-based model [53] and a smaller transformer model. We show that the transformer models can be trained on the collagen dataset, even though the dataset is very small.[53] The small transformer model has an R^2 similar to the larger ProtBERT model for the test data, but the ProtBERT model outperforms on the validation data. We believe this is the first demonstration of the transformer model to a smaller protein dataset for the prediction of physical properties, and shows that despite the relatively small size of this dataset, the transformer model is able to accurately predict T_m values of collagen sequences.

5.4.1 Transformer model training

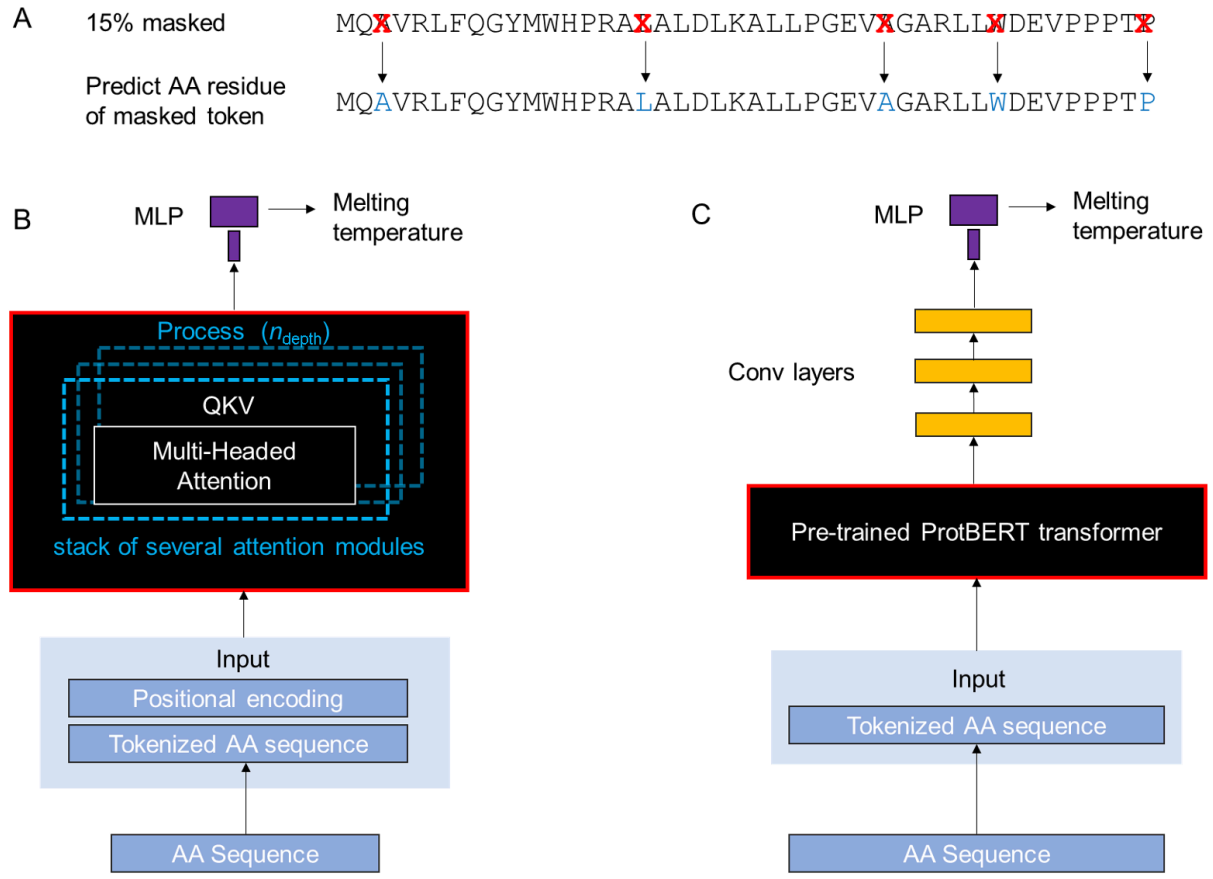


Figure 5-21. Overview of the small and ProtBERT transformer models trained and applied in this section. A) Masking is widely used to pretrain natural language processing methods. In this method, a random sample of amino acids has been replaced by a mask placeholder. The machine learning model then predicts the masked amino acid. This masked language modeling allows the machine learning model to develop a statistical understanding of the amino acid sequences, and is used in ProtBERT pretraining, as reported in [53]. A) Small transformer model (details in Materials and Methods and **Table 5-6**). A tokenizer is trained to translate amino acid sequences into a list of integer numbers, which are then processed by an embedding layer to generate input embeddings. B) Larger transformer model based on ProtBERT with a convolutional head for decoding, fine-tuned against our dataset. Both models use multilayer perceptron (MLP) blocks, consisting of fully connected dense layers, to yield the final prediction dimensionality. In all cases considered in this section, the whole sequence is mapped to a single, scalar value, T_m. For the ProtBERT model, we use the pretrained ProtBERT tokenizer. Positional encoding are added within the pretrained ProtBERT model.

We train two transformer models to determine performance and appropriate strategies for collagen predictions. Transformer models were recently developed for NLP to process sequential input data such as sentences all at once, unlike recurrent neural networks which can only process the input data in series. Transformer models use attention to determine the context of each word (amino acid) within the larger sentence (protein sequence). This allows transformer models to be more parallel, thus reducing training times, while also allowing relationships throughout the entire sequence to be more accurately discovered. **Figure 5-21** shows a basic overview of the two transformer models in this work. **Figure 5-21b** depicts a small transformer model with a linear head consisting of a multilayer perceptron (MLP), featuring fully connected feedforward artificial neural network layers that will be trained from

scratch, with the details in **5.4.4 Materials and Methods** and **Table 5-6**. The small transformer model is not pretrained. Our dataset, discussed below, is directly input into this small transformer model during training to yield a model that can provide T_m outputs based on sequence inputs.

Figure 5-21c illustrates the structure of the larger transformer model based on ProtBERT with a deep convolution head followed by an MLP, fine-tuned against our dataset.[62] The ProtBERT model, introduced in 2021, is based on BERT[52], which is a Bidirectional Encoder Representation for Transformers, introduced by Google in 2018 to understand natural language. The ProtBERT model consists of a very large number of transformer blocks, similar to the multiple self-attention blocks shown in **Figure 5-21b**. Our larger transformer model, utilizes a pretrained ProtBERT. Pretraining is an attractive strategy for NLP models as it does not require labeled data. In this step, conducted in earlier work, [53] the strategy of randomly masking part of the input is used; in this case, randomly mask amino acids) and then predict the masked amino acids (**Figure 5-21a**). This process is repeated over a very large number of sequences (e.g. Uniref100 with 217 million sequences as is the case for the ProtBERT model used in this section).[63] Through this training process ProtBERT learns the structure and representations of general proteins before being fine-tuned to our specific collagen protein dataset. This pretrained ProtBERT effectively acts as an embedding function in our larger transformer model that relates input sequences with a very large embedding dimension. The resulting tensor is of dimension 64x1024, where 64 amino acids is the maximum sequence length considered, and 1024 the embedding dimension produced by the ProtBERT model. The maximum length of 64 amino acids was selected as a maximum length as the collagen mimetic peptides in our dataset are typically around, or shorter than this length. Any sequence shorter than 64 amino acids is padded with zeroes, which does not affect the overall training.

The training data, discussed in the next paragraph, is used to fine-tune the larger ProtBERT transformer model together with the head. In this case, the weights of the entire model are adjusted based on the new input training data.

The above small and large transformer models are trained with a collagen dataset used for ColGen, and discussed below in the **5.4.4. Materials and Methods** section. Data normalization is important to improve model accuracy, and as such, the Scikit-learn StandardScaler model is used here. The StandardScaler model zeroes the mean and scales the data to unit variance. We note that the QuantileTransformer could have also been applied instead of the StandardScaler to rescale the data. The QuantileTransformer normalizes the data such that the data with the most frequent values is more spread out. The QuantileTransformer typically results in better transformer performance in regions with a large quantity of training data, but yields worse results in the low high T_m regions where there is less training data. Conversely, the StandardScaler does not have this same problem. As such, the section proceeds with the StandardScaler, but one could conceive of using the QuantileTransformer if they were looking to target higher performance in the median T_m range.

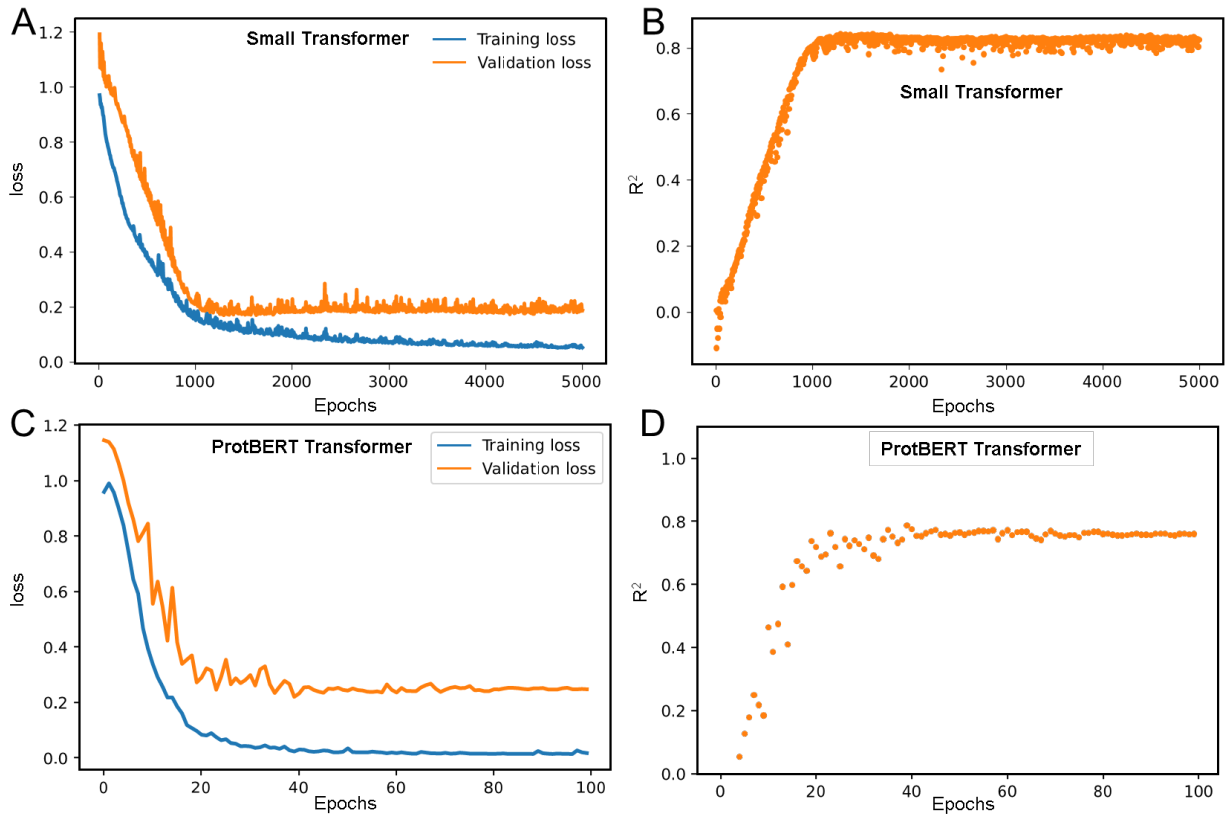


Figure 5-22: Training performance of the two transformer models. A,C) Training and validation loss over epochs shows the both models reach convergence over a reasonable number of epochs. B,D) R^2 value over epochs also shows convergence. A,B) Results for the small transformer model. C,D) Results for the larger ProtBERT based model fine-tuned against the collagen dataset.

To determine whether our transformer models could predict the T_m values of collagen sequences, we trained the model using our training set and show the training, testing, and validation performance in **Figure 5-22.** The small transformer model reaches convergence around 1300 epochs, whereas the ProtBERT-based model converges much faster around 40 epochs (**Figure 5-22a,c**). This convergence speed is unsurprising given that the ProtBERT-based model has already been pretrained on a larger set of protein sequences. However, the ~1500 epochs required to train the small transformer model is still within reasonable computational requirements especially since training proceeds quickly, and the epochs listed for the ProtBERT-based model do not take into account the computational time and cost required to train the larger model.

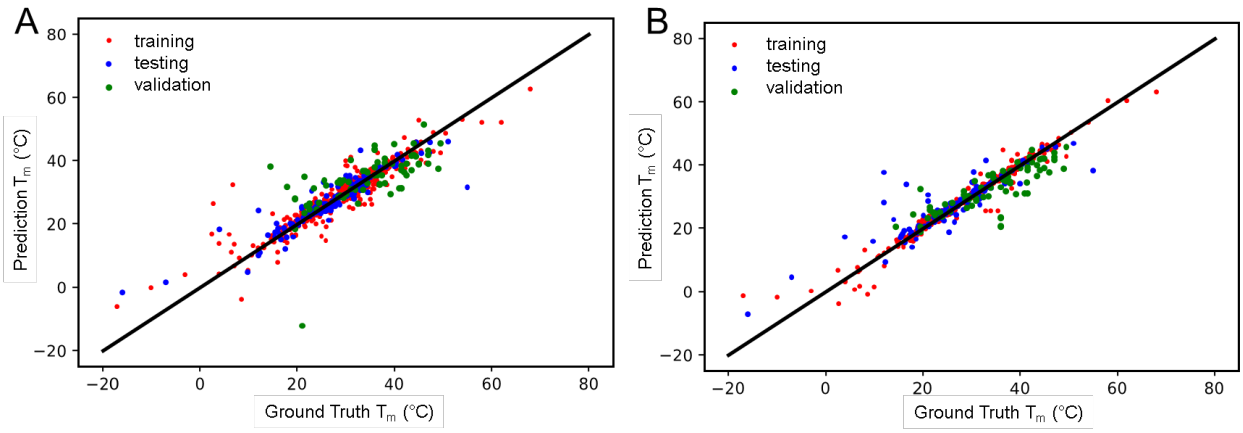


Figure 5-23. Prediction vs. ground truth, Tm. A,B) Unscaled training, testing and validation data, temperatures in °C. The small transformer model reaches an R^2 value of 0.84 and the large model 0.79 for the test dataset. For the validation dataset (green points), the R^2 values are 0.28 and 0.61, respectively.

Table 5-5: R^2 accuracy of models. Summary of R^2 accuracy measure, and number of parameters for the transformer model trained from scratch and the ProtBERT based model.

Feature	**Small Transformer Model**	**ProtBERT, Conv head, and finetuning**
R^2 accuracy (train/test/validation)	0.88/0.84/0.28	0.97/0.79/0.62
Number of parameters	108,069	423,079,813
Epochs until convergence	~1300	~40

The models reach a convergence R^2 value as the number of epochs increases (**Figure 5-22b,d**). Despite the speed of convergence of the ProtBERT-based model, the small transformer model is able to achieve a reasonable accuracy in prediction compared to ProtBERT-based model. **Figure 5-23** depicts predictions vs. ground truth of T_m of the unscaled training, testing and validation data in °C (**a,b**), and a summary of the performance of both models is listed in **Table 5-5**. The validation set, also discussed in the **5.4.4 Materials and Methods** section, [58–61] consists of recently published collagen sequences and their T_m values and were not included in the training or testing dataset. The small transformer model reaches an R^2 value of 0.84 and the large model 0.79 for the test dataset. This is surprising given that the small transformer model uses only 0.026% of the parameters required in the ProtBERT-based model. Despite this good R^2 value for the testing data, for the validation dataset, the R^2 values are 0.27 and 0.42 for the small transformer model and ProtBERT-based model, respectively. Given that the validation dataset has a slightly different mean and standard deviation than the training dataset, we also calculated R^2 values using the training dataset as baseline. This calculation can be found in the **5.4.4 Materials and Methods** section and results in an R^2 value of 0.46 for the small transformer model and 0.71 for the ProtBERT-based model. We believe this lower R^2 value is due to two factors. First, predicting T_m values of very low temperature sequences is challenging due to the lack of a significant quantity of training data at these values. Second, the distribution of amino acids in the validation dataset is slightly different than that of the training dataset; the validation data has a higher frequency of the amino acid phenylalanine (F), which makes an important difference in the R^2 values given that that ~20% of the validation sequences have F in them.

The significant difference between the R^2 values for the smaller transformer model versus the ProtBERT is likely due to the general relationships between amino acid sequence and protein structure that the ProtBERT learned in pretraining and that the small transformer model did not. As such, though the R^2 values for the testing data are very good for the small transformer, it seems that the larger pretrained ProtBERT is more helpful in extrapolating to other collagen sequences.

5.4.2 Transformer model T_m prediction

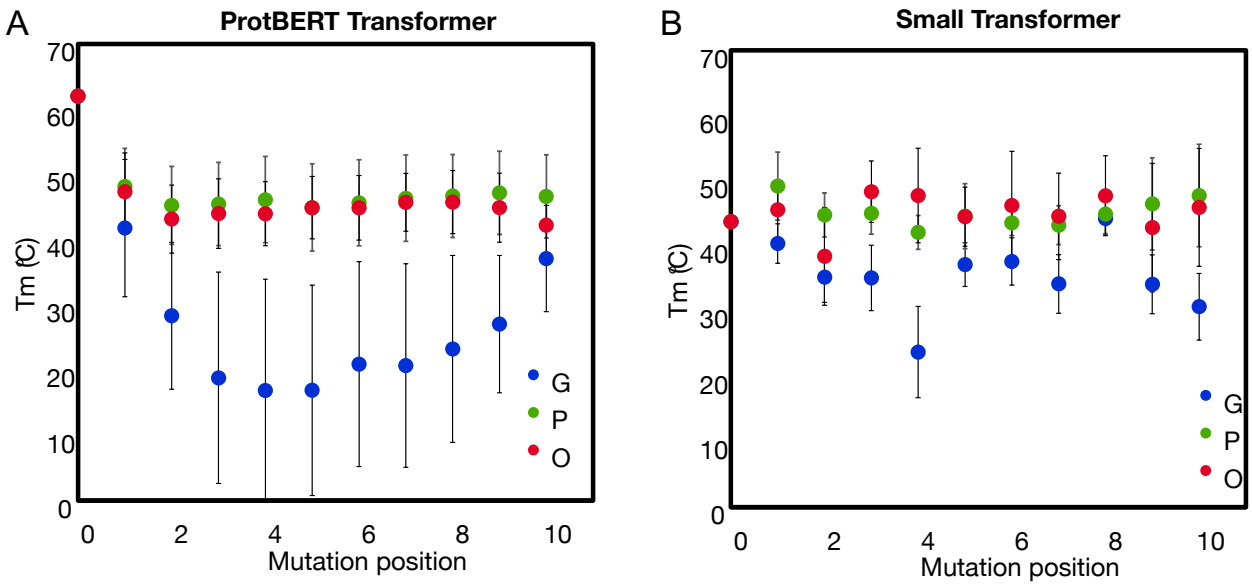


Figure 5-24. Transformer predictions for Tm based on mutation position and disorder parameter. Characterization of mutations in positions G, P, or O along the $(GPO)_{10}$ model peptide triplets for the A) ProtBERT and B) small transformer model. Mutation substitutions are made with every amino acid, and values demonstrate the mean and standard deviation of these substitutions.

Using the transformer models, we demonstrate the effect of various mutations on T_m values. We start with a model $(GPO)_{10}$ sequence and determine how mutations in either the G, P, or O position along the 10 triplets of the peptide affect the resulting T_m values (**Figure 5-24a,b**). Consistent with the earlier ColGen model, we find that there is a greater T_m destabilization if there is a mutation towards the middle of the sequence. This indicates that there is a critical transition point along the length of the sequence that is critical in maintaining the structural stability of the protein. Further, we also find that disruptions in glycines are most destabilizing. This is consistent with experimental findings that glycine disruptions impact stability and often constitute disease states.[64–67] Interestingly, the ProtBERT model predicts a more similar trend to our original LSTM-based ColGen model compared to the small transformer model. This is likely because the ProtBERT model pretraining is helpful in extrapolating to other collagen sequences as discussed above.

We note that the goal of this section was to demonstrate how ML and specifically the transformer model can be used to predict the stability of collagen sequences given the small dataset. While the transformer models do not have as good of an R^2 as the recent physics-based model by Walker et al. (0.84, 0.79 for the small transformer and ProtBert transformer, 0.95 for Walker), it is notable that ML model is able to have strong prediction capability without having explicit physico-chemical principles guiding the prediction. One benefit to the ML model

is that the model evolves as more data points are added, and does not need to be rebuilt to find physical laws that guide the new data points. Similarly, when comparing to the excellent model by Persikov et al., the ML model here allows the prediction of thermal stability, even for sequences which do not follow the traditional Gly-Xaa-Yaa. In addition, the performance of both transformer models reported in this section is better than the original LSTM based ColGen approach, where the R^2 was 0.67 for the test dataset. This improvement is likely because the LSTM approach only includes short-range interactions, and has practical limitations with how deep the neural network can be built to account for all relevant interactions. In contrast, the transformer models have no limitations on the length of their interactions, can provide a more accurate description of the sequences, and thus has a larger capacity for training.

5.4.3 Discussions, Implications, & Conclusions

Our analysis showed that we could successfully develop transformer models to predict key physical parameters of collagen molecules, here based on T_m. We explored two strategies. One was the design of a transformer model from scratch, with relatively few parameters (**Figure 5-22A**). The model was able to adequately capture the relationship between sequence and T_m, but did not perform as well for very low temperatures (likely due to the lack of significant sequences in that range) or for amino acids not well represented in the training set. The second model, developed based on a fine-tuning strategy based on the ProtBERT model (**Figure 5-22B**)[62] achieved adequate accuracy and performed better for the validation dataset. Key performance metrics are summarized in **Figures 5-23, 5-24** and **Table 5-5**.

The results show that transformer models can be successfully developed, either from scratch, or by using pretraining and fine-tuning. This opens up many exciting possibilities to capture other structure-function relationships in biomaterials. While the ProtBERT based model performs better for validation, it features a much larger number of parameters and requires extensive and expensive pretraining to be developed. It may be most useful when predicting T_m values of sequences that do not follow the amino acid distribution of the training data. Notably, the small transformer model requires only 0.026% of the number of parameters compared to the much larger model, but reaches almost the same accuracy for the test data. Further, even though fewer training epochs are needed to fine-tune the large model, a significantly higher computational burden was expended on training the original pretrained model. On the other hand, the ProtBERT model can be adapted quite easily to meet a variety of downstream tasks, as was shown in this section for mechanical properties.

Given that the goal of our work is demonstrate the transformer methodology to the prediction of physical properties of small protein datasets, we have included two Jupyter Notebooks on the small transformer model and larger ProtBERT model. Our hope is that the Jupyter Notebooks (**Appendix S5-4, S5-5, GitHub**) can serve as a good starting point for others to adapt and develop these transformer models further, especially for small protein datasets or the prediction of physical properties such as the T_m value demonstrated here.

Given the power of these transformer models to capture relevant properties of small protein datasets, one could envision their use in predicting the mechanical behavior of collagen-based diseases such as osteogenesis imperfecta, where 90% of diseases are caused by mutations in the type I collagen genes. Beyond collagen, the transformer models here can also be applied to other protein science and engineering challenges where large datasets are often not easily

accessible, such as the design of silk-elastin composites for use as biomaterials with specific mechanical properties or predicting the lethality of coronavirus variants given the amino acid composition of its spike protein.

5.4.4 Materials and Methods

Data preparation and dataset

The dataset is identical to the one reported in an earlier publication from this lab **Appendix S5-1**.[41] The dataset consists primarily of collagen mimetic peptides. The T_m value of collagen is affected by the exact specifications of the experiment such as solvent, heating rate, instrument, and concentration,[68–70] which could cause discrepancies in T_m between the same amino acid sequence under different conditions. However, we chose the majority of our sequences that collected T_m values under aqueous solvent conditions at pH 7 (water, PBS, ammonium acetate), heating rates from 0.1-1 °C/min, using CD spectroscopy. A more refined dataset would include T_m measurements from collagen sequences under the same test conditions, but this is beyond the scope of this current work. We use Scikit-learn's Standard Scaler model[71] to scale the input data by zeroing the mean and scaling to unit variance. The whole dataset includes 633 pairings of sequences and T_m. We use a 80% training and 20% testing data split. The validation dataset consists of the *de novo* homotrimeric peptides from Walker et al., Mekkat et al, and Kohler et al,[58–61] and the experimental and predicted T_m values are included in **Appendix S5-6.**

Table 5-6: Summary of model parameters for the small CollagenTransformer model trained from scratch, built on a transformer encoder model. We also summarize parameters for the Adam optimizer.[73] The model has a total of 108,069 parameters.

Parameter	Value
Batch size	128
Maximum length	64 amino acids
Number of tokens	23
Transformer depth	3
Number of attention heads	10
Embedding dimension (AA sequence)	8
Embedding dimension, positional encoding (learnable)	8
Learning rate	0.00005
Beta1, Beta2	0.99, 0.999

We develop a small transformer model[43] from scratch, as schematically shown in **Figure 5-22A.** Parameters and hyperparameters are summarized in **Table 5-6**. Trainable position encoding is used. To keep the number of parameters small, the transformer only features 3 layers, to encode the input sequence data into a higher-dimensional space. The input data is processed using a tokenizer trained based on the amino acid sequence data in the collagen dataset. It is processed by an embedding layer. We use learnable positional encodings, fed to the model via an embedding layer. The output from the transformer model is processed using an MLP block to yield the desired output dimensionality; which is 1 scalar value for each sequence. The model is trained using an Adam optimizer. All parameters for these aspects are listed in **Table 5-6**. All code is written in Python, using PyTorch.[72]

Table 5-7: Key parameters and head design for the pretrained ProtBERT based fine-tuning model. We use three convolutional layers, followed by MLP, to process the embeddings predicted by the ProtBERT model.[62] The first layer takes the 64×1024-dimensional output of the ProtBERT model (sequence length × embedding depth) and processes it, ultimately through all convolutional layers and the MLP block to yield a scalar output for each sequence input. We use an Adam optimizer[73] for fine-tuning, with a relatively small learning rate.

Parameter	**Value**
Batch size	12
Maximum length	64 amino acids
Learning rate	0.000005
Beta1, Beta2	0.99, 0.999
LR decay rate	0.96
Warmup steps	500

Conv Layer	**Kernel Size**	**In channels**	**Out channels**
3 (head)	3	512	512
2	3	512	512
1	3	1024	512

As comparison and alternative strategy, we built a larger transformer model based on a pretrained model using the BERT architecture, as schematically shown in **Figure 5-22B.** Parameters and hyperparameters are summarized in **Table 5-7**. We use a BERT tokenizer and the pretrained model's processing of embeddings and positional encoding, as reported in [62]. All code is written in Python, using PyTorch.[72] The pretraining was accomplished by randomly masking 15% of the amino acids in the input, using all sequences in the UniRef100 dataset [63]. The pretraining-fine tuning strategy offers the possibility to learn fundamental insights about a system of interest, here proteins, from a much larger but unlabeled dataset (UniRef100) and then fine-tune the model to be able to solve a particular task, such as predicting T_m for collagen sequences, as downstream task.

<u>R^2 Calculation for Transformer Models with Training Dataset Baseline</u>

The mean and standard deviation for the training/testing (μ=29.08°C, σ=9.46°C) and validation datasets (μ=34.24°C, σ=8.74°C) are slightly different. Because the model is trained on scaled data from the training/testing set, we calculated R^2 with respect to this dataset baseline to yield the following:

$$R^2 = 1 - \frac{\sum\left(y_{validation\_truth} - y_{valiation\_prediction}\right)^2}{\sum\left(y_{valid\_truth} - \widehat{y_{training}}\right)^2} \quad (5\text{-}1)$$

CHAPTER 6

CONCLUSIONS AND OUTLOOK

6.1 Major conclusions

The goal of this thesis was to help in the design of sustainable materials that are fracture-resistant, self-healing, and have exceptional mechanical properties, by looking to biological systems for inspiration. In doing so, we also aimed to help elucidate some of the complex mechanistic principles underlying biomaterials, which are still not well understood. We focused on metal-coordination bonds, which are a promising chemistry for designing materials with specific dynamic mechanical properties that ultimately enable fracture-resistance or self-healing. Their increased kinetic lability compared to standard covalent bonds enables dynamic, reversible mechanical properties. More remarkably, metal-coordination bonds can be tuned over a wide range of bond strengths and dissociation times using simple levers such as metal ions, pH, ligands and counterions. These bonds have opened up new opportunities beyond polymer physical structure to engineer time-dependent mechanical properties. In this thesis, we advance the ability to engineer metal-coordination bonds into proteins and polymers for desired mechanical behavior.

In Chapter 1, we discussed the emergence of metal-coordination bonds as a exceptional chemical motif to control the mechanical properties of polymers and proteins.

In Chapter 2, we predicted the macroscopic dynamic mechanical behavior of metal-coordinated polymers using simulated free energy landscapes of the metal-coordinate complex crosslinks. We incorporated metal-coordinate bonds as crosslinkers in ideal polymer network hydrogels, such that the experimental relaxation time of the hydrogel was controlled by the relaxation of the coordination complex. Using metadynamics implemented in MD simulations, we found that the energy landscape of the coordination complexes could be used to predict the experimental macroscopic relaxation time of the ideal metal-coordinated network using an empirical equation. We then expanded the system by constructing a coarse-grained model to show how the dynamic mechanical properties of the network could be controlled by the crosslinker chemistry.

In Chapter 3, we examined strength of clusters of metal-coordination bonds, to more closely mimic coordination bonds found in natural systems. We sought to probe whether clusters of coordination bonds act cooperatively to yield the remarkable stiffness and mechanical properties seen in natural materials. We conducted both experiment and simulation on various de novo metal-coordinated proteins to probe this question, and found that metal-coordination bonds can rupture cooperatively, but have highly heterogeneous deformation behavior. Our results suggested that metal-coordination bonds may have been evolved by biology for primarily energy dissipative means.

In Chapter 4, we contextualized our insights from model systems by analyzing native biological metal-coordinated materials. We characterized the role of metal ions on the structural folding

and mechanics of Nvjp-1, the major protein of the *Nereis virens* worm jaw which uses metal-coordination to produce a sclerotized, hard jaw. As one of the only highly metal-coordinated proteins with a computationally resolved structure, Nvjp-1 offered the opportunity develop a detailed nanostructural understanding of the role of metal ions in the native biological environment. Using various simulation techniques, we found that the quantity of metal ions plays a large role in the structural folding and compactness of the proteins, but that the distribution of metal-ions and hydrogen bonds has a large influence on mechanics.

In Chapter 5, we sought to develop machine learning methods to predict the mechanical properties of bio and bioinspired materials. We expanded our focus beyond metal-coordination systems to the biomaterial collagen, due to the larger amount of available data required for machine learning. We showed that by using deep learning, natural language processing, generative algorithms, and transformer models, we could effectively predict the thermal denaturation mechanical properties of collagen molecules. We experimentally validated our natural language processing model and used it to design new collagen sequences with desired thermal properties. These important first demonstrations serve as a basis for developing machine learning models for other bioinspired materials. Once more data is collected on metal-coordination bonds, these first demonstrations would help advance the study of metal-coordinated proteins more broadly using machine learning methods.

Altogether, this thesis made important contributions in the design of biomaterials, and specifically metal-coordination bonds and collagen, towards advanced mechanical function using both simulation and experiment.

6.2 Immediate future directions of this thesis

Several important future directions emerge from the foundations laid out in this thesis. These future directions are primarily related to the experimental or computational validation required to strengthen the design principles presented in this work for the rational mechanical use of metal-coordination bonds. For example, the hydrogel relaxation time predictions in this thesis were for the nitrogen-family ligands with Ni^{2+}. A more universal equation would characterize additional metal-coordination complexes and polymers to further validate or improve the robustness of the empirical equation presented in Chapter 2. Such an expansion should also closely model buffer effects in metal-coordinated materials, as even non metal-coordinating buffers have been found to produce different relaxation times in metal-coordinated hydrogels. A universal relationship between energy landscape and mechanical properties would have resoundingly large consequences on the field of soft matter. This ability to directly relate chemical energy landscapes to resulting mechanical properties is still missing despite similar relationships in the field of hard condensed matter.[302,476]

The general concept to expand the applicability of our design principles applies to the concepts of rupture strength presented in Chapter 3 as well. Testing additional *de novo* peptides would help indicate how to design proteins for specific rupture and fracture behavior. Currently, the largest bottleneck is knowing the protein structure, as the structure is extremely critical in characterizing the mechanical properties of metal-coordinated materials. For example, advanced structural resolution would allow the clear differentiation between the contribution of metal-coordination bonds and hydrogen bonds to the resulting protein

mechanics such that materials engineers could select the most appropriate bond type for their desired application. While techniques like AlphaFold used in this work offer a good first approximation of the structure, they cannot resolve the folding structure of multimers, especially in the context of metal ions. Advanced folding techniques that predict structure with metal ions present would help make the predictions in this work more robust. Alternatively, experimental testing methods with rapid peptide synthesis and high-throughput SMFS could also achieve a similar output.

Within the context of native biological materials, only a handful of resolved structures of proteins with high amounts of metal-coordination exist, limiting the understanding of coordinated biological proteins. Resolving more protein structures through experimental techniques such as crystallography or computational techniques such as replica exchange MD in the presence of metal ions would enable a clearer nanostructural understanding of the mechanical role of metal ions. We can also further increase existing data on metal-coordinated proteins for mechanical function by conducting a thorough mechanical characterization of proteins in databases such as MetalPDB.[287,477]

More broadly, increasing the accuracy and efficiency of computational methods would unlock the ability to universally predict the properties of metal-coordinated materials. Currently, the largest limitation in modeling metal ions is the force field, and the force field is absolutely critical to MD. Metal coordination bonds exhibit a number of chemistry challenges, such as polarization or charge transfer between the metal and ligand, which are difficult to capture in conventional force fields. These challenges are one of the core reasons why this thesis only focused on one metal ion chemistry (Ni^{2+}). Developing metal ion parameters is nontrivial,[113] and advances in metal parameters for mechanical function as well as easy implementation methods for widespread adoption, will be required to progress this field. In contrast, machine learning can be applied to systems even when the force field is unknown, due to its reliance on data rather than underlying physical principles. However, for problems such as the metal-coordination mechanics in this work, machine learning methods that can train on smaller datasets will be required.

6.3 Broader outlook of metal-coordination bonds for mechanical function

Continued progress in the understanding and application of metal-coordinate chemistry requires biological, materials and engineering research with an emphasis on mechanical properties (Error! Reference source not found.**6-1**). Further exploration at all length and time scales is required to realize the notable potential of metal-coordination bonds to enable dynamic structural properties, just as in natural systems such as spider webs, in which chemistry and web architecture both contribute to deriving variegated functionalities for biological survival.[22,478–480] To this end, characterization of more metal-coordinating biological materials is required to reveal where metal-coordination complexes exist and how they function. This understanding can then be translated into a synthetic context to design new ligand chemistries and enable the formation of robust networks under multiple environmental conditions. Metal ions should also be studied in the context of gradients, well-studied in biology, but not well implemented in synthetic systems, to build temporal[25] or compositional[56] gradients. Further, positioning of metal-coordinate bonds within proteins or block copolymers

may enable even greater control over mechanical properties. This design work must be accompanied by simulations and characterizations that directly link chemical bond properties with macroscopic mechanical properties or help explain natural biological materials designs.[214,481] We may even start to answer questions such as where the strength of a bond comes from or how it relates to static or nonlinear mechanical properties such as toughness, a significantly underexplored area in metal-coordination bonds. With these understandings, it will be possible to design metal-coordinated materials that operate over a wider range of tunable mechanical properties.

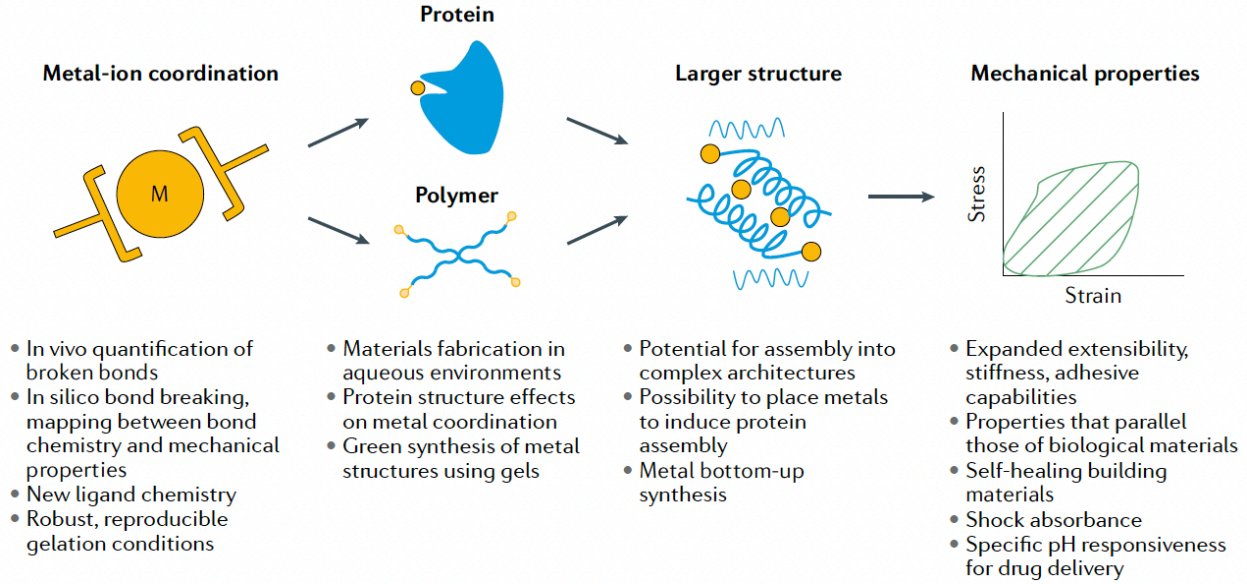


Figure 6-1. Future directions in the study of metal-coordination bonds. Key open questions and future milestones in the study of metal-coordination bonds at different length scales, from that of the individual bonds to that of the mechanical properties of the macroscopic structures they enable.

The dynamic, tunable nature of metal-coordination bonds allows several futuristic applications. For biological tissue-engineering applications, strictly controlling the relaxation times of gels would enable specific cell growth and differentiation or allow a precise trigger for drug release *in vivo*. Self-healing bonds may also be incorporated in structural building materials to repair cracks or prevent earthquake damage. If metal aggregation in the network can be controlled, these materials can potentially grow small metallic structures or minerals for biomineralization[482–484] in a controlled manner. Finally, it is hypothesized that biological organisms prefer metal-coordination motifs to other bonds, such as hydrogen bonds, when secreting mechanically robust materials outside of the body in aqueous environments. This preference may provide clues for developing materials that require mechanical strength and limited susceptibility to water, such as environmentally friendly biomass-based materials that are not humidity sensitive or materials for fetal surgeries in liquid environments. Overall, the wide range of metal-coordination complex interactions and their tunable parameters will elicit a wide variety of new insights and applications.

APPENDIX

GLOSSARY

AFM	atomic force microscopy
AlphaFold	protein folding tool
CD	circular dichroism spectroscopy
CGMD	coarse grained molecular dynamics
CHARMM22	chemistry at Harvard Macromolecular Mechanics Force Field
ColGen	developed collagen NLP model
CV	collective variable
DOPA	L-3,4-dihydroxyphenylalanine
DSC	differential scanning calorimetry
Ea	activation energy
ELP	elastin-like polypeptide
G'	storage modulus
G"	loss modulus
GA	generative algorithm
ITC	isothermal titration calorimetry
lp	persistence length
LSTM	long short-term memory
MD	molecular dynamics
NAMD	nanoscale molecular dynamics
Ncr	critical number of bonds that rupture together
NLP	natural language processing
NPT	isothermal- isobaric ensemble
NVT	canonical ensemble
PDB	protein data bank file
PEG	polyethylene glycol
ProtBERT	protein bidirectional encoder representations from transformers model
REMD	replica exchange molecular dynamics
SMD	steered molecular dynamics
SMFS	single molecule force spectroscopy
TIP3P	TIP3P water model
Tm	melting point
UV-Vis	UV-Visible spectroscopy
τ	relaxation time / bond dissociation time

CHAPTER 2—SUPPORTING INFORMATION

Entanglement Fraction

The following section is summarized from the calculations in the supplementary information of S. Tang et al.[227] Key calculations of from their work relevant to the calculation of entanglement fraction are presented here:

The entanglement volume fraction is estimated by classic reptation theory to be:

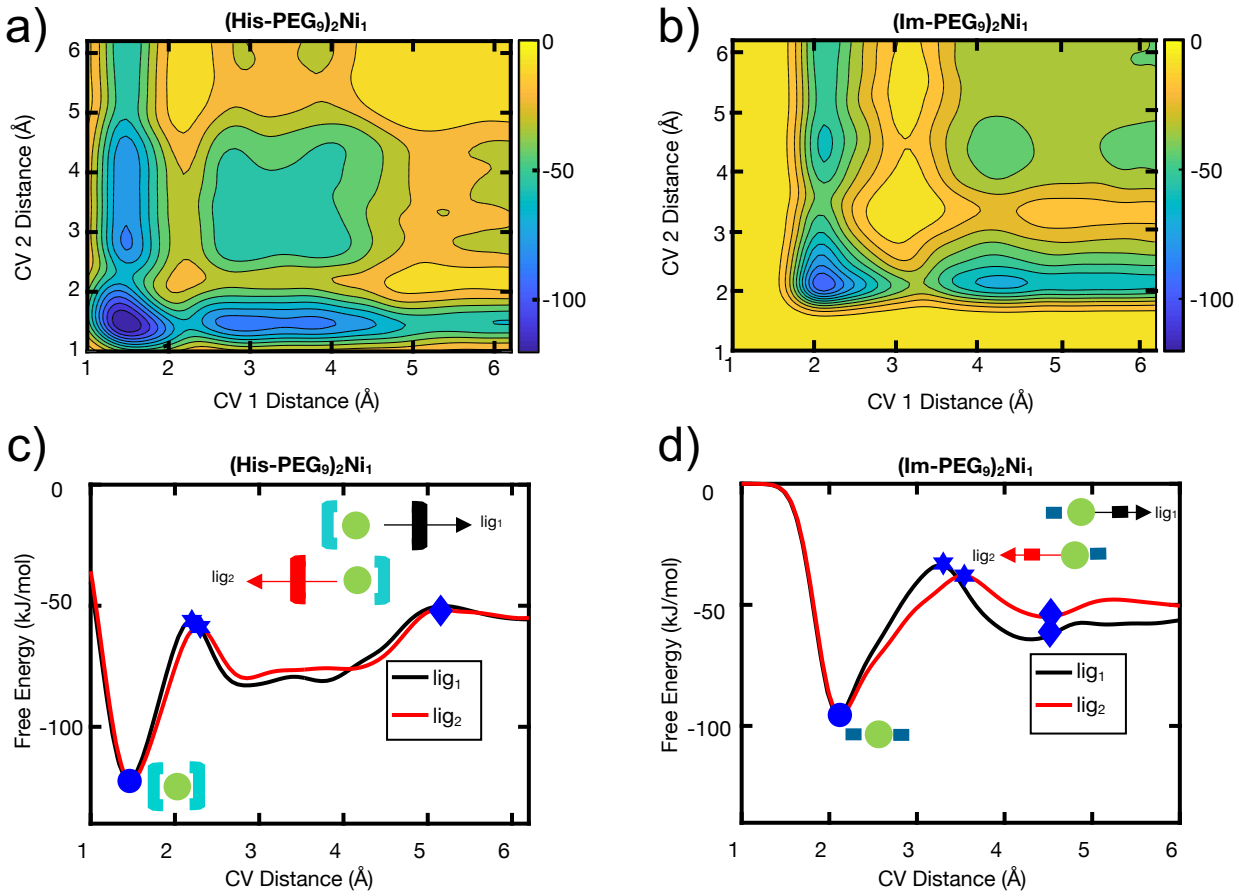
$$\varphi_e = \left(\frac{N_e}{N}\right)^{3v-1} = \left(\frac{M_e}{M}\right)^{3v-1} \quad \text{(S7)}$$

where N_e is the number of monomers between entanglements in a melt, N is the degree of polymerization of the polymer, and v is the Flory exponent. Correspondingly, M_e is the entanglement molecular weight of PEG, which is 1730 g/mol, and M is the molecular weight of the segment between two crosslinks, or 5000 g/mol. As such, the entanglement volume fraction is 0.44.

In our experiment, we use 20% w/v for the histidine and imidazole gels. Volume fraction is calculated by

$$\varphi = \frac{m_p/\rho_p}{m_p/\rho_p + V_s} \quad \text{(S8)}$$

where p is polymer and s is solvent, m is mass, ρ is density, and V is volume. The density of 4-arm PEG is 1.128 g/cm3. For a 20% w/v gel, the corresponding volume fraction is 0.151, which is less than the entanglement volume fraction.



	His_2Ni_1 Complex	**His_2Ni_1 -PEG**	**Im_2Ni_1 Complex**	**Im_2Ni_1 -PEG**
Eb (kJ/mol)	50.14	66.55	33.76	41.96
Ea (kJ/mol)	53.96	64.34	47.54	59.66

Figure S2-1. Energy landscapes of Ni2+-ligand-PEG are similar to Ni2+-ligand landscapes with minor quantitative differences. Energy landscapes of a) $(His\text{-}PEG9)_2Ni_1$ and b) $(Im\text{-}PEG9)_2Ni_1$ are qualitatively similar to landscapes with only the metal-coordinate complex (**Figure 3**): the position and shapes of the deepest binding well, metastable states, and activation barriers are preserved across the PEG-ligand and ligand-only landscapes. PEG9 is used to capture the interaction of PEG with the metal-coordinate complex and water solution while being computationally tractable compared to PEG53 (degree of polymerization in the experiment). 2D energy landscapes (c,d) are derived from the 3D energy surface where lig_x is held at a constant distance while the other ligand dissociates (see diagram on c,d). E_a is computed as the energy difference between the minimum energy (blue circle) and highest transition energy (blue hexagram) before the bond is broken. The van der Waals interaction cutoff between the coordinating nitrogen atoms and the metal ion is represented by the blue diamond. The table shows how the simulations with PEG and without PEG scale by: $E_{PEG\text{-}ligand} \sim 1.2 * E_{ligand}$. Figure from Ref.[216]

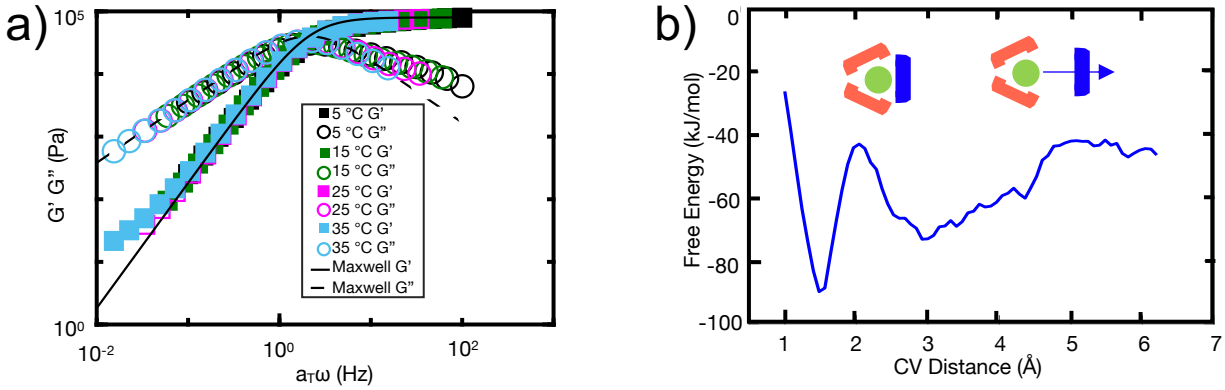


Figure S2-2. Frequency sweep and energy landscape data for His3Ni1. a) Time-temperature super-positioned storage (G') and loss moduli (G') from frequency sweeps collected over several temperatures for a 1M:2L ratio of Ni^{2+}–4arm PEG-histidine at pH 10. At pH 10, the His_3Ni_1 complex (inset) is expected to be dominant coordination complex. The gel demonstrates Maxwellian behavior. b) In the 2D free energy landscape demonstrated lig_1 and lig_2 remain bound to the ligand at a constant distance, corresponding to a state where the lowest energy ML_2 complex remains intact, while lig_3 dissociates. Figure from Ref.[216]

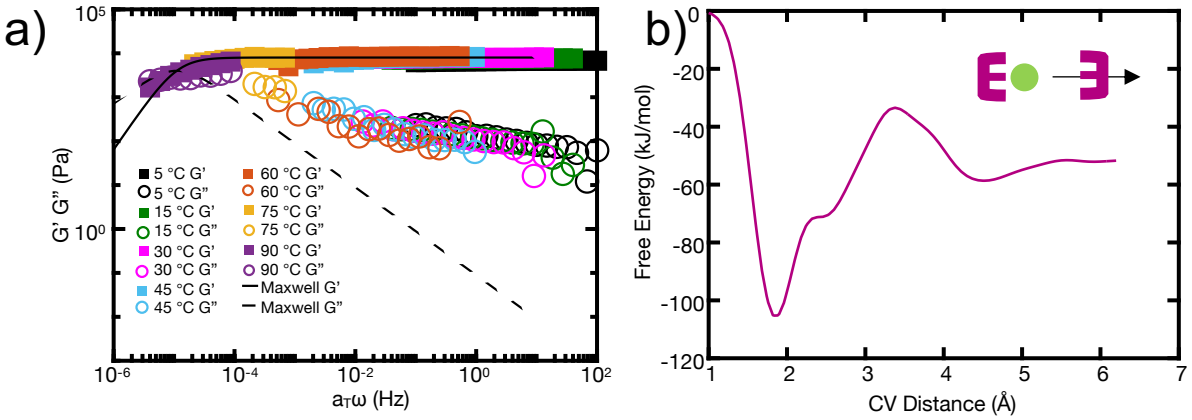


Figure S2-3. Frequency sweep for Tpy2Ni1. a) Time-temperature super-positioned storage (G') and loss moduli (G'') from frequency sweeps collected over several temperatures for a 1M:2L ratio of Ni^{2+}–4arm PEG-terpyridine. Tpy_2Ni_1 is expected to be dominant coordination complex. The gel demonstrates deviation from Maxwellian behavior in the high frequency regime. b) 2D free energy landscape of Tpy_2Ni_1. Figure from Ref.[216]

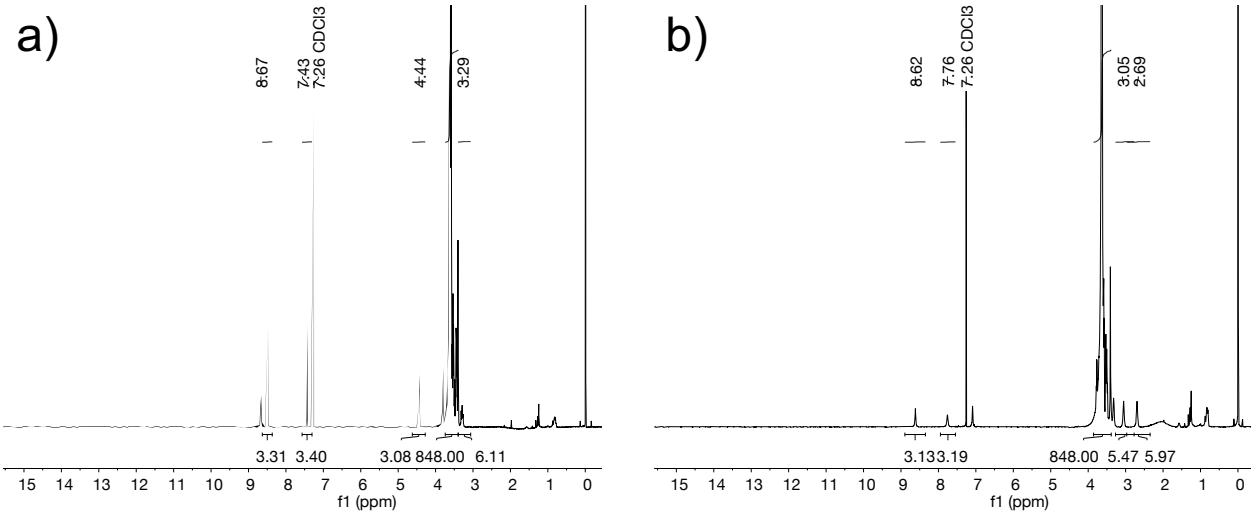


Figure S2-4. NMR of a) 4PEG-His, b) 4PEG-Imidazole, c) 4PEG-Terpyridine. Figure from Ref.[216]

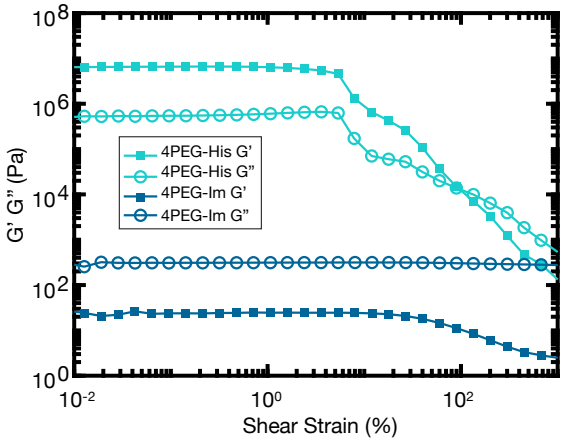


Figure S2-5. Strain sweep. Strain sweep of 1M:2L ratio of Ni^{2+} – 4-arm PEG-histidine and 1M:2L ratio of Ni^{2+} – 4-arm PEG-imidazole at pH ~8 at ~20% w/v at 5 °C confirms that the 1% shear strain amplitude used for the frequency sweeps is within the linear elastic regime of the hydrogels. Strain sweep at 1 rad/s. Figure from Ref.[216]

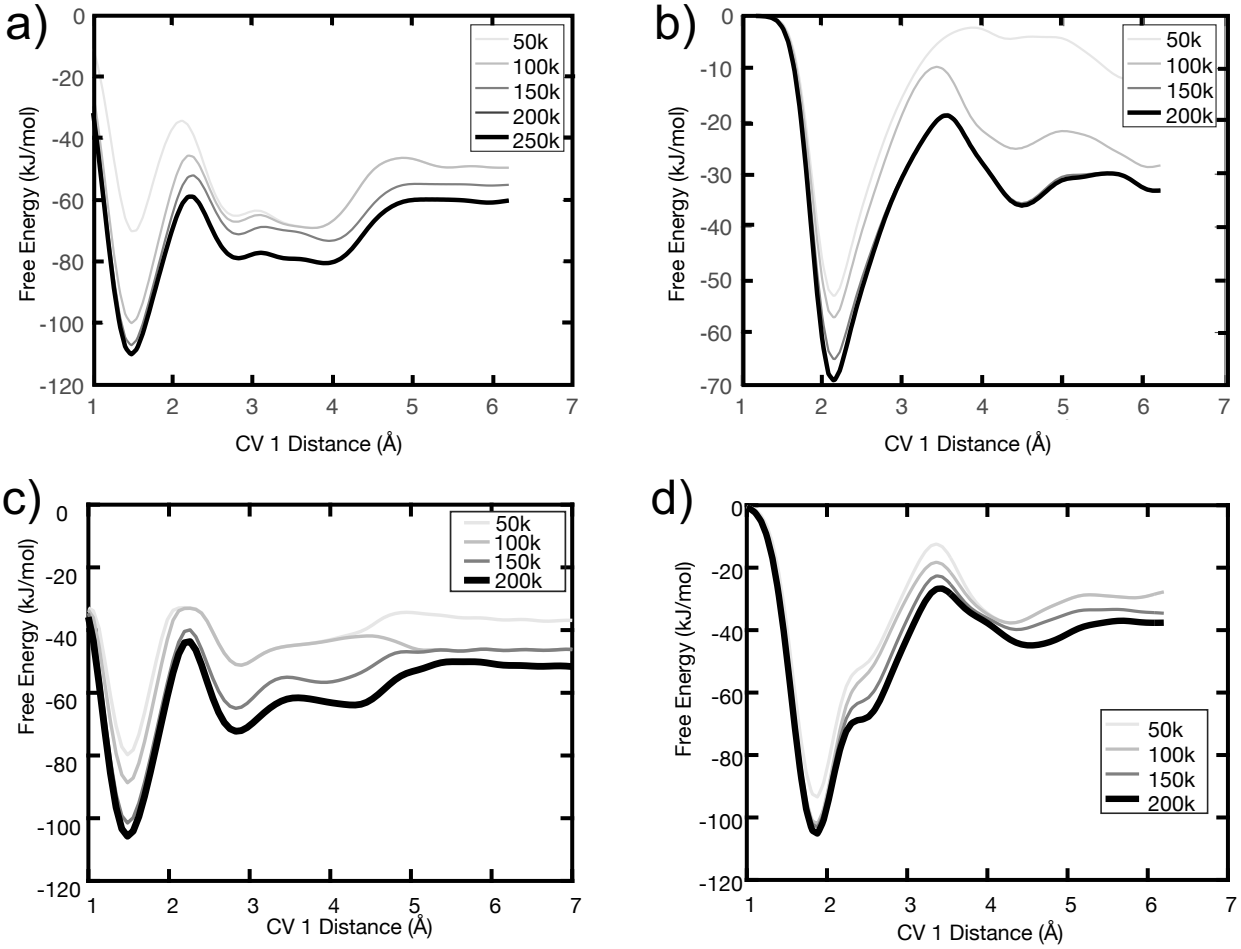


Figure S2-6. 2D energy landscapes of the ligand dissociation from the metal-coordination complex demonstrate convergence. Evolution of free energy landscape of a) His_2Ni_1 and b) Im_2Ni_1 as more Gaussian hills are added (grayscale lines). The collective variable (CV) is the distance between the center of mass of the coordinating nitrogen atoms of each ligand and metal ion. The free energy landscape is considered converged when the addition of more Gaussian hills does not change the shape of the resulting energy landscape. Convergence of the lowest energy well is reached quickly, while the whole energy landscape is converged once ~200,000 Gaussian hills have been added. Similar convergence is demonstrated for c) His_3Ni_1, where a representative collective variable d1 is shown, and d) Tpy_2Ni_1 upon the addition of at least 200,000 Gaussian hills. In this paper, energy landscapes are consistently compared at 200 ns of simulation, to ensure that at least 200,000 Gaussian hills are added. Figure in Ref.[216]

Codes

PLUMED LAMMPS run file:

```
# Created by charmm2lammps.pl ff_name structure_name

units           real
neigh_modify    delay 0 every 1
atom_style      full
```

```
#############################################
# POTENTIAL INFORMATION
bond_style      harmonic
angle_style     charmm
dihedral_style  charmm
improper_style  harmonic

pair_style      lj/charmm/coul/long 10.0 12.0
kspace_style      pppm 1.0e-4
pair_modify     mix arithmetic

read_restart       restart.equil2 # read from equilibrated structure

special_bonds   charmm

#################################################
timestep        2.0
thermo          100
thermo_style   custom step temp pe ke etotal press pxx pyy pzz lx ly lz

group       watH type 3
fix         fwatH watH shake 0.0001 20 10 t 3       #required for 2 fs timestep

#############################################
variable        temperature equal 300.0 # Simulation temperature in Kelvin
variable        tempDamp equal 100.0 # Relaxation time of thermostat
variable        pressure equal 1.0 # Pressure of box at 1 atm
variable        pressDamp equal       1000 # Relaxation time of pressure

fix             therm all nvt temp ${temperature} ${temperature} ${tempDamp}

dump        myDump all dcd 10000 out.dcd
dump_modify  myDump unwrap yes

reset_timestep          0
group           his id 1:20
fix             moment his momentum 1 linear 1 1 1
fix         plume all plumed plumedfile plumed.dat outfile plumed.out # Use plumed

run           5000000
unfix           plume
unfix           moment
```

PLUMED.dat file:

```
# vim:ft=plumed

RESTART
```

```
COM ATOMS=1,9 LABEL=com1 # desired atoms for collective variable
COM ATOMS=21,29 LABEL=com2
d1:  DISTANCE ATOMS=com1,41
d2:  DISTANCE ATOMS=com2,41

METAD ...
 LABEL=metad
 ARG=d1,d2
 SIGMA=0.02,0.02
 HEIGHT=1.
 PACE=500
 TEMP=300.0
 BIASFACTOR=6.0
... METAD

PRINT ARG=d1,d2 STRIDE=500 FILE=COLVAR
```

CHAPTER 3—SUPPORTING INFORMATION

Appendix S3-1 ; SI File 1 De Novo PDB strucctures. The set of initial representative *de novo* PDB structures and sample equilibration file, all included in a ZIP file. The PDB files are named based on the molecule name (H1.pdb, H2.pdb … HT4.pdb). A sample equilibration file is also included for reference.

Appendix S3-2 ; SI File 2 MetalPDB Database. MetalPDB database with the analysis of structures with 5-10 biologically-relevant metal ions is included in the Excel file.

Appendix Method 3-S1: DNA and amino acid sequences of the MC-ELP constructs

H1: **H1**-ELP-*linker*-***C*** (MW 14.78 kDa)
Encoding gene:
ATGGCTAGC**GGTCATGGTGTTCCA**GGTGTTGGTGTTCCAGGTAGCGGTGTGCCTGGTGTTG
GCGTGCCTGGCGTGGGCGTTCCCGGTGTAGGTGTGCCAGGCACCGGTGTTCCTGGCGTTG
GAGTACCAGGTGTCGGAGTTCCTGGTGTCGGTGTACCCGGTAGCGGCGTGCCAGGTGTTGG
TGTACCTGGTGTTGGTGTTCCAGGTAGCGGTGTGCCTGGTGTTGGCGTGCCTGGCGTGGGC
GTTCCCGGTGTAGGTGTGCCAGGCACCGGTGTTCCTGGCGTTGGAGTACCAGGTGTCGGAG
TTCCTGGTGTCGGTGTACCCGGTAGCGGCGTGCCAGGTGTTGGTGTACCTGGTGTTGGTGTT
CCAGGTAGCGGTGTGCCTGGTGTTGGCGTGCCTGGCGTGGGCGTTCCCGGTGTAGGTGTGC
CAGGCACCGGTGTTCCTGGCGTTGGAGTACCAGGTGTCGGAGTTCCTGGTGTCGGTGTACC
CGGTAGCGGCGTGCCAGGTGTTGGTGTACCTGGCCC*GGGGGTGGAGGCAGT****TGC***
Amino acid sequence:
MAS**GHGVP**GVGVPGSGVPGVGVPGVGVPGVGVPGTGVPGVGVPGVGVPGVGVPGSGVPGVGV
PGVGVPGSGVPGVGVPGVGVPGVGVPGTGVPGVGVPGVGVPGVGVPGSGVPGVGVPGVGVPG
SGVPGVGVPGVGVPGVGVPGTGVPGVGVPGVGVPGVGVPGSGVPGVGVPGP*GGGGS****C***

H2: **H2**-ELP-*linker*-***C*** (MW 15.22 kDa)
Encoding gene:
ATGGCTAGC**GGTCATGGTGTTCCAGGTCATGGTGTTCCA**GGTGTTCCAGGTGTTGGTGTTC
CAGGTAGCGGTGTGCCTGGTGTTGGCGTGCCTGGCGTGGGCGTTCCCGGTGTAGGTGTGCC
AGGCACCGGTGTTCCTGGCGTTGGAGTACCAGGTGTCGGAGTTCCTGGTGTCGGTGTACCC
GGTAGCGGCGTGCCAGGTGTTGGTGTACCTGGTGTTGGTGTTCCAGGTAGCGGTGTGCCTG
GTGTTGGCGTGCCTGGCGTGGGCGTTCCCGGTGTAGGTGTGCCAGGCACCGGTGTTCCTGG
CGTTGGAGTACCAGGTGTCGGAGTTCCTGGTGTCGGTGTACCCGGTAGCGGCGTGCCAGGT
GTTGGTGTACCTGGTGTTGGTGTTCCAGGTAGCGGTGTGCCTGGTGTTGGCGTGCCTGGCGT
GGGCGTTCCCGGTGTAGGTGTGCCAGGCACCGGTGTTCCTGGCGTTGGAGTACCAGGTGTC
GGAGTTCCTGGTGTCGGTGTACCCGGTAGCGGCGTGCCAGGTGTTGGTGTACCTGGCCC*GG*
*GGGTGGAGGCAGT****TGC***
Amino acid sequence:
MAS**GHGVPGHGVP**GVGVPGSGVPGVGVPGVGVPGVGVPGTGVPGVGVPGVGVPGVGVPGSG
VPGVGVPGVGVPGSGVPGVGVPGVGVPGVGVPGTGVPGVGVPGVGVPGVGVPGSGVPGVGVP
GVGVPGSGVPGVGVPGVGVPGVGVPGTGVPGVGVPGVGVPGVGVPGSGVPGVGVPGP*GGGGS*
C

H3: **H3**-ELP-*linker*-***C*** (MW 15.67 kDa)
Encoding Gene:
ATGGCTAGC**GGTCATGGTGTTCCAGGTCATGGTGTTCCAGGTCATGGTGTTCCA**GGTGTTC
CAGGTGTTGGTGTTCCAGGTAGCGGTGTGCCTGGTGTTGGCGTGCCTGGCGTGGGCGTTCC
CGGTGTAGGTGTGCCAGGCACCGGTGTTCCTGGCGTTGGAGTACCAGGTGTCGGAGTTCCT
GGTGTCGGTGTACCCGGTAGCGGCGTGCCAGGTGTTGGTGTACCTGGTGTTGGTGTTCCAG
GTAGCGGTGTGCCTGGTGTTGGCGTGCCTGGCGTGGGCGTTCCCGGTGTAGGTGTGCCAGG
CACCGGTGTTCCTGGCGTTGGAGTACCAGGTGTCGGAGTTCCTGGTGTCGGTGTACCCGGT
AGCGGCGTGCCAGGTGTTGGTGTACCTGGTGTTGGTGTTCCAGGTAGCGGTGTGCCTGGTG
TTGGCGTGCCTGGCGTGGGCGTTCCCGGTGTAGGTGTGCCAGGCACCGGTGTTCCTGGCGT
TGGAGTACCAGGTGTCGGAGTTCCTGGTGTCGGTGTACCCGGTAGCGGCGTGCCAGGTGTT
GGTGTACCTGGCCC*GGGGGTGGAGGCAGT****TGC***
Protein sequence:
MAS**GHGVPGHGVPGHGVP**GVGVPGSGVPGVGVPGVGVPGVGVPGTGVPGVGVPGVGVPGVG
VPGSGVPGVGVPGVGVPGSGVPGVGVPGVGVPGVGVPGTGVPGVGVPGVGVPGVGVPGSGVP
GVGVPGVGVPGSGVPGVGVPGVGVPGVGVPGTGVPGVGVPGVGVPGVGVPGSGVPGVGVPGP
*GGGGS****C***

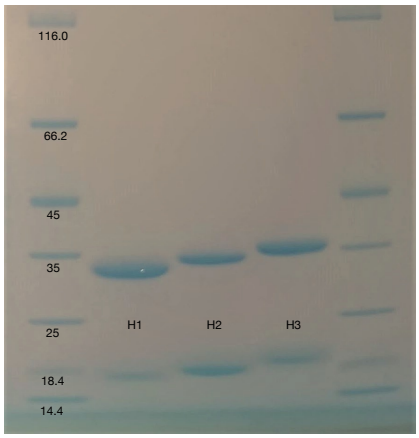


Figure S3-1. Non-reducing SDS-PAGE of purified MC-ELPs. The gel shows two bands for each of the MC-ELPs. The band with the lower molecular weight corresponds to the calculated molecular weight of H1 to H3. The second band has double the molecular weight and corresponds to disulfide-linked dimers. Note that the molecular weights are slightly higher than expected. This originates from the high content of hydrophobic amino acids in the MC-ELP sequences, which slow the electrophoretic mobility of the polypeptides in the gel.[485]

Table S3-1 Binding curves collected and analyzed. For each experiment, the total number of binding curves meeting the evaluation criteria (Kuhn lengths between 0.4 and 1.6 nm and a contour length cut-off of 100 nm), the total number of approach-retract cycles collected for each experiment and the binding frequency is listed.

MC-ELP	Retract Velocity (nm s^{-1})	Number of Binding Curves Meeting Criteria	Total Number of Approach-Retract Cycles Collected	Binding Frequency (%)
H1	200	335	7320	4.6
	400	172	5924	2.9

	1000	114	7907	1.4
	2500	89	6107	1.5
	5000	113	7047	1.6
	200	193	4697	4.1
	400	389	4472	8.7
	1000	301	5010	6.0
	2500	324	5008	6.5
	5000	240	5396	4.4
	200	204	3065	6.7
	400	166	4075	4.1
	1000	268	5178	5.2
	2500	211	5253	4.0
	5000	189	6150	3.1
H2	200	143	3427	4.2
	400	219	6271	3.5
	1000	233	4064	5.7
	2500	460	4791	9.6
	5000	680	4954	13.7
	200	271	3797	7.1
	400	496	4372	11.3
	1000	424	4793	8.8
	2500	530	5035	10.5
	5000	649	5249	12.4
	200	156	3323	4.7
	400	308	5766	5.3
	1000	322	5100	6.3
	2500	376	5440	6.9
	5000	477	5522	8.6
H3	200	345	6670	5.2
	400	213	5965	3.6
	1000	134	5968	2.2
	2500	293	5265	5.6
	5000	870	5120	17.0
	200	438	5991	7.3
	400	447	6012	7.4
	1000	536	5965	9.0
	2500	723	5053	14.3
	5000	190	5463	3.5
	200	256	5945	4.3
	400	468	5866	8.0
	1000	308	6528	4.7
	2500	342	6053	5.7
	5000	443	6158	7.2
H1-H3	200	221	3421	6.5
	400	364	3651	10.0
	1000	550	4431	12.4
	2500	410	4518	9.1

	5000	445	4379	10.2
	200	208	3166	6.6
	400	356	4199	8.5
	1000	568	4422	12.8
	2500	369	4448	8.3
	5000	466	4230	11.0
	200	316	3303	9.6
	400	249	4039	6.2
	1000	431	4429	9.7
	2500	394	4563	8.6
	5000	428	4260	10.0
EDTA	1000	0	973	0.0
	1000	0	971	0.0
	1000	3	1036	0.3
FsPl	1000	45	966	4.7
	1000	19	1289	1.5
	1000	12	966	1.2
PsFl	1000	0	960	0.0
	1000	6	1059	0.6
	1000	10	1250	0.8
PsPl	1000	1	1069	0.1
	1000	1	2576	0.0
	1000	2	2061	0.1

Codes

Example NAMD Equilibration File

```
#############################################################
## JOB DESCRIPTION                                         ##
#############################################################

# Minimization and Equilibration

#############################################################
## ADJUSTABLE PARAMETERS                                   ##
#############################################################

structure          solvate.psf
coordinates        solvate.pdb

set temperature    300 ; #K
set outputname     outnvt

# Continuing a job from the restart files
if {0} {
set inputname      equil3
binCoordinates     $inputname.restart.coor
```

```
binVelocities      $inputname.restart.vel  ;# remove the "temperature" entry if you use this!
extendedSystem          $inputname.restart.xsc
}

firsttimestep      0

#############################################################
## SIMULATION PARAMETERS                                   ##
#############################################################

# Input
paraTypeCharmm          on
parameters          par_all27_prot_na_babu.prm
temperature         $temperature

# Force-Field Parameters
exclude             scaled1-4
1-4scaling          1.0
cutoff              12.0
switching           on
switchdist          10.0
pairlistdist        16.0

# Integrator Parameters
timestep            2.0  ;# 2fs/step
rigidBonds          all  ;# needed for 2fs steps
nonbondedFreq       1
fullElectFrequency  2
stepspercycle       10


# Constant Temperature Control
if {1} {
langevin            on    ;# do langevin dynamics
langevinDamping     1     ;# damping coefficient (gamma) of 1/ps
langevinTemp        $temperature
langevinHydrogen    off    ;# don't couple langevin bath to hydrogens
}

# PBC
if {1} {
cellBasisVector1 90 0 0
cellBasisVector2 0 50 0
cellBasisVector3 0 0 50
}
cellOrigin 0 0 0
wrapWater on

# PME (for full-system periodic electrostatics)
```

```
PME                yes
PMEGridSpacing     1.0

# Constant Pressure Control (variable volume) -- for NPT
if {1} {
useGroupPressure      yes ;# needed for rigidBonds
useFlexibleCell       no
useConstantArea       no

langevinPiston        on
langevinPistonTarget  1.01325 ;#  in bar -> 1 atm
langevinPistonPeriod  100.0
langevinPistonDecay   50.0
langevinPistonTemp    $temperature
}

# Output
outputName          $outputname

restartfreq         1000
dcdfreq                  10000
xstFreq             1000
outputEnergies      1000
outputPressure      1000

minimize 100
run 3000000 ; #6ns
```

Example NAMD SMD File

```
structure          ionized.psf
coordinates        ionized.pdb

set temperature    300 ; #room temperature
set outputname     outsmd

# Continuing a job from the restart files
if {1} {
set inputname      outnvt2
binCoordinates     $inputname.restart.coor
binVelocities      $inputname.restart.vel  ;# remove the "temperature" entry if you use this!
extendedSystem         $inputname.restart.xsc
}

firsttimestep      0

#############################################################
## SIMULATION PARAMETERS                                   ##
#############################################################
```

```
# Input
paraTypeCharmm          on
parameters          par_all27_prot_na.prm

# Force-Field Parameters
exclude             scaled1-4
1-4scaling          1.0
cutoff              12.0
switching           on
switchdist          10.0
pairlistdist        14.0

# Integrator Parameters
timestep            2.0  ;# 2fs/step
rigidBonds          all  ;# needed for 2fs steps
nonbondedFreq       1
fullElectFrequency  2
stepspercycle       10


# Constant Temperature Control
if {1} {
  langevin            off    ;# do not do langevin dynamics
  langevinDamping     1     ;# damping coefficient (gamma) of 1/ps
  langevinTemp        $temperature
  langevinHydrogen    off    ;# don't couple langevin bath to hydrogens
}

# Periodic Boundary Conditions
if {0} {
cellBasisVector1 61 0 0
cellBasisVector2 0 82 0
cellBasisVector3 0 0 107
cellOrigin 37 -55 -45
}
wrapWater on

# PME (for full-system periodic electrostatics)
PME                 yes
PMEGridSpacing      1.0

# Constant Pressure Control (variable volume) -- for NPT
if {0} {
 useGroupPressure      yes ;# needed for rigidBonds
 useFlexibleCell       no
 useConstantArea       no

 langevinPiston        on
```

```
 langevinPistonTarget  1.01325 ;#  in bar -> 1 atm
 langevinPistonPeriod  100.0
 langevinPistonDecay   50.0
 langevinPistonTemp    $temperature
}

# Output
outputName          $outputname

restartfreq         5000     ;# 500steps = every 5ps
dcdfreq             500     ;# 500steps = every 5ps
xstFreq             5000
outputEnergies      5000
outputPressure      5000

#############################################################
## EXECUTION SCRIPT                                        ##
#############################################################
# Fixed Atoms Constraint (set PDB beta-column to 1)
if {1} {
fixedAtoms          on
fixedAtomsFile      fixed.ref
fixedAtomsCol       B
}

SMD    on
SMDFile fixed.ref
SMDk    7
SMDVel  0.0005
SMDDir 0.4050393894904376 -0.8826684563719812 0.2384103376263656
SMDOutputFreq 100

run 250000
```

CHAPTER 4—SUPPORTING INFORMATION

1. Initial PDB structure from Chou et al.[47]
2. Lowest energy PDB structures of simulation A, B, C
3. REMD simulation configuration files: See file from Supplementary Information of Ref (Khare, Luo, Buehler. *Soft Matter*)

CHAPTER 5—SUPPORTING INFORMATION

Appendix S5-1 ; Collagen dataset for training. Only 100 points included for demonstration purposes. See additional sequences in Ref.[486,487]

Sequences	Tm
GPOGPOGPOGPOGPQGPGGPPGPOGPOGPOGPOGPO	68
GPOGPOGPOGPOGPOGPOGPOGPOGPO	62
GPOGPOGPOGPOGPOGPOGPOGPOGPO	58
GPOGPOGPOGPOGAKGDAGPOGPOGPOGPOGPO	55
GPOGPOGPOGPOGPVGAAGATGPOGPOGPOGPOGPO	54
GPOGPOGPKGEOGPOGPOGPO	51
GPOGPOGPOGPOGPOGPOGPO	50.5
GPOGPOGPKGSOGPOGPOGPO	49.5
GPOGPOGPOGQOGLOGLOGPOGPOGPOGPO	48
GPOGPOGPKGDOGPOGPOGPO	48
GPOGPOGPOGPOGPOGPOGPOGPO	47.3
GPOGPOGPOGPOGPOGPOGPOGPO	47.3
GPOGPOGPOGPOGPOGPOGPOGPO	47.3
GPOGPOGPOGPRGPOGPOGPOGPO	47.2
GPOGPOGPOGPRGPOGPOGPOGPO	47.2
GPOGPOGPOGTOGPOGPOGPO	46.5
GPOGPOGPOGPOGPOGPOGPOGPO	45.5
GPOGPOGPOGPAGPOGPOGPOGPO	45.5
GPOGPOGPOGSOGPOGPOGPO	45.5
GPOGPOGPOGPPGPOGPOGPOGPO	45.5
GPOGPOGPOGPOGPOGPGSPPGPOGPOGPOGPOGPO	45
GPOGPOGPOGPOGPOGPOGPOGPO	44.5
GPOGPOGPOGPRGPOGPOGPOGPO	44.5
GPOGPOGPOGPOGPOGPOGPOGPO	44.5
GPOGPOGPOGPOGPOGPOGPOGPO	44.5
GPOGPOGPOGGQOGLOGLOGPOPOGPOGPOGGY	44
GPOGPOGPOGSOGPOGPOGPO	44
GPOGPOGPOGEOGPOGPOGPOGPO	42.9
GPOGPOGPOGEOGPOGPOGPOGPO	42.9
GPOGPOGPOGEOGPOGPOGPOGPO	42.9
GPOGPOGPOGAKGDAGPOGPAGPOGPOGPOGY	42.8
GPOGPOGPAGPOGPOGPAGPOGPO	42.8
GPOGPOGPOGPMGPOGPOGPOGPO	42.6
GPOGPOGPOGPMGPOGPOGPOGPO	42.6
GPOGPOGPOGPAGPOGPAGPOGPO	42.2
GPOGPOGPKGPOGPOGPOGPO	42
GPOGPOGPOGAOGPOGPOGPOGPO	41.7
GPOGPOGPOGAOGPOGPOGPOGPO	41.7
GPOGPOGPOGKOGPOGPOGPOGPO	41.5
GPOGPOGPOGKOGPOGPOGPOGPO	41.5
GPOGPOGPOGPIGPOGPOGPOGPO	41.5
GPOGPOGPOGKOGPOGPOGPOGPO	41.5
GPOGPOGPOGPIGPOGPOGPOGPO	41.5
GPOGPOGQOGLOGLOGPOGPOGPOGPO	41.3
GPOGPOGPOGPQGPOGPOGPOGPO	41.3
GPOGPOGPOGPQGPOGPOGPOGPO	41.3
GPOGPOGPOGEPGPOGPOGPOGPO	41.1
GPOGPOGPOGPAGPOGPOGPOGPO	40.9
GPOGPOGPOGPAGPOGPOGPOGPO	40.9
GPOGPOGPOGAOGPOGPOGPOGPO	40.6
GPOGPOGPOGROGPOGPOGPOGPO	40.6
GPOGPOGPOGROGPOGPOGPOGPO	40.6
GPOGPOGPOGROGPOGPOGPOGPO	40.6
GPOGPOGPOGPOGPAGPAGPOGPO	40.4
GPOGPOGPOGERGPOGPOGPOGPO	40.4
GPOGPOGPOGQOGPOGPOGPOGPO	40.4
GPOGPOGPOGQOGPOGPOGPOGPO	40.4
GPOGPOGPOGERGPOGPOGPOGPO	40.4
GPOGPOGPOGDOGPOGPOGPOGPO	40.1
GPOGPOGPOGDOGPOGPOGPOGPO	40.1
GPOGPOGPOGDOGPOGPOGPOGPO	40.1
GPOGPOGPOGPVGPOGPOGPOGPO	40
GPOGPOGPOGPVGPOGPOGPOGPO	40
GPOGPOGPOGAOGPOGPOGPOGPO	39.9
GPOGPOGPOGAOGPOGPOGPOGPO	39.9
GPOGPOGPOGAOGPOGPOGPOGPO	39.9
GPOGPOGPOGPEGPOGPOGPOGPO	39.7
GPOGPOGPOGPEGPOGPOGPOGPO	39.7

GPOGPOGPOGPTGPOGPOGPOGPO	39.7
GPOGPOGPOGKPGPOGPOGPOGPO	39.7
GPOGPOGPOGPEGPOGPOGPOGPO	39.7
GPOGPOGPOGPTGPOGPOGPOGPO	39.7
GPOGPOGPOGTOGPOGPOGPO	39.5
GPOGPOGPOGQRGPOGPOGPOGPO	39.5
GPOGPOGPOGKRGPOGPOGPOGPO	39.1
GPOGPOGPOGLOGPOGPOGPOGPO	39
GPOGPOGPOGLOGPOGPOGPOGPO	39
GPOGPOGPOGLOGPOGPOGPOGPO	39
GPOGPOGPOGVOGPOGPOGPOGPO	38.9
GPOGPOGPOGVOGPOGPOGPOGPO	38.9
GPOGPOGPOGKQGPOGPOGPOGPO	38.9
GPOGPOGPOGRPGPOGPOGPOGPO	38.8
GPOGPOGPOGMOGPOGPOGPOGPO	38.6
GPOGPOGPOGMOGPOGPOGPOGPO	38.6
GPOGPOGPOGQPGPOGPOGPOGPO	38.6
GPOGPOGPKGTOGPOGPOGPO	38.5
GLPGPRGEQGPTGPTGPAGPRGLQGLQGLQGERGEQGPTGPAGPRGLPGERGEQGPTGLAGKAGEAGAKGETGPAGPQGPRGEQGPQGLPGKDGEAGAQGPAGP MGPAGERGEKGEPGTQGAKGDRGETGPVGPRGERGEAGPAGKDGERGPVGPAGKDGQNGQDGLPGKDGKDGQNGKDGPGKDGKDGQNGKDGLPGKDGKDGQD GKDGLPGKDGKDGLPGKDGKDGQPGKPGKY	38.5
GPOGPOGPOGIOGPOGPOGPOGPO	38.4
GPOGPOGPOGIOGPOGPOGPOGPO	38.4
GPOGPOGPOGPAGPOGPOGPOGPO	38.3
GPOGPOGPOGPAGPOGPOGPOGPO	38.3
GPOGPOGPOGNOGPOGPOGPOGPO	38.3
GPOGPOGPOGNOGPOGPOGPOGPO	38.3
GPOGPOGPOGDPGPOGPOGPOGPO	38.3
GPOGPOGPOGSOGPOGPOGPOGPO	38.3
GLPGPRGEQGPTGPMGPAGPRGLQGLQGLQGERGEQGPTGPAGPRGLPGERGEQGPTGLAGKAGEAGAKGETGPAGPQGPRGEQGPQGLPGKDGEAGAQGPAG PMGPAGERGEKGEPGTQGAKGDRGETGPVGPRGERGEAGPAGKDGERGPVGPAGKDGQNGQDGLPGKDGKDGQNGKDGPGKDGKDGQNGKDGLPGKDGKDGQ DGKDGLPGKDGKDGLPGKDGKDGQPGKPGKY	38.3
GPOGPOGPOGAAGPOGPOGPOGPO	38.2
GPOGPOGPOGARGPOGPOGPOGPO	38.2
GPOGPOGPOGEMGPOGPOGPOGPO	38.2
GPOGPOGPOGSOGPOGPOGPOGPO	38

Appendix S5-2. Dataset of de novo collagen sequences generated for Tm = 22C by ColGen-GA. Only 100 points included for demonstration purposes. See additional sequences in Ref.[487]

Generated Sequence	Normalized Tm
GPOGPOGPOGMFGPOGPOGPOGPP	-0.7658614
GPVGLOGPOGGPOGPOGPO	-0.7398325
GPOGAOGPOGGFGLOGPOGPOGPO	-0.743013
GPOGPOGPOGSDGPNGDOGMYGPOGPOGPOGPO	-0.7581793
GERGPOGLAGAOGLRGGYGPOGPOGPOGWOGV	-0.7694818
GQQGPOGPOGNNGPOGPOGEOGLOGPOGPOGSOGPOGY	-0.7773197
GPFGPOGPOGWAGPEGPOGPOGPO	-0.7298877
GPOGPOGPOGSCGEOGPOGPOGPO	-0.7610745
GPOGAKGEOGDAGAKGDAGPOGPOGPOGPOGPO	-0.7520819
GPOGPOGYKGPOGPOGPOGPN	-0.7486301
GDRGETGPAGPAGPVGICGPOGPOGPOGPO	-0.7468727
GPOGPOGPSGDYGPOGPOGPOGPO	-0.7524183
GPOGPOGPOGHNGPOGSPGPOGPOGPOGPO	-0.7395569
GSOGPOGPOGLSGPOGPOGPOGPO	-0.7382498
GPOGPOGPOGGGGPAGPOGHOGPO	-0.7456985
GSOGPOGPOGYIGNOGPOGPOGPO	-0.7503122
GSOGPOGGCGKOGPLGPOGPO	-0.7236552
GPOGPOGYGGPOGPOGPOGPO	-0.7756709
GPOGPOGPVGNTGPOGPOGYOGPO	-0.7438831
GPOGPOGPOGGGGPOGHOGPOGPO	-0.7267381
GPOGIEGPOGSLGPOGPOGPOGPO	-0.7613491
GPOGPOGPOGFGGPOGVFGPOGPVGPOGPO	-0.7090974
GPDGPOGPTGRFGPOGPOGPOGPO	-0.7459955
GPOGPOGPDGIIGPOGPOGPOGPO	-0.7658652
GPOGPOGPCGWWGPOGPOGPCGPO	-0.7370111
GPOGPOGPOGFGGPOGWOGPOGPO	-0.7498701
GPNGPOGPOGWRGPOGOOGPOGPO	-0.7538912
GPOGPOGPOGNTGPOGIOGPOGPO	-0.7466145
GPOGPOGPOGSCGPOGPOGPOGPO	-0.7675912
GPOGLAGAOGLRGGAGROGPOGNOGPOGQOGV	-0.7301593
GPVGHOGPOGYFGPOGPOGPOGKO	-0.7654637
GPOGPOGPOGCFGPMGPOGPOGPO	-0.7743458
GAOGPOGDVGKOGPOGPTGPO	-0.7518646
GPOGPOGPOGSWGPOGPEGPFGPO	-0.7490631
GPOGPOGHOGQTGPOGPOGPOGPO	-0.7159497
GPNGPOGPVGGRGPOGPOGPOGPO	-0.7478068
GPOGPEGPOGHTGPWGPOGPOGPK	-0.7320225
GPOGPOGEOGGFGPOGPOGPOGPO	-0.7268328
GPOGDOGPCGPEGKTGPOGPOGPOGPO	-0.7407224
GDRGETGPAGPFGPVGPAGARGPOGPOGPOGPO	-0.7658292

GPOGPOGPOGWIGPOGPV	-0.7387197
GEOGQOGPOGMGGPOGPOGPOGPO	-0.7432636
GPOGAKGEOGDAGAKGDAGPOGPOGPOGPOGPO	-0.7520819
GPOGPOGPOGDOGAKADAGPOGPOGPOGPOGPO	-0.7369985
GPOGAOGROGAOGKVAPAGARGPOGPOGPOGPO	-0.7922283
GPOGAKGEOGDAGAKGDAGPOGPOGPOGPOGPO	-0.7520819
GPOGPOGPOGLVGPIGWOGPOGQO	-0.7466161
GPRGPRGPRGORGPRGPRGDRGPR	-0.7322026
GPOGPOGPOGKRGIOGPMGMOGPOGG	-0.7468322
GPOGPOGPOGSHGPHGPOGPPGPO	-0.7599311
GPOGPLGPOGKRGKPGPOGPOGPOGG	-0.7446287
GPOGPFGROGRLGROGPOGPOGPOGG	-0.747282
GAWGPAGPOGPOGCOGPOGPT	-0.735636
GPOGPOGPOGFGGPOGWOGPOGPO	-0.7498701
GPOGPLGPOGFRGPOGPOGPNGPO	-0.7777403
GPPGVOGPOGCNGPOGPOGCOGPY	-0.7532425
GDSGPRGLOGPOGMOGPQGFQGPOGEOGPOGVOGPM	-0.7557499
GPOGPOGPSGDYGPOGPVGPOGPO	-0.7760955
GLAGEOGKOGIOGNEGRAGPOGKRGKOGPOGPOGPOGG	-0.7530578
GPWGPOGPOGQTGPOGPOGPOGPO	-0.773415
GPOGAOGAOGAOGPVATAGARGPOGPOGPOGPO	-0.7708261
GLAGHDGKOGIOGLOGRAGPOGPOGPOGPOGV	-0.7546377
GPOGPOGPOGSKGLMGPOGPOGPO	-0.7408754
GPOGPSGPOGGNGPOGPOGPQGPO	-0.7657645
GEOGPOGTOGNIGPOGPOGPVGPOGPO	-0.7136348

Appendix S5-3. Dataset of de novo collagen sequences generated for Tm = 37C by ColGen-GA. Only 100 points included for demonstration purposes. See additional sequences in Ref.[487]

Generated Sequence	Normalized - Tm
GPOGPOGPOGTPGPOGKOGKOGPO	0.8976266
GPLGPOGPOGIOGPOGPOGPDGPOGPO	0.89387107
GPOGPOGPOGEPGPOGKOGIPGPOGPOGPPGYO	0.9225265
GPOGPOGPOGHEGQOGPOGPOGPOGPVGPO	0.8924112
GPOGPOGPOGFOGPOGPOGPOGFO	0.89561224
GPRGPOGPOGRPGPOGPOGPOGPY	0.91124725
GPOGEFGPHGLOGPOGPOGPOGPOGPOGPO	0.91380763
GPOGPSGPOGLAGPDGPOGPOGPOGHEGPOGHOGPOGPO	0.8967152
GPOGPOGPOGPMGPOGDOGPOGPO	0.8974596
GPOGPOGPOGPPGPOGPFGPOGPE	0.888
GPOGOOGPOGELGPOGPOGPOGPO	0.86755306
GPOGPOGPOGPQGPOGPOGPWGPO	0.94261396
GPOGPOGPOGCOGPOGPOGPOGPOGVO	0.89037347
GPOGPOGHOGPOGPOGPOGPO	0.8849506
GPOGMOGPOGVOGPOGPOGIOGOO	0.8328903
GPOGPOGPYGPOGPMGOOGPO	0.8727951
GPOGPOGPOGOPGPOGPOGPOGPH	0.88454163
GPOGDOGPOGPGPOGPOGPLGOGPO	0.90652144
GPOGPOGIOGPOGPOGPOGPO	0.9004248
GPOGPOGPOGVVGPOGPOGPOGPOGPO	0.89499044
GPOGPOGPOGIOGPOGPOGPOGPOGPOGSOGPEGPOGPO	0.90283203
GPOGPWGPOGROGPOGPOGPWGPO	0.88049686
GPOGPOGPSGPOGPOGWQGPOGPHGPOGPO	0.8764247
GPOGPOGPOGPTGPOGPAGPOGPOGPOGPOGSOGPP	0.93828905
GPOGPOGPOGPCGPOGPOGQOGPO	0.8695917
GPOGPOGPOGPVGPOGLYGPOGPO	0.93197936
GPOGPOGPOGPDGPOGPOGPOGPO	0.90410554
GPOGPOGPOGKOGPOGPOGPOGRO	0.9120938
GPYGPOGPOGRHGPOGROGDOGPWGPOGATGPOGPOGPOGPM	0.91180974
GPRGPOGPOGPVGOOGQOGPOGPO	0.85163856
GPDGPOGPOGPVGHOGPOGPFGPO	0.8786826
GPOGPDGPOGAOGPOGPOGPOGPO	0.937035
GYOGPOGPOGELGPOGPQGPQGPO	0.90398335
GPOGPOGPOGWVGPOGPOGPO	0.86969304
GPOGPOGPOGPDGPOGPOGPOGPO	0.90410554
GPOGAOGPOGDOGPOGPOGPO	0.8907305
GPOGPOGPOGPAGPRGPOGPOGCOGPO	0.9090091
GPOGPOGPOGDMGPOGPAGPOGPOGPOGPO	0.8784678
GPOGPEGPOGROGPOGPOGPOGPO	0.8967092
GPOGPOGPOGROGPOGPIGWOGPO	0.89294904
GPOGPOGPOGPOGPOGQHGPOGPOGPOGPO	0.8938109
GPOGPOGPOGROGPDGPOGPOGTO	0.8918484
GPOGPOGPOGPYGPOGPOGPOGPW	0.90054035
GPOGPOGTOGRPGPOGPKGPOGWO	0.87852454
GPLGPOGPOGPDGPOGWOGPOGPO	0.9125881
GPNGPOGPOGLOGPOGPOGPRGPO	0.9235561
GPIGPOGPOGQOGPOGPOGPOGNO	0.86403644
GPOGPOGPOGIOGPOGPOGPAGPO	0.913179
GPOGPOGPOGPDGPOGPOGPOGPO	0.90410554
GPOGPOGPOGELGPOGPOGPOGPV	0.91464484
GPOGPOGPOGPLGPAGPOGPOGPOGPO	0.9151588

GPOGPOGPOGEGPOGMOGPGGPOGPOGPDGQO	0.89381975
GPOGPOGPOGPEGPOGPOGPOGPO	0.9146085
GPOGPOGPOGPEGPOGPOGPOGPO	0.91460866
GPSGPOGPOGIQGPOGPOGKLGPOGPOGPOGPO	0.88644147
GPOGPOGPOGFEGPOGPOGQOGPTGPOGPOGPF	0.88502705
GPOGPOGPOGHLGPOGPOGPWGPO	0.8812223
GPOGPOGPOGPDGPOGPOGPOGPO	0.90410554
GMOGPSGPOGOOGPOGPYGPOGPO	0.93688023
GPOGPOGPOGPEGPOGPOGPOGPO	0.91460866
GPOGPOGPOGPDGPOGPOGPOGPO	0.9041054
GPOGPOGPOGPDGPOGPOGPOGPO	0.90410554
GAOGPOGPOGPDGPOGPFGPOGPO	0.9057503
GROGPOGPOGPAGPOGPOGPOGPO	0.9086806
GPOGPOGPRGDWGPOGPOGVMGPOGPOGPOGPO	0.9641918
GPOGPOGPOGPEGPOGPOGPOGPO	0.91460866
GPTGPOGPOGEPGPOGPOGPOGPO	0.87310445
GPOGPOGPOGILGPOGPMGPOGPO	0.8078369
GPOGPOGPOGPRGPOGPOGPOGPO	0.8756664
GPAGPOGPOGLOGPOGPOGPOGPO	0.91208255
GPOGPOGPOGPEGPOGPOGPOGPO	0.91460866
GPOGPOGPOGTVGPOGNOGPO	0.8756932
GPOGPOGPOGPEGPOGPOGPOGPO	0.91460866
GPVGPOGPOGERGPOGPQGPOGQFGLOGLOGPOGPOGPOGPO	0.92844445
GPOGPOGPOGAOGPOGDOGPOGPO	0.90816784
GPOGPOGHOGPOGPOGPOGPO	0.8849506
GPOGPOGPOGRPGPOGPOGPOGPO	0.92474985
GPOGPRGPOGPHGPOGPIGPOGPWGQMGPOGPOGPO	0.9045195
GSOGPOGPOGCMGPOGPOGPOGPW	0.86980855
GPOGPOGPOGAOGPOGDOGPOGPO	0.90816784
GPOGPOGPOGPOGPAGCAGPOGPO	0.8935522
GPOGPOGPOGPDGPOGPOGPOGPO	0.90410554
GPOGPOGPOGPDGPOGPOGPOGPO	0.90410554
GPOGFOGPOGERGPOGPOGFOGPOGPOGPOGPO	0.89464045
GPOGPOGPOGPLGTOGPOGPWGYOGPO	0.91672325
GPOGPOGPOGPRGPMGPOGPOGPO	0.8776943
GPDGPOGPOGPLGPOGDOGPOGPA	0.90623176
GPOGPOGPOGDOGPOGPOGPOGTO	0.91527426
GPOGPOGPOGEHGPOGPOGPOGPOGPO	0.8977125
GPOGPOGPOGEAGPOGPOGHOGPO	0.9273832
GPOGPOGPOGAOGPOGDOGPOGPO	0.90816784
GPAGPOGPOGPAGPOGPOGPOGPO	0.9142337
GPEGPOGPOGPDGPOGPOGPOGPY	0.88918996
GPOGPOGPOGLKGPFGPOGPOGPOGPWGPFGYOGPOGPO	0.9020096
GPOGPOGPOGPEGPOGPYGPOGPO	0.8511671
GPOGAOGPOGLOGPOGPOGPOGPO	0.92471576
GPOGPOGPOGLHGPOGPOGPOGPW	0.89278805
GSOGPOGPOGEAGPOGPOGPOGPO	0.894419
GPOGPOGPOGPEGPOGPOGPOGPO	0.91460866
GROGPOGPOGMAGPOGTOGPO	0.88845265

Appendix S5-4; Jupyter Notebook Transformer 1. See file from Supplementary Information of Ref.[486]

Appendix S5-5; Jupyter Notebook Transformer 2. See file from Supplementary Information of Ref.[486]

Appendix S5-6: Validation Sequences for Transfomer

Sequences	Tm
GPOGPOGPDGPOGKPGPOGPOGPOG	38
GPOGPOGPDGKOGPOGPOGPOGPOG	46
GPOGPOGPKGFOGPOGPOGPOG	28.8
GPOGPOGPKGYOGPOGPOGPOG	28.4
GPOGPOGPKGWOGPOGPOGPOG	26
GPOGPOGPRGFOGPOGPOGPOG	41
GPOGPOGPRGYOGPOGPOGPOG	41.41
GPOGPOGPRGWOGPOGPOGPOG	39.2
GPOGPOGPFGROGPOGPOGPOG	21.5
GPRGPOGPRGPOGPRGPOGPRGPOG	49
GPOGPOGPOGPOGPOGPOGPOG	49.5
GPOGPOGPOGDOGPOGPOGPOG	41.5
GPOGPOGPOGEOGPOGPOGPOG	44.5
GPOGPOGPOGFOGPOGPOGPOG	37
GPOGPOGPKGPOGPOGPOGPOG	40
GPOGPOGPRGPOGPOGPOGPOG	47
GPOGPOGPOGDKGPOGPOGPOG	32
GPOGPOGPOGEKGPOGPOGPOG	37
GPOGPOGPOGFKGPOGPOGPOG	22
GPOGPOGPOGDRGPOGPOGPOG	38
GPOGPOGPOGERGPOGPOGPOG	42.5

GPOGPOGPOGFRGPOGPOGPOG	29.5
GPOGPOGPKGDOGPOGPOGPOG	47
GPOGPOGPKGEOGPOGPOGPOG	44
GPOGPOGPKGFOGPOGPOGPOG	29.5
GPOGPOGPRGDOGPOGPOGPOG	42
GPOGPOGPRGEOGPOGPOGPOG	42.5
GPOGPOGPRGFOGPOGPOGPOG	41.5
GPOGPOGPOGDPGPOGPOGPOG	36.5
GPOGPOGPOGEPGPOGPOGPOG	39.5
GPOGPOGPOGFPGPOGPOGPOG	31
GPOGPOGPOGPPGPOGPOGPOG	47
GPOGPOGPOGWOGPOGPOGPOG	33.5
GPOGPOGPHGPOGPOGPOGPOG	38
GPOGPOGWHGPOGPOGPOGPOG	14.5
GPOGPOGPHGWOGPOGPOGPOG	22.5
GPOGPOGPOGAOGPOGPOGPOG	45
GPOGPOGPOGLOGPOGPOGPOG	43.5
GPOGPOGPOGVOGPOGPOGPOG	41.5
GPOGPOGPAGPOGPOGPOGPOG	44
GPOGPOGPFGPOGPOGPOGPOG	31
GPOGPOGPVGPOGPOGPOGPOG	43.5
GPOGPOGPLGPOGPOGPOGPOG	36
GPOGPOGPAGAOGPOGPOGPOG	36.5
GPOGPOGPAGFOGPOGPOGPOG	29
GPOGPOGPAGVOGPOGPOGPOG	33.5
GPOGPOGPAGLOGPOGPOGPOG	34.4
GPOGPOGPFGAOGPOGPOGPOG	25.6
GPOGPOGPFGFOGPOGPOGPOG	19.5
GPOGPOGPFGLOGPOGPOGPOG	23.5
GPOGPOGPFGVOGPOGPOGPOG	18
GPOGPOGPLGAOGPOGPOGPOG	29
GPOGPOGPLGFOGPOGPOGPOG	24
GPOGPOGPLGLOGPOGPOGPOG	28.5
GPOGPOGPLGVOGPOGPOGPOG	19.5
GPOGPOGPVGAOGPOGPOGPOG	37.5
GPOGPOGPVGFOGPOGPOGPOG	30.1
GPOGPOGPVGLOGPOGPOGPOG	34.5
GPOGPOGPVGVOGPOGPOGPOG	30.6
GPKGFOGPOGFKGFKGPKGPOGFKGPOG	21
GPOGPOGPOGKDGKDGQPGKPGTPGPQGIAGQRGVVGLPGPRGEQGPTGPTGPOGPOGPO	36
GPOGPOGPOGKDGKDGQPGKPGTPGPQGLLGAPGILGLPGPRGEQGPTGPTGPOGPOGPO	36
GPOGPOGPOGKDGKDGQPGKPGPPGPQGLAGQRGIVGLPGPRGEQGPTGPTGPOGPOGPO	36
GPOGPOGPOGKDGKDGQPGKPGAPGPLGIAGITGARGLAGPRGEQGPTGPTGPOGPOGPO	36
GPOGPOGPQGPOGPOGPOGPOG	45
GPOGPOGPNGPOGPOGPOGPOG	34
GPOGPOGPOGFAGPOGPOGPOG	22.5
GPOGPOGPOGFQGPOGPOGPOG	24.5
GPOGPOGPQGFOGPOGPOGPOG	32.5
GPOGPOGPOGFNGPOGPOGPOG	19.5
GPOGPOGPNGFOGPOGPOGPOG	21.5

Bibliography

1 H. D. Espinosa, J. E. Rim, F. Barthelat, M. J. Buehler and H. D. Espinosa, *Prog. Mater. Sci.*, 2009, **54**, 1059–1100.
2 R. O. Ritchie, *Nat. Mater.*, 2011, **10**, 817–822.
3 P. Ball, *Nature*, 2001, **409**, 413–416.
4 R. J. . Williams, *Biol. Rev.*, 1953, **28**, 381–412.
5 R. J. Sundberg and R. B. Martin, *Chem. Rev.*, 1974, **74**, 471–517.
6 A. M. Pyle, *J. Biol. Inorg. Chem.*, 2002, **7**, 679–690.
7 J. H. Waite, in *Integrative and Comparative Biology*, 2002, vol. 42, pp. 1172–1180.
8 C. J. Sun and J. H. Waite, *J. Biol. Chem.*, 2005, **280**, 39332–39336.
9 N. Holten-Andersen, M. J. Harrington, H. Birkedal, B. P. Lee, P. B. Messersmith, K. Y. C. Lee and J. H. Waite, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 2651–5.
10 D. E. Fullenkamp, L. He, D. G. Barrett, W. R. Burghardt and P. B. Messersmith, *Macromolecules*, 2013, **46**, 1167–1174.
11 C. H. Li, C. Wang, C. Keplinger, J. L. Zuo, L. Jin, Y. Sun, P. Zheng, Y. Cao, F. Lissel, C. Linder, X. Z. You and Z. Bao, *Nat. Chem.*, 2016, **8**, 618–624.
12 M. A. Hartmann and P. Fratzl, *Nano Lett.*, 2009, **9**, 3603–3607.
13 C. C. Broomell, S. F. Chase, T. Laue and J. H. Waite, *Biomacromolecules*, 2008, **9**, 1669–1677.
14 C. C. Broomell, F. W. Zok and J. H. Waite, *Acta Biomater.*, 2008, **4**, 2045–2051.
15 C. C. Broomell, M. A. Mattoni, F. W. Zok and J. H. Waite, *J. Exp. Biol.*, 2006, **209**, 3219–3225.
16 J. H. Waite, N. H. Andersen, S. Jewhurst and C. Sun, *J. Adhes.*, 2005, **81**, 297–317.
17 E. Degtyar, M. J. Harrington, Y. Politi and P. Fratzl, *Angew. Chemie - Int. Ed.*, 2014, 53, 12026–12044.
18 U. G. K. Wegst, H. Bai, E. Saiz, A. P. Tomsia and R. O. Ritchie, *Nat. Mater.*, 2015, **14**, 23–36.
19 M. E. Launey and R. O. Ritchie, *Adv. Mater.*, 2009, **21**, 2103–2110.
20 C. Creton, *Macromolecules*, 2017, 50, 8297–8316.
21 G. E. Fantner, T. Hassenkam, J. H. Kindt, J. C. Weaver, H. Birkedal, L. Pechenik, J. A. Cutroni, G. A. G. Cidade, G. D. Stucky, D. E. Morse and P. K. Hansma, *Nat. Mater.*, 2005, **4**, 612–616.
22 F. Barthelat, Z. Yin and M. J. Buehler, *Nat. Rev. Mater.*, 2016, 1, 16007.
23 G. X. Gu, M. Takaffoli and M. J. Buehler, *Adv. Mater.*, 2017, **29**, 1–7.
24 N. Holten-Andersen, A. Jaishankar, M. J. Harrington, D. E. Fullenkamp, G. Dimarco, L. He, G. H. McKinley, P. B. Messersmith and K. Y. C. Lee, *J. Mater. Chem. B*, 2014, **2**, 2467–2472.
25 S. C. Grindy, R. Learsch, D. Mozhdehi, J. Cheng, D. G. Barrett, Z. Guan, P. B. Messersmith and N. Holten-Andersen, *Nat. Mater.*, 2015, **14**, 1210–1216.
26 E. Khare, N. Holten-Andersen and M. J. Buehler, *Nat. Rev. Mater.*, 2021, 1–16.
27 D. Mozhdehi, J. A. Neal, S. C. Grindy, Y. Cordeau, S. Ayala, N. Holten-Andersen and Z. Guan, *Macromolecules*, 2016, **49**, 6310–6321.
28 M. J. Harrington, A. Masic, N. Holten-Andersen, J. H. Waite and P. Fratzl, *Science (80-.).*, 2010, **328**, 216–220.
29 S. C. Grindy, M. Lenz and N. Holten-Andersen, *Macromolecules*, 2016, **49**, 8306–8312.
30 H. Li, P. Yang, P. Pageni and C. Tang, *Macromol. Rapid Commun.*, 2017, 38.
31 S. A. Cazzell and N. Holten-Andersen, *Proc. Natl. Acad. Sci.*, 2019, **22**, 201906349.
32 M. J. Buehler, *MRS Bull.*, 2013, 38, 169–176.
33 T. Giesa, R. Jagadeesan, D. I. Spivak and M. J. Buehler, *ACS Biomater. Sci. Eng.*, 2015, **1**, 1009–1015.
34 W. Huang, D. Ebrahimi, N. Dinjaski, A. Tarakanova, M. J. Buehler, J. Y. Wong and D. L. Kaplan, *Acc. Chem. Res.*, 2017, **50**, 866–876.
35 D. Mozhdehi, S. Ayala, O. R. Cromwell and Z. Guan, *J. Am. Chem. Soc.*, 2014, **136**, 16128–16131.
36 S. Y. Zheng, H. Ding, J. Qian, J. Yin, Z. L. Wu, Y. Song and Q. Zheng, *Macromolecules*, 2016, **49**, 9637–9646.
37 Q. Tang, D. Zhao, H. Yang, L. Wang and X. Zhang, *J. Mater. Chem. B*, 2019, **7**, 30–42.
38 J. Yeo, G. S. Jung, A. Tarakanova, F. J. Martín-Martínez, Z. Qin, Y. Cheng, Y. W. Zhang and M. J. Buehler, *Extrem. Mech. Lett.*, 2018, **20**, 112–124.
39 K. Bertoldi and M. C. Boyce, *J. Mater. Sci.*, 2007, **42**, 8943–8956.
40 N. Cohen, J. H. Waite, R. M. McMeeking and M. T. Valentine, *Philos. Trans. R. Soc. B Biol. Sci.*, 2019, **374**, 20190202.
41 T. Priemel, E. Degtyar, M. N. Dean and M. J. Harrington, *Nat. Commun.*, 2017, **8**, 14539.
42 E. C. Bell and J. M. Gosline, *J. Exp. Biol.*, 1996, **199**, 1005–1017.

43 E. Carrington and J. M. Gosline, *Am. Malacol. Bull.*, 2004, **18**, 135–142.
44 J. H. Waite, H. C. Lichtenegger, G. D. Stucky and P. Hansma, *Biochemistry*, 2004, 43, 7653–7662.
45 Z. Qin and M. J. Buehler, *Nat. Commun.*, 2013, **4**, 1–8.
46 M. J. Harrington and J. H. Waite, *J. Exp. Biol.*, 2007, **210**, 4307–4318.
47 C. C. Chou, F. J. Martin-Martinez, Z. Qin, P. B. Dennis, M. K. Gupta, R. R. Naik and M. J. Buehler, *ACS Nano*, 2017, **11**, 1858–1868.
48 H. Birkedal, R. K. Khan, N. Slack, C. C. Broomell, H. C. Lichtenegger, F. Zok, G. D. Stucky and J. H. Waite, *ChemBioChem*, 2006, **7**, 1392–1399.
49 H. Zhao and J. H. Waite, *Biochemistry*, 2006, **45**, 14223–14231.
50 E. Vaccaro and J. H. Waite, *Biomacromolecules*, 2001, **2**, 906–911.
51 S. Zechel, M. Hager, T. Priemel and M. J. Harrington, *Biomimetics*, 2019, **4**, 20.
52 C. N. Z. Schmitt, Y. Politi, A. Reinecke and M. J. Harrington, *Biomacromolecules*, 2015, **16**, 2852–2861.
53 S. Schmidt, A. Reinecke, F. Wojcik, D. Pussak, L. Hartmann and M. J. Harrington, *Biomacromolecules*, 2014, **15**, 1644–1652.
54 S. Krauss, T. H. Metzger, P. Fratzl and M. J. Harrington, *Biomacromolecules*, 2013, **14**, 1520–1528.
55 A. Reinecke, L. Bertinetti, P. Fratzl and M. J. Harrington, *J. Struct. Biol.*, 2016, **196**, 329–339.
56 K. U. Claussen, R. Giesa, T. Scheibel and H. W. Schmidt, *Macromol. Rapid Commun.*, 2011, **33**, 206–211.
57 N. Holten-Andersen, T. E. Mates, M. S. Toprak, G. D. Stucky, F. W. Zok and J. H. Waite, *Langmuir*, 2009, **25**, 3323–3326.
58 J. Saiz-Poseu, J. Mancebo-Aracil, F. Nador, F. Busqué and D. Ruiz-Molina, *Angew. Chemie - Int. Ed.*, 2019, 58, 696–714.
59 A. H. Hofman, I. A. van Hees, J. Yang and M. Kamperman, *Adv. Mater.*, 2018, 30, 1704640.
60 P. Kord Forooshani and B. P. Lee, *J. Polym. Sci. Part A Polym. Chem.*, 2017, 55, 9–33.
61 H. C. Lichtenegger, T. Schöberl, J. T. Ruokolainen, J. O. Cross, S. M. Heald, H. Birkedal, J. H. Waite and G. D. Stucky, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 9144–9149.
62 H. C. Lichtenegger, T. Schoberl, M. H. Bartl, H. Waite and G. D. Stucky, *Science (80-.).*, 2002, **298**, 389–392.
63 R. K. Khan, P. K. Stoimenov, T. E. Mates, J. H. Waite and G. D. Stucky, *Langmuir*, 2006, **22**, 8465–8471.
64 Y. Politi, M. Priewasser, E. Pippel, P. Zaslansky, J. Hartmann, S. Siegel, C. Li, F. G. Barth and P. Fratzl, *Adv. Funct. Mater.*, 2012, **22**, 2519–2528.
65 R. M. S. Schofield, M. H. Nesson, K. A. Richardson and P. Wyeth, *J. Insect Physiol.*, 2003, **49**, 31–44.
66 N. Horbelt, M. Eder, L. Bertinetti, P. Fratzl and M. J. Harrington, *Biomacromolecules*, 2019, **20**, 3094–3103.
67 D. Pasche, N. Horbelt, F. Marin, S. Motreuil, E. Macías-Sánchez, G. Falini, D. S. Hwang, P. Fratzl and M. J. Harrington, *Soft Matter*, 2018, **14**, 5654–5664.
68 T.-M. Fung, C. Gallego Lazo and A. M. Smith, *Philos. Trans. R. Soc. B Biol. Sci.*, 2019, **374**, 20190201.
69 W. N. Lipscomb and N. Sträter, *Chem. Rev.*, 1996, **96**, 2375–2433.
70 S. E. Hromada, A. M. Hilbrands, E. M. Wolf, J. L. Ross, T. R. Hegg, A. G. Roth, M. T. Hollowell, C. E. Anderson and D. E. Benson, *J. Inorg. Biochem.*, 2017, **176**, 168–174.
71 M. Yu, J. Hwang and T. J. Deming, *J. Am. Chem. Soc.*, 1999, **121**, 5825–5826.
72 H. Yamamoto, *J. Chem. Soc. Perkin Trans. 1*, 1987, 613–618.
73 G. A. Parada and X. Zhao, *Soft Matter*, 2018, **14**, 5186–5196.
74 Q. Tang, D. Zhao, Q. Zhou, H. Yang, K. Peng and X. Zhang, *Macromol. Rapid Commun.*, 2018, **39**, 1800109.
75 M. Yu and T. J. Deming, *Macromolecules*, 1998, **31**, 4739–4745.
76 G. E. Sanoja, N. S. Schauser, J. M. Bartels, C. M. Evans, M. E. Helgeson, R. Seshadri and R. A. Segalman, *Macromolecules*, 2018, **51**, 2017–2026.
77 W. Guo, X. Wang, X. Lu, X. Li, Y. Li and J. Sun, *J. Mater. Chem. A*, 2019, **7**, 21927–21933.
78 A. Srivastava, N. Holten-Andersen, G. D. Stucky and J. H. Waite, *Biomacromolecules*, 2008, **9**, 2873–2880.
79 M. Enke, S. Bode, J. Vitz, F. H. Schacher, M. J. Harrington, M. D. Hager and U. S. Schubert, *Polymer (Guildf).*, 2015, **69**, 274–282.
80 M. Enke, R. K. Bose, S. Zechel, J. Vitz, R. Deubler, S. J. Garcia, S. Van Der Zwaag, F. H. Schacher, M. D. Hager and U. S. Schubert, *Polym. Chem.*, 2018, **9**, 3543–3551.
81 R. Wang, M. K. Sing, R. K. Avery, B. S. Souza, M. Kim and B. D. Olsen, *Acc. Chem. Res.*, 2016, **49**, 2786–2795.
82 T. Rossow and S. Seiffert, *Polym. Chem.*, 2014, **5**, 3018–3029.
83 S. Tang and B. D. Olsen, *Macromolecules*, 2016, **49**, 9163–9175.
84 W. Yu, W. Sun, Q. Fan, B. Xue, Y. Li, M. Qin, Y. Li, B. Chen, W. Wang and Y. Cao, *Soft Matter*, 2019, **15**, 4423–4427.
85 M. Ahmadi, A. Jangizehi, E. van Ruymbeke and S. Seiffert, *Macromolecules*, 2019, **52**, 5255–5267.
86 D. S. Hwang, H. J. Yoo, J. H. Jun, W. K. Moon and H. J. Cha, *Appl. Environ. Microbiol.*, 2004, **70**, 3352–

3359.
87 D. S. Hwang, S. B. Sim and H. J. Cha, *Biomaterials*, 2007, **28**, 4039–4046.
88 B. J. Kim, Y. S. Choi and H. J. Cha, *Angew. Chemie - Int. Ed.*, 2012, **51**, 675–678.
89 B. J. Kim, S. Kim, D. X. Oh, A. Masic, H. J. Cha and D. S. Hwang, *J. Mater. Chem. B*, 2013, 1–3.
90 B. J. Kim, D. X. Oh, S. Kim, J. H. Seo, D. S. Hwang, A. Masic, D. K. Han and H. J. Cha, *Biomacromolecules*, 2014, **15**, 1579–1585.
91 M. K. Gupta, K. A. Becknell, M. G. Crosby, N. M. Bedford, J. Wright, P. B. Dennis and R. R. Naik, *ACS Appl. Mater. Interfaces*, 2018, **10**, 31928–31937.
92 A. R. Statz, R. J. Meagher, A. E. Barron and P. B. Messersmith, *J. Am. Chem. Soc.*, 2005, **127**, 7972–7973.
93 W. Yang, X. Gao, G. Springsteen and B. Wang, *Tetrahedron Lett.*, 2002, **43**, 6339–6342.
94 A. Trapaidze, M. D'Antuono, P. Fratzl and M. J. Harrington, *Eur. Polym. J.*, 2018, **109**, 229–236.
95 F. Jehle, P. Fratzl and M. J. Harrington, *ACS Nano*, 2018, **12**, 2160–2168.
96 X.-W. Wang, D. Liu, G.-Z. Yin and W.-B. Zhang, in *Bioinspired Materials Science and Engineering*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2018, pp. 295–309.
97 A. Reinecke, G. Brezesinski and M. J. Harrington, *Adv. Mater. Interfaces*, 2017, **4**, 1600416.
98 D. E. Przybyla and J. Chmielewski, *J. Am. Chem. Soc.*, 2010, **132**, 7866–7867.
99 E. N. Salgado, X. I. Ambroggio, J. D. Brodin, R. A. Lewis, B. Kuhlman and F. A. Tezcan, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 1827–1832.
100 E. N. Salgado, R. J. Radford and F. A. Tezcan, *Acc. Chem. Res.*, 2010, **43**, 661–672.
101 J. D. Brodin, X. I. Ambroggio, C. Tang, K. N. Parent, T. S. Baker and F. A. Tezcan, *Nat. Chem.*, 2012, **4**, 375–382.
102 I. Tunn, M. J. Harrington and K. G. Blank, *Biomimetics*, 2019, **4**, 25.
103 W. Sun, B. Xue, Q. Fan, R. Tao, C. Wang, X. Wang, Y. Li, M. Qin, W. Wang, B. Chen and Y. Cao, *Sci. Adv.*, 2020, **6**, eaaz9531.
104 I. Tunn, A. S. De Léon, K. G. Blank and M. J. Harrington, *Nanoscale*, 2018, **10**, 22725–22729.
105 Y. Cao, X. Wei, Y. Lin and F. Sun, *Mol. Syst. Des. Eng.*, 2019, **5**, 117–124.
106 C. Mirabello and B. Wallner, *bioRxiv*, 2018, 394437.
107 C. H. Yu, Z. Qin, F. J. Martin-Martinez and M. J. Buehler, *ACS Nano*, 2019, **13**, 7471–7482.
108 S. Sjöberg, *Pure Appl. Chem.*, 1997, **69**, 1549–1570.
109 H. Irving and R. J. P. Williams, *J. Chem. Soc.*, 1953, 3192–3210.
110 M. Remko, D. Fitz and B. M. Rode, *Amino Acids*, 2010, **39**, 1309–1319.
111 R. M. Smith and A. E. Martell, *Critical Stability Constants*, 1976, vol. 6.
112 M. Gensler, C. Eidamshaus, A. Galstyan, E. W. Knapp, H. U. Reissig and J. P. Rabe, *J. Phys. Chem. C*, 2015, **119**, 4333–4343.
113 P. Li and K. M. Merz, *Chem. Rev.*, 2017, **117**, 1564–1686.
114 E. Duboué-Dijon, P. E. Mason, H. E. Fischer and P. Jungwirth, *J. Phys. Chem. B*, 2018, **122**, 3296–3306.
115 L. Rodríguez-Santiago, J. Alí-Torres, P. Vidossich and M. Sodupe, *Phys. Chem. Chem. Phys.*, 2015, **17**, 13582–13589.
116 A. Assifaoui, A. Lerbret, H. T. D. Uyen, F. Neiers, O. Chambin, C. Loupiac and F. Cousin, *Soft Matter*, 2015, **11**, 551–560.
117 S. Tetteh, *Bioinorg. Chem. Appl.*, 2018, **2018**, 1.
118 T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama and A. C. T. Van Duin, *npj Comput. Mater.*, 2016, 2, 15011.
119 D. Raymand, A. C. T. van Duin, M. Baudin and K. Hermansson, *Surf. Sci.*, 2008, **602**, 1020–1031.
120 D. Raymand, A. C. T. Van Duin, M. Baudin and K. Hermansson, 2008, **602**, 1020–1031.
121 A. C. T. Van Duin, V. S. Bryantsev, M. S. Diallo, W. A. Goddard, O. Rahaman, D. J. Doren, D. Raymand and K. Hermansson, *J. Phys. Chem. A*, 2010, **114**, 9507–9514.
122 L. Huang, K. L. Joshi, A. C. T. V. Duin, T. J. Bandosz and K. E. Gubbins, *Phys. Chem. Chem. Phys.*, 2012, **14**, 11327–11332.
123 L. Zeng, M. Song, J. Gu, Z. Xu, B. Xue, Y. Li and Y. Cao, *Biomimetics*, 2019, **4**, 1–13.
124 H. Shabbir and M. A. Hartmann, *New J. Phys.*, 2017, **19**, 093024.
125 H. Shabbir, C. Dellago and M. Hartmann, *Biomimetics*, 2019, **4**, 12.
126 L. Rulíšek and J. Vondrášek, *J. Inorg. Biochem.*, 1998, **71**, 115–127.
127 M. J. Sever and J. J. Wilker, *Dalt. Trans.*, 2006, 813–822.
128 Y. Li, J. Wen, M. Qin, Y. Cao, H. Ma and W. Wang, *ACS Biomater. Sci. Eng.*, 2017, **3**, 979–989.
129 Z. Xu, *Sci. Rep.*, 2013, **3**, 2914.
130 C. H. Langford and H. B. Gray, *Ligand substitution processes*, 1966.
131 C. P. Li and M. Du, *Chem. Commun.*, 2011, **47**, 5958–5972.
132 P. C. Redfern, P. Zapol, L. A. Curtiss, T. Rajh and M. C. Thurnauer, *J. Phys. Chem. B*, 2003, **107**, 11419–

11427.
133 L. Helm and A. E. Merbach, *Chem. Rev.*, 2005, 105, 1923–1959.
134 C. C. Buck, P. B. Dennis, M. K. Gupta, M. T. Grant, M. G. Crosby, J. M. Slocik, P. A. Mirau, K. A. Becknell, K. K. Comfort and R. R. Naik, *Macromol. Biosci.*, 2019, **19**, 1800238.
135 G. P. Maier, M. V Rapp, J. H. Waite, J. N. Israelachvili and A. Butler, *Science (80-.).*, 2015, **349**, 628–632.
136 Y. Li, Y. Li, T. Wang, L. Xia, L. Wang, M. Qin, W. Wang and Y. Cao, *J. Mater. Chem. B*, 2017, **5**, 4416–4420.
137 Y. Li, C. Liang, L. Gao, S. Li, Y. Zhang, J. Zhang and Y. Cao, *Mater. Chem. Front.*, 2017, **1**, 2664–2668.
138 M. J. Sever, J. T. Weisser, J. Monahan, S. Srinivasan and J. J. Wilker, *Angew. Chemie - Int. Ed.*, 2004, **43**, 448–450.
139 D. G. Barrett, D. E. Fullenkamp, L. He, N. Holten-Andersen, K. Y. C. Lee and P. B. Messersmith, *Adv. Funct. Mater.*, 2013, **23**, 1111–1119.
140 J. T. Auletta, G. J. Ledonne, K. C. Gronborg, C. D. Ladd, H. Liu, W. W. Clark and T. Y. Meyer, *Macromolecules*, 2015, **48**, 1736–1747.
141 C. Mou, F. Ali, A. Malaviya and C. J. Bettinger, *J. Mater. Chem. B*, 2019, **7**, 1690–1696.
142 S. V. Wegner, F. C. Schenk, S. Witzel, F. Bialas and J. P. Spatz, *Macromolecules*, 2016, **49**, 4229–4235.
143 S. Okamoto and L. D. Eltis, *Metallomics*, 2011, 3, 963–970.
144 F. W. Outten and E. C. Theil, *Antioxidants Redox Signal.*, 2009, 11, 1029–1046.
145 A. Campanella, D. Döhler and W. H. Binder, *Macromol. Rapid Commun.*, 2018, 39, 1700739.
146 E. Duboué-Dijon, P. E. Mason, H. E. Fischer and P. Jungwirth, *J. Chem. Phys.*, 2017, **146**, 185102.
147 S. Kim, A. M. Peterson and N. Holten-Andersen, *Chem. Mater.*, 2018, **30**, 3648–3655.
148 Q. Liu, X. Lu, L. Li, H. Zhang, G. Liu, H. Zhong and H. Zeng, *J. Phys. Chem. C*, 2016, **120**, 21670–21677.
149 J. Castillo-Tejas, O. Castrejón-González, S. Carro, V. González-Coronel, J. F. J. Alvarado and O. Manero, *Colloids Surfaces A Physicochem. Eng. Asp.*, 2016, **491**, 37–49.
150 J. Castillo-Tejas, S. Carro and O. Manero, *ACS Macro Lett.*, 2017, **6**, 190–193.
151 D. Liu, F. Liu, W. Zhou, F. Chen and J. Wei, *J. Mol. Liq.*, 2018, **268**, 131–139.
152 Q. Zhang, X. Zhu, C. H. Li, Y. Cai, X. Jia and Z. Bao, *Macromolecules*, 2019, **52**, 660–668.
153 H. Ozaki, T. Indei, T. Koga and T. Narita, *Polymer (Guildf).*, 2017, **128**, 363–372.
154 K. Sénéchal, O. Maury, H. Le Bozec, I. Ledoux and J. Zyss, *J. Am. Chem. Soc.*, 2002, **124**, 4560–4561.
155 H. Qin, T. Zhang, H. N. Li, H. P. Cong, M. Antonietti and S. H. Yu, *Chem*, 2017, **3**, 691–705.
156 M. S. Menyo, C. J. Hawker and J. H. Waite, *Soft Matter*, 2013, **9**, 10314–10323.
157 K. Zhan, C. Kim, K. Sung, H. Ejima and N. Yoshie, *Biomacromolecules*, 2017, **18**, 2959–2966.
158 D. P. Wang, J. C. Lai, H. Y. Lai, S. R. Mo, K. Y. Zeng, C. H. Li and J. L. Zuo, *Inorg. Chem.*, 2018, **57**, 3232–3242.
159 W. Schmolke, M. Ahmadi and S. Seiffert, *Phys. Chem. Chem. Phys.*, 2019, **21**, 19623–19638.
160 D. S. Hwang, H. Zeng, Q. Lu, J. Israelachvili and J. H. Waite, *Soft Matter*, 2012, **8**, 5640–5648.
161 N. Budisa and T. Schneider, *ChemBioChem*, 2019, 20, 2163–2190.
162 F. Liu, Y. Long, Q. Zhao, X. Liu, G. Qiu, L. Zhang, Q. Ling and H. Gu, *Polymer (Guildf).*, 2018, **143**, 212–227.
163 J. C. Lai, X. Y. Jia, D. P. Wang, Y. B. Deng, P. Zheng, C. H. Li, J. L. Zuo and Z. Bao, *Nat. Commun.*, 2019, **10**, 1164.
164 J. P. Park, I. T. Song, J. Lee, J. H. Ryu, Y. Lee and H. Lee, *Chem. Mater.*, 2015, **27**, 105–111.
165 Q. Li, D. G. Barrett, P. B. Messersmith and N. Holten-Andersen, *ACS Nano*, 2016, **10**, 1317–1324.
166 L. Du, Z. Y. Xu, C. J. Fan, G. Xiang, K. K. Yang and Y. Z. Wang, *Macromolecules*, 2018, **51**, 705–715.
167 J. H. Xu, S. Ye and J. J. Fu, *J. Mater. Chem. A*, 2018, **6**, 24291–24297.
168 Z. Xu, J. Li, G. Gao, Z. Wang, Y. Cong, J. Chen, J. Yin, L. Nie and J. Fu, *J. Polym. Sci. Part B Polym. Phys.*, 2018, **56**, 865–876.
169 E. Filippidi, T. R. Cristiani, C. D. Eisenbach, J. Herbert Waite, J. N. Israelachvili, B. Kollbe Ahn and M. T. Valentine, *Science (80-.).*, 2017, **358**, 502–505.
170 Y. Zhang, Z. Zeng, X. Y. D. Ma, C. Zhao, J. M. Ang, B. F. Ng, M. P. Wan, S. C. Wong, Z. Wang and X. Lu, *Appl. Surf. Sci.*, 2019, **479**, 700–708.
171 M. A. Bellinger, J. A. Sauer and M. Hara, *Macromolecules*, 1994, **27**, 6147–6155.
172 L. Shi, F. Wang, W. Zhu, Z. Xu, S. Fuchs, J. Hilborn, L. Zhu, Q. Ma, Y. Wang, X. Weng and D. A. Ossipov, *Adv. Funct. Mater.*, 2017, **27**, 1700591.
173 M. Burnworth, L. Tang, J. R. Kumpfer, A. J. Duncan, F. L. Beyer, G. L. Fiore, S. J. Rowan and C. Weder, *Nature*, 2011, **472**, 334–337.
174 Z. Zhang, Z. Gao, Y. Wang, L. Guo, C. Yin, X. Zhang, J. Hao, G. Zhang and L. Chen, *Macromolecules*, 2019, **52**, 2531–2541.
175 Y. L. Rao, A. Chortos, R. Pfattner, F. Lissel, Y. C. Chiu, V. Feig, J. Xu, T. Kurosawa, X. Gu, C. Wang, M. He, J. W. Chung and Z. Bao, *J. Am. Chem. Soc.*, 2016, **138**, 6020–6027.
176 Q. Zhu, L. Zhang, K. Van Vliet, A. Miserez and N. Holten-Andersen, *ACS Appl. Mater. Interfaces*, 2018, **10**,

10409–10418.
177 B. P. Lee and S. Konst, *Adv. Mater.*, 2014, **26**, 3415–3419.
178 B. P. Lee, M. H. Lin, A. Narkar, S. Konst and R. Wilharm, *Sensors Actuators, B Chem.*, 2015, **206**, 456–462.
179 L. Yang, G. Zhang, N. Zheng, Q. Zhao and T. Xie, *Angew. Chemie - Int. Ed.*, 2017, **56**, 12599–12602.
180 J. A. Foster, R. M. Parker, A. M. Belenguer, N. Kishi, S. Sutton, C. Abell and J. R. Nitschke, *J. Am. Chem. Soc.*, 2015, **137**, 9722–9729.
181 L. Li, W. Smitthipong and H. Zeng, *Polym. Chem.*, 2015, 6, 353–358.
182 N. Sharma, P. K. Sharma, Y. Singh and C. M. Nagaraja, *ACS Omega*, 2019, **4**, 1354–1363.
183 B. J. Kim, H. Cheong, B. H. Hwang and H. J. Cha, *Angew. Chemie - Int. Ed.*, 2015, **54**, 7318–7322.
184 Y. Jeong, Y. K. Jo, B. J. Kim, B. Yang, K. Il Joo and H. J. Cha, *ACS Nano*, 2018, **12**, 8909–8919.
185 E. R. L. Brisson, J. C. Griffith, A. Bhaskaran, G. V Franks and L. A. Connal, *J. Polym. Sci. Part A Polym. Chem.*, 2019, **57**, 1964–1973.
186 W. C. Huang, C. J. Bettinger, K. Rhee and C. J. Bettinger, *Biomacromolecules*, 2017, **18**, 1162–1171.
187 X. Yi, J. He, X. Wang, Y. Zhang, G. Tan, Z. Zhou, J. Chen, D. Chen, R. Wang, W. Tian, P. Yu, L. Zhou and C. Ning, *ACS Appl. Mater. Interfaces*, 2018, **10**, 6190–6198.
188 J. Yang, R. Bai, B. Chen and Z. Suo, *Adv. Funct. Mater.*, 2020, 30, 1901693.
189 P. Kord Forooshani and B. P. Lee, *J. Polym. Sci. Part A Polym. Chem.*, 2017, 55, 9–33.
190 Y. Li and Y. Cao, *Nanoscale Adv.*, 2019, **1**, 4246–4257.
191 D. X. Oh, S. Kim, D. Lee and D. S. Hwang, *Acta Biomater.*, 2015, **20**, 104–112.
192 E. Prajatelistia, S. W. Ju, N. D. Sanandiya, S. H. Jun, J. S. Ahn and D. S. Hwang, *Adv. Healthc. Mater.*, 2016, **5**, 919–927.
193 S. W. Ju, E. Prajatelistia, S. H. Jun, D. S. Hwang, J. S. Ahn and N. D. Sanandiya, *RSC Adv.*, 2017, **7**, 87–94.
194 B. K. Ahn, S. Das, R. Linstadt, Y. Kaufman, N. R. Martinez-Rodriguez, R. Mirshafian, E. Kesselman, Y. Talmon, B. H. Lipshutz, J. N. Israelachvili and J. H. Waite, *Nat. Commun.*, 2015, **6**, 8663.
195 M. A. Gebbie, W. Wei, A. M. Schrader, T. R. Cristiani, H. A. Dobbs, M. Idso, B. F. Chmelka, J. H. Waite and J. N. Israelachvili, *Nat. Chem.*, 2017, **9**, 473–479.
196 C. Dai, Z. Zhou, Z. Guan, Y. Wu, Y. Liu, J. He, P. Yu, L. Tu, F. Zhang, D. Chen, R. Wang, C. Ning, L. Zhou and G. Tan, *Macromol. Mater. Eng.*, 2018, **303**, 1800305.
197 H. Zeng, D. S. Hwang, J. N. Israelachvili and J. H. Waite, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 12850–12853.
198 B. Yang, C. Lim, D. S. Hwang and H. J. Cha, *Chem. Mater.*, 2016, **28**, 7982–7989.
199 Y. Chan Choi, J. S. Choi, Y. J. Jung and Y. W. Cho, *J. Mater. Chem. B*, 2014, **2**, 201–209.
200 H. Kang, K. Zhang, H. J. Jung, B. Yang, X. Chen, Q. Pan, R. Li, X. Xu, G. Li, V. P. Dravid and L. Bian, *Adv. Mater.*, 2018, **30**, 1803591.
201 E. Lai, B. Keshavarz and N. Holten-Andersen, *Langmuir*, 2019, **35**, 15979–15984.
202 N. S. Schauser, G. E. Sanoja, J. M. Bartels, S. K. Jain, J. G. Hu, S.-I. Han, L. M. Walker, M. E. Helgeson, R. Seshadri and R. A. Segalman, *Chem. Mater.*, 2018, **30**, 5759–5769.
203 Y. Shi, M. Wang, C. Ma, Y. Wang, X. Li and G. Yu, *Nano Lett.*, 2015, **15**, 6276–6281.
204 H. Gao, Y. Sun, J. Zhou, R. Xu and H. Duan, *ACS Appl. Mater. Interfaces*, 2013, **5**, 425–432.
205 L. Yan, M. Lv, C. Su, L. Zheng, J. Li and Z. Ye, *Soft Matter*, 2017, **13**, 8772–8780.
206 Q. Lin, Q. P. Yang, B. Sun, Y. P. Fu, X. Zhu, T. B. Wei and Y. M. Zhang, *Mater. Lett.*, 2014, **137**, 444–446.
207 Q. Lin, B. Sun, Q. P. Yang, Y. P. Fu, X. Zhu, Y. M. Zhang and T. B. Wei, *Chem. Commun.*, 2014, **50**, 10669–10671.
208 Q. Zhu, K. Van Vliet, N. Holten-Andersen and A. Miserez, *Adv. Funct. Mater.*, 2019, **29**, 1808191.
209 Y. Ma, P. She, K. Y. Zhang, H. Yang, Y. Qin, Z. Xu, S. Liu, Q. Zhao and W. Huang, *Nat. Commun.*, 2018, **9**, 3.
210 S. Bhowmik, B. N. Ghosh, V. Marjomäki and K. Rissanen, *J. Am. Chem. Soc.*, 2014, **136**, 5543–5546.
211 H. Ejima, J. J. Richardson, K. Liang, J. P. Best, M. P. Van Koeverden, G. K. Such, J. Cui and F. Caruso, *Science (80-.).*, 2013, **341**, 154–157.
212 G. Yun, J. J. Richardson, M. Biviano and F. Caruso, *ACS Appl. Mater. Interfaces*, 2019, **11**, 6404–6410.
213 M. Björnmalm, J. Cui, N. Bertleff-Zieschang, D. Song, M. Faria, M. A. Rahim and F. Caruso, *Chem. Mater.*, 2017, **29**, 289–306.
214 Y. Vidavsky, M. R. Buche, Z. M. Sparrow, X. Zhang, S. J. Yang, R. A. Distasio and M. N. Silberstein, *Macromolecules*, 2020, **53**, 2021–2030.
215 Q. Daniel, P. Huang, T. Fan, Y. Wang, L. Duan, L. Wang, F. Li, Z. Rinkevicius, F. Mamedov, M. S. G. Ahlquist, S. Styring and L. Sun, *Coord. Chem. Rev.*, 2017, **346**, 206–215.
216 E. Khare, S. A. Cazzell, J. Song, N. Holten-Andersen and M. J. Buehler, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2213160120.
217 W. C. Yount, D. M. Loveless and S. L. Craig, *Angew. Chemie - Int. Ed.*, 2005, **44**, 2746–2748.

218 A. Barducci, G. Bussi and M. Parrinello, *Phys. Rev. Lett.*, 2008, **100**, 020603.
219 M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia and M. Parrinello, *Comput. Phys. Commun.*, 2009, **180**, 1961–1972.
220 P. Tiwary and M. Parrinello, *Phys. Rev. Lett.*, 2013, **111**, 230602.
221 S. A. Cazzell, Massachusetts Institute of Technology, 2020.
222 T. Rossow, A. Habicht and S. Seiffert, *Macromolecules*, 2014, **47**, 6473–6482.
223 D. U. Ferreiro, E. A. Komives and P. G. Wolynes, *Q. Rev. Biophys.*, 2014, 47, 285–363.
224 P. Zhang, J. Han, P. Cieplak and M. S. Cheung, *J. Chem. Phys.*, 2021, **154**, 124104.
225 J. Kahlen, L. Salimi, M. Sulpizi, C. Peter and D. Donadio, *J. Phys. Chem. B*, 2014, **118**, 3960–3972.
226 P. Hänggi, P. Talkner and M. Borkovec, *Rev. Mod. Phys.*, 1990, **62**, 251–341.
227 S. Tang, A. Habicht, S. Li, S. Seiffert and B. D. Olsen, *Macromolecules*, 2016, **49**, 5599–5608.
228 S. Plimpton, *J. Comput. Phys.*, 1995, **117**, 1–19.
229 A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, 1998, **102**, 3586–3616.
230 A. D. Mackerell, M. Feig and C. L. Brooks, *J. Comput. Chem.*, 2004, **25**, 1400–1415.
231 C. S. Babu and C. Lim, *J. Phys. Chem. A*, 2006, **110**, 691–699.
232 H. Lee, R. M. Venable, A. D. MacKerell and R. W. Pastor, *Biophys. J.*, 2008, **95**, 1590–1599.
233 R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*, 1988.
234 H. C. Andersen, *J. Comput. Phys.*, 1983, **52**, 24–34.
235 G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni and G. Bussi, *Comput. Phys. Commun.*, 2014, **185**, 604–613.
236 K. W. Plaxco and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 13591–13596.
237 N. Hori, N. A. Denesyuk and D. Thirumalai, *J. Phys. Chem. B*, 2018, **122**, 11279–11288.
238 G. Arya, *Mol. Simul.*, 2016, **42**, 1102–1115.
239 N. Berglund, *Polymath*, 2013, **19**, 459–490.
240 M. Chupeau, J. Gladrow, A. Chepelianskii, U. F. Keyser and E. Trizac, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 1383–1388.
241 S. Adhikari and K. S. D. Beach, *Phys. Rev. Res.*, 2020, **2**, 23276.
242 M. Arrayás, I. K. H. Kaufman, D. G. Luchinsky, P. V. E. Mc Clintock and S. M. Soskin, *Phys. Rev. Lett.*, 2000, **84**, 2556–2559.
243 S. Guo, A. K. Efremov and J. Yan, *Commun. Chem.*, 2019, **2**, 30.
244 V. Barsegov and D. Thirumalai, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 1835–1839.
245 K. L. Sebastian and A. K. R. Paul, *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, 2000, **62**, 927–939.
246 O. V. Gendelman and G. Karmi, *Nonlinear Dyn.*, 2019, **98**, 2775–2792.
247 X. Zhang, Y. Vidavsky, S. Aharonovich, S. J. Yang, M. R. Buche, C. E. Diesendruck and M. N. Silberstein, *Soft Matter*, 2020, **16**, 8591–8601.
248 J. Zhao, T. Narita and C. Creton, in *Advances in Polymer Science*, 2020, vol. 285, pp. 1–20.
249 G. M. Ross, I. L. Shamovsky, S. B. Woo, J. I. Post, P. N. Vrkljan, G. Lawrance, M. Solc, S. M. Dostaler, K. E. Neet and R. J. Riopelle, *J. Neurochem.*, 2001, **78**, 515–523.
250 R. Ge, Y. Zhang, X. Sun, R. M. Watt, Q. Y. He, J. D. Huang, D. E. Wilcox and H. Sun, *J. Am. Chem. Soc.*, 2006, **128**, 11330–11331.
251 J. Song, E. Khare, L. Rao, M. J. Buehler and N. Holten-Andersen, .
252 C. F. Quinn, M. C. Carpenter, M. L. Croteau and D. E. Wilcox, *Isothermal Titration Calorimetry Measurements of Metal Ions Binding to Proteins*, Elsevier Inc., 1st edn., 2016, vol. 567.
253 C. M. H. Ferreira, I. S. S. Pinto, E. V. Soares and H. M. V. M. Soares, *RSC Adv.*, 2015, 5, 30989–31003.
254 A. Tripathi, K. C. Tam and G. H. McKinley, *Macromolecules*, 2006, **39**, 1981–1999.
255 F. Tanaka and S. F. Edwards, *J. Nonnewton. Fluid Mech.*, 1992, **43**, 247–271.
256 F. Tanaka and S. F. Edwards, *J. Nonnewton. Fluid Mech.*, 1992, **43**, 289–309.
257 M. Rubinstein and R. H. Colby, *Polymers physics*, Oxford University Press New York, 2003, vol. 23.
258 L. M. Franklin, S. M. Walker and G. Hill, *J. Mol. Model.*, 2020, **26**, 116.
259 D. Vaslensin, Ł. Szyrwiel, F. Camponeschi, M. Rowińska-Zyrek, E. Molteni, E. Jankowska, A. Szymanska, E. Gaggelli, G. Valensin and H. Kozłowski, *Inorg. Chem.*, 2009, **48**, 7330–7340.
260 J. H. Viles, F. E. Cohen, S. B. Prusiner, D. B. Goodin, P. E. Wright and H. J. Dyson, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 2042–2047.
261 P. Deschamps, P. P. Kulkarni, M. Gautam-Basak and B. Sarkar, *Coord. Chem. Rev.*, 2005, **249**, 895–909.
262 P. Ferrer, F. Jiménez-Villacorta, J. Rubio-Zuazo, I. Da Silva and G. R. Castro, *J. Phys. Chem. B*, 2014, **118**,

2842–2850.
263 L. Zhou, S. Li, Y. Su, X. Yi, A. Zheng and F. Deng, *J. Phys. Chem. B*, 2013, **117**, 8954–8965.
264 A. Gergely and I. Sóvágó, *Inorganica Chim. Acta*, 1976, **20**, 19–25.
265 M. Itabashi, K. Shoji and K. Itoh, *Inorg. Chem.*, 1982, **21**, 3484–3489.
266 Y. Altun and F. Köseoĝlu, *J. Solution Chem.*, 2005, **34**, 213–231.
267 T. Vander Hoogerstraete, N. R. Brooks, B. Norberg, J. Wouters, K. Van Hecke, L. Van Meervelt and K. Binnemans, *CrystEngComm*, 2012, **14**, 4902–4911.
268 T. Otieno, M. J. Hatfield, S. L. Asher, A. I. McMullin, B. O. Patrick and S. Parkin, *Synth. React. Inorg. Met. Chem.*, 2001, **31**, 1587–1598.
269 J. E. Bauman and J. C. Wang, *Imidazole Complexes of Nickel(II), Copper(II), Zinc(II), and Silver(I)*, 1963.
270 T. Dudev and C. Lim, *J. Am. Chem. Soc.*, 2000, **122**, 11146–11153.
271 R. R. Conry, in *Encyclopedia of Inorganic and Bioinorganic Chemistry*, John Wiley & Sons, Ltd, Chichester, UK, 2011.
272 R. R. Roe and Y.-P. Pang, *J. Mol. Model.*, 1999, **5**, 134–140.
273 H. Keppler, in *Encyclopedia of Geochemistry*, Springer, 19th edn., 1998, pp. 118–120.
274 M. Maeder and Y.-M. Neuhold, *Practical Data Analysis in Chemistry*, Elsevier, 2007.
275 J. Da Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.
276 A. V Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.
277 A. C. Balazs, *Mater. Today*, 2007, 10, 18–23.
278 B. V. S. Iyer, V. V. Yashin, M. J. Hamer, T. Kowalewski, K. Matyjaszewski and A. C. Balazs, *Prog. Polym. Sci.*, 2015, **40**, 121–137.
279 G. V Kolmakov, K. Matyjaszewski and A. C. Balazs, *ACS Nano*, 2009, **3**, 885–892.
280 H. Lee, A. H. De Vries, S. J. Marrink and R. W. Pastor, *J. Phys. Chem. B*, 2009, **113**, 13186–13194.
281 M. Zhong, R. Wang, K. Kawamoto, B. D. Olsen and J. A. Johnson, *Science (80-.).*, 2016, **353**, 1264–1268.
282 F. Lange, K. Schwenke, M. Kurakazu, Y. Akagi, U. Il Chung, M. Lang, J. U. Sommer, T. Sakai and K. Saalwächter, *Macromolecules*, 2011, **44**, 9666–9674.
283 M. H. Chen, L. L. Wang, J. J. Chung, Y. H. Kim, P. Atluri and J. A. Burdick, *ACS Biomater. Sci. Eng.*, 2017, **3**, 3146–3160.
284 S. A. Cazzell, B. Duncan, R. Kingsborough and N. Holten-Andersen, *Adv. Funct. Mater.*, 2021, **31**, 2009118.
285 H. Fu, B. Wang, J. Li, J. Xu, J. Li, J. Zeng, W. Gao and K. Chen, *Mater. Horizons*, 2022, **9**, 1412–1421.
286 X. Fan, Y. Fang, W. Zhou, L. Yan, Y. Xu, H. Zhu and H. Liu, *Mater. Horizons*, 2021, **8**, 997–1007.
287 C. Andreini, G. Cavallaro, S. Lorenzini and A. Rosato, *Nucleic Acids Res.*, 2013, **41**, D312–D319.
288 J. Watly, E. Simonovsky, N. Barbosa, M. Spodzieja, R. Wieczorek, S. Rodziewicz-Motowidlo, Y. Miller and H. Kozlowski, *Inorg. Chem.*, 2015, **54**, 7692–7702.
289 J. Wątły, A. Hecel, M. Rowińska-Żyrek and H. Kozłowski, *Inorganica Chim. Acta*, 2018, **472**, 119–126.
290 D. Brasili, J. Watly, E. Simonovsky, R. Guerrini, N. A. Barbosa, R. Wieczorek, M. Remelli, H. Kozlowski and Y. Miller, *Dalt. Trans.*, 2016, **45**, 5629–5639.
291 J. Qian and C. Berkland, *iScience*, 2019, **21**, 124–134.
292 M. Mammen, S.-K. Choi and G. M. Whitesides, *Angew. Chemie - Int. Ed.*, 1998, 2754–2794.
293 E. Zumbro, J. Witten and A. Alexander-Katz, *Biophys. J.*, 2019, **117**, 892–902.
294 M. Hebel, A. Riegger, M. M. Zegota, G. Kizilsavas, J. Gačanin, M. Pieszka, T. Lückerath, J. A. S. Coelho, M. Wagner, P. M. P. Gois, D. Y. W. Ng and T. Weil, *J. Am. Chem. Soc.*, 2019, **141**, 14026–14031.
295 I. Mahmad Rasid, C. Do, N. Holten-Andersen and B. D. Olsen, *Soft Matter*, 2021, **17**, 8960–8972.
296 Z. Jiang, A. Bhaskaran, H. M. Aitken, I. C. G. Shackleford and L. A. Connal, *Macromol. Rapid Commun.*, 2019, **40**, 1900038.
297 K. Breul, S. Kissel and S. Seiffert, *Macromolecules*, 2021, **54**, 8407–8422.
298 S. Knecht, D. Ricklin, A. N. Eberle and B. Ernst, *J. Mol. Recognit.*, 2009, **22**, 270–279.
299 J. Nomata, M. Kitashima, K. Inoue and Y. Fujita, *FEBS Lett.*, 2006, **580**, 6151–6154.
300 S. Keten and M. J. Buehler, in *Nano Letters*, American Chemical Society, 2008, vol. 1061, pp. 743–748.
301 A. A. Griffith, *Philos. Trans. R. Soc. london. Ser. A, Contain. Pap. a Math. or Phys. character*, 1921, **221**, 163–198.
302 M. J. Buehler and H. Gao, *Nature*, 2006, **439**, 307–310.
303 C. Chen, Z. Wang and Z. Suo, *Extrem. Mech. Lett.*, 2017, **10**, 50–57.
304 W. Ott, M. A. Jobst, M. S. Bauer, E. Durner, L. F. Milles, M. A. Nash and H. E. Gaub, *ACS Nano*, 2017, **11**, 6346–6354.
305 G. Stirnemann, D. Giganti, J. M. Fernandez and B. J. Berne, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 3847–3852.
306 S. Fluegel, K. Fischer, J. R. McDaniel, A. Chilkoti and M. Schmidt, *Biomacromolecules*, 2010, **11**, 3216–3218.

307 Z. Qin and M. J. Buehler, *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 2010, **82**, 061906.
308 R. Rohs, C. Etchebest and R. Lavery, *Biophys. J.*, 1999, **76**, 2760–2768.
309 K. Fujiwara, H. Toda and M. Ikeguchi, *BMC Struct. Biol.*, 2012, **12**, 18.
310 C. K. Smith, L. Regan and J. M. Withka, *Biochemistry*, 1994, **33**, 5510–5517.
311 N. Bhattacharjee and P. Biswas, *BMC Struct. Biol.*, 2010, **10**, 29.
312 H. Zheng, M. Chruszcz, P. Lasota, L. Lebioda and W. Minor, *J. Inorg. Biochem.*, 2008, **102**, 1765–1776.
313 S. Barber-Zucker, B. Shaanan and R. Zarivach, *Sci. Rep.*, 2017, **7**, 16381.
314 J. Crowe, B. S. Masone and J. Ribbe, *Methods Mol. Biol.*, 1996, **58**, 491–510.
315 F. Kienberger, G. Kada, H. J. Gruber, V. P. Pastushenko, C. Riener, M. Trieb, H. G. Knaus, H. Schindler and P. Hinterdorfer, *Single Mol.*, 2000, **1**, 59–65.
316 R. J. Young and P. A. Lovell, *Introduction to Polymers*, Taylor & Francis Group, Boca Raton, 3rd edn., 2011.
317 D. D. Root, V. K. Yadavalli, J. G. Forbes and K. Wang, *Biophys. J.*, 2006, **90**, 2852–2866.
318 J. Van Noort, T. Van der Heijden, M. De Jager, C. Wyman, R. Kanaar and C. Dekker, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 7581–7586.
319 M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier and C. Bustamante, *Science (80-.).*, 1997, **276**, 1112–1116.
320 S. Choe and S. X. Sun, *J. Chem. Phys.*, 2005, **122**, 244912.
321 C. C. Vandenakker, M. F. M. Engel, K. P. Velikov, M. Bonn and G. H. Koenderink, *J. Am. Chem. Soc.*, 2011, **133**, 18030–18033.
322 Y. R. Luo, *Comprehensive handbook of chemical bond energies*, CRC Press, 2007.
323 L. F. Zou, Y. Fu, K. Shen and Q. X. Guo, *J. Mol. Struct. THEOCHEM*, 2007, **807**, 87–92.
324 F. H. Seubold, *J. Chem. Phys.*, 1954, *22*, 945–946.
325 S. Keten, MIT, 2010.
326 T. Ackbarow, X. Chen, S. Keten and M. J. Buehler, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 16410–16415.
327 R. Wieczorek and J. J. Dannenberg, *J. Am. Chem. Soc.*, 2003, **125**, 8124–8129.
328 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.
329 R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper and D. Hassabis, *bioRxiv*, 2021, 2021.10.04.463034.
330 S. C. Grindy and N. Holten-Andersen, *Soft Matter*, 2017, **13**, 4057–4065.
331 L. E. Valenti, C. P. De Pauli and C. E. Giacomelli, *J. Inorg. Biochem.*, 2006, **100**, 192–200.
332 E. Zumbro and A. Alexander-Katz, *ACS Omega*, 2020, **5**, 10774–10781.
333 E. Khare, D. S. Grewal and M. J. Buehler, *Nanoscale*, 2023.
334 S. C. Shen, E. Khare, N. A. Lee, M. K. Saad, D. L. Kaplan and M. J. Buehler, *Chem. Rev.*, 2022, 123, 2242–2275.
335 G. I. Bell, *Science (80-.).*, 1978, 200, 618–627.
336 E. Evans and K. Ritchie, *Biophys. J.*, 1997, **72**, 1541–1555.
337 S. Sumiyoshi, K. Suyama, N. Tanaka, T. Andoh, A. Nagata, K. Tomohara, S. Taniguchi, I. Maeda and T. Nose, *Sci. Rep.*, 2022, **12**, 19414.
338 Y. Wei, A. A. Thyparambil and R. A. Latour, *Biochim. Biophys. Acta - Proteins Proteomics*, 2014, **1844**, 2331–2337.
339 C. A. Brautigam, *Methods*, 2015, **76**, 124–136.
340 *Microcal iTC200 System User Manual*, Malvern, Worcestershire, 2014.
341 J. Watly, E. Simonovsky, R. Wieczorek, N. Barbosa, Y. Miller and H. Kozlowski, *Inorg. Chem.*, 2014, **53**, 6675–6683.
342 Y. B. Zeng, D. M. Zhang, H. Li and H. Sun, *J. Biol. Inorg. Chem.*, 2008, **13**, 1121–1131.
343 H. Li and P. Zheng, *Curr. Opin. Chem. Biol.*, 2018, 43, 58–67.
344 M. Conti, G. Falini and B. Samorì, *Angew. Chemie - Int. Ed.*, 2000, **39**, 215–218.
345 C. Verbelen, H. J. Gruber and Y. F. Dufrêne, *J. Mol. Recognit.*, 2007, **20**, 490–494.
346 L. Schmitt, M. Ludwig, H. E. Gaub and R. Tampé, *Biophys. J.*, 2000, **78**, 3275–3285.
347 A. Dittmore, D. B. McIntosh, S. Halliday and O. A. Saleh, *Phys. Rev. Lett.*, 2011, **107**, 148301.
348 F. Oesterhelt, M. Rief and H. E. Gaub, *New J. Phys.*, 1999, **1**, 6–6.
349 R. W. Friddle, A. Noy and J. J. De Yoreo, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 13573–13578.
350 P. López-García, A. D. de Araujo, A. E. Bergues-Pupo, I. Tunn, D. P. Fairlie and K. G. Blank, *Angew. Chemie Int. Ed.*, 2020, **60**, 232–236.
351 M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez and H. E. Gaub, *Science (80-.).*, 1997, **276**, 1109–1112.

352 A. Pohl, F. Berger, R. M. A. Sullan, C. Valverde-Tercedor, K. Freindl, N. Spiridis, C. T. Lefèvre, N. Menguy, S. Klumpp, K. G. Blank and D. Faivre, *Nano Lett.*, 2019, **19**, 8207–8215.
353 P. Zheng, S. J. Takayama, A. G. Mauk and H. Li, *J. Am. Chem. Soc.*, 2013, **135**, 7992–8000.
354 M. S. Menyo, C. J. Hawker and J. H. Waite, *ACS Macro Lett.*, 2015, **4**, 1200–1204.
355 S. Roberts, M. Dzuricky and A. Chilkoti, *FEBS Lett.*, 2015, **589**, 2477–2486.
356 D. J. Callahan, W. Liu, X. Li, M. R. Dreher, W. Hassouneh, M. Kim, P. Marszalek and A. Chilkoti, *Nano Lett.*, 2012, **12**, 2165–2170.
357 Z. Martín-Moldes, Q. Spey, T. Bhatacharya and D. L. Kaplan, *Macromol. Biosci.*, 2022, **22**, 2200122.
358 Q. Wang, X. Xia, W. Huang, Y. Lin, Q. Xu and D. L. Kaplan, *Adv. Funct. Mater.*, 2014, **24**, 4303–4310.
359 J. L. Zimmermann, T. Nicolaus, G. Neuert and K. G. Blank, *Nat. Protoc.*, 2010, **5**, 975–985.
360 J. L. Hutter and J. Bechhoefer, *Rev. Sci. Instrum.*, 1993, **64**, 1868–1873.
361 J. Rodriguez-Ramos and F. Rico, *Rev. Sci. Instrum.*, 2021, **92**, 045001.
362 D. J. Price and C. L. Brooks, *J. Chem. Phys.*, 2004, **121**, 10096–10103.
363 B. W. Cribb, C. L. Lin, L. Rintoul, R. Rasch, J. Hasenpusch and H. Huang, *Acta Biomater.*, 2010, **6**, 3152–3156.
364 H. Quan, D. Kisailus and M. A. Meyers, *Nat. Rev. Mater.*, 2021, 6, 264–283.
365 S. Bekele, K. Singh, E. Helton, S. Farajollahi, R. R. Naik, P. Dennis, N. Kelley-Loughnane and R. Berry, *J. Phys. Chem. B*, 2022, **126**, 6614–6623.
366 L. Banci, I. Bertini, F. Cramaro, R. Del Conte and M. S. Viezzoli, *Biochemistry*, 2003, **42**, 9543–9553.
367 W. Li, J. Wang, J. Zhang and W. Wang, *Curr. Opin. Struct. Biol.*, 2015, 30, 25–31.
368 K. Pagel, T. Seri, H. von Berlepsch, J. Griebel, R. Kirmse, C. Böttcher and B. Koksch, *ChemBioChem*, 2008, **9**, 531–536.
369 W. T. Chen, Y. H. Liao, H. M. Yu, I. H. Cheng and Y. R. Chen, *J. Biol. Chem.*, 2011, **286**, 9646–9656.
370 K. Mahnam, B. Saffar, M. Mobini-Dehkordi, A. Fassihi and A. Mohammadi, *Res. Pharm. Sci.*, 2014, **9**, 69–82.
371 F. Jehle, E. Macías-Sánchez, P. Fratzl, L. Bertinetti and M. J. Harrington, *Nat. Commun.*, 2020, **11**, 862.
372 G. W. Bryan and P. E. Gibbs, *J. Mar. Biol. Assoc. United Kingdom*, 1979, **59**, 969–973.
373 W. H. Roos, M. M. Gibbons, A. Arkhipov, C. Uetrecht, N. R. Watts, P. T. Wingfield, A. C. Steven, A. J. R. Heck, K. Schulten, W. S. Klug and G. J. L. Wuite, *Biophys. J.*, 2010, **99**, 1175–1181.
374 A. N. Masir, A. Darvizeh and A. Zajkani, *Adv. Compos. Lett.*, 2019, **28**, 0963693519860162.
375 O. Kononova, K. A. Marx and V. Barsegov, in *Applied Nanoindentation in Advanced Materials*, John Wiley & Sons, Ltd, Chichester, UK, 2017, pp. 393–428.
376 Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141–151.
377 J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, *J. Comput. Chem.*, 2005, 26, 1781–1802.
378 M. Feig, J. Karanicolas and C. L. Brooks, in *Journal of Molecular Graphics and Modelling*, Elsevier, 2004, vol. 22, pp. 377–395.
379 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graph.*, 1996, **14**, 33–38.
380 D. Frishman and P. Argos, *Proteins Struct. Funct. Bioinforma.*, 1995, **23**, 566–579.
381 R. Benjamin, F. Z. Carl and M. Ashlie, 2010.
382 Z. Qin, L. Wu, H. Sun, S. Huo, T. Ma, E. Lim, P. Y. Chen, B. Marelli and M. J. Buehler, *Extrem. Mech. Lett.*, 2020, **36**, 100652.
383 A. Sorushanova, L. M. Delgado, Z. Wu, N. Shologu, A. Kshirsagar, R. Raghunath, A. M. Mullen, Y. Bayon, A. Pandit, M. Raghunath and D. I. Zeugolis, *Adv. Mater.*, 2019, 31, 1801651.
384 H. Lodish, A. Berk and S. Zipursky, in *Molecular Cell Biology*, W. H. Freeman, 2000, p. Section 22.3.
385 J. P. R. O. Orgel, T. C. Irving, A. Miller and T. J. Wess, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 9001–9005.
386 G. K. G.N. Ramachandran, *Nature*, 1955, 593–595.
387 F. C. A. Rich, *Nature*, 1955, 915–916.
388 M. J. Buehler and S. Y. Wong, *Biophys. J.*, 2007, **93**, 37–43.
389 A. Bhattacharjee and M. Bansal, *IUBMB Life*, 2005, 57, 161–172.
390 R. Puxkandl, I. Zizak, O. Paris, J. Keckes, W. Tesch, S. Bernstorff, P. Purslow and P. Fratzl, *Philos. Trans. R. Soc. B Biol. Sci.*, 2002, **357**, 191–197.
391 D. J. Prockop and K. I. Kivirikko, *Annu. Rev. Biochem.*, 1995, 64, 403–434.
392 R. K. Nalla, J. J. Kruzic, J. H. Kinney and R. O. Ritchie, *Biomaterials*, 2005, **26**, 217–231.
393 R. O. Ritchie, J. J. Kruzic, C. L. Muhlstein, R. K. Nalla and E. A. Stach, *Int. J. Fract.*, 2004, **128**, 1–15.
394 M. J. Buehler, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 12285–12290.
395 A. Gautieri, S. Vesentini, A. Redaelli and M. J. Buehler, *Matrix Biol.*, 2012, **31**, 141–149.
396 Z. L. Shen, M. R. Dodge, H. Kahn, R. Ballarini and S. J. Eppell, *Biophys. J.*, 2008, **95**, 3956–3963.
397 J. A. J. Van Der Rijt, K. O. Van Der Werf, M. L. Bennink, P. J. Dijkstra and J. Feijen, *Macromol. Biosci.*, 2006, **6**, 697–702.

398 L. Yang, K. O. Van Der Werf, C. F. C. Fitié, M. L. Bennink, P. J. Dijkstra and J. Feijen, *Biophys. J.*, 2008, **94**, 2204–2211.
399 R. B. Svensson, T. Hassenkam, C. A. Grant and S. P. Magnusson, *Biophys. J.*, 2010, **99**, 4020–4027.
400 C. H. Lee, A. Singla and Y. Lee, *Int. J. Pharm.*, 2001, **221**, 1–22.
401 R. Parenteau-Bareil, R. Gauvin and F. Berthod, *Materials (Basel).*, 2010, **3**, 1863–1887.
402 M. Milazzo, N. Contessi Negrini, S. Scialla, B. Marelli, S. Farè, S. Danti and M. J. Buehler, *Adv. Funct. Mater.*, 2019, 29, 1903055.
403 A. V. Persikov, J. A. M. Ramshaw, A. Kirkpatrick and B. Brodsky, *Biochemistry*, 2005, **44**, 1414–1422.
404 T. V. Burjanadze, *Biopolymers*, 1979, **18**, 931–938.
405 K. Gekko and S. Koga, *J. Biochem.*, 1983, **94**, 199–205.
406 B. J. Rigby, *Nature*, 1968, 219, 166–167.
407 A. Mohs, T. Silva, T. Yoshida, R. Amin, S. Lukomski, M. Inouye and B. Brodsky, *J. Biol. Chem.*, 2007, **282**, 29757–29765.
408 K. Inouye, Y. Kobayashi, Y. Kyogoku, Y. Kishida, S. Sakakibara and D. J. Prockop, *Arch. Biochem. Biophys.*, 1982, **219**, 198–203.
409 S. Sakakibara, K. Inouye, K. Shudo, Y. Kishida, Y. Kobayashi and D. J. Prockop, *BBA - Protein Struct.*, 1973, **303**, 198–202.
410 M. J. Buehler, *J. Mater. Res.*, 2006, **21**, 1947–1961.
411 A. V. Persikov, J. A. M. Ramshaw, A. Kirkpatrick and B. Brodsky, *Biochemistry*, 2000, **39**, 14960–14967.
412 A. V. Persikov, J. A. M. Ramshaw, A. Kirkpatrick and B. Brodsky, *J. Mol. Biol.*, 2002, **316**, 385–394.
413 A. V. Persikov, J. A. M. Ramshaw and B. Brodsky, *J. Biol. Chem.*, 2005, **280**, 19343–19349.
414 J. Wang, H. Cao, J. Z. H. Zhang and Y. Qi, *Sci. Rep.*, 2018, **8**, 6349.
415 A. Al-Shahib, R. Breitling and D. R. Gilbert, *BMC Genomics*, 2007, **8**, 78.
416 C. H. Yu, E. Khare, O. P. Narayan, R. Parker, D. L. Kaplan and M. J. Buehler, *J. Mech. Behav. Biomed. Mater.*, 2022, **125**, 104921.
417 E. S. Hwang, G. Thiagarajan, A. S. Parmar and B. Brodsky, *Protein Sci.*, 2010, **19**, 1053–1064.
418 D. Whitley, *Stat. Comput.*, 1994, **4**, 65–85.
419 I. Goldberga, R. Li and M. J. Duer, *Acc. Chem. Res.*, 2018, **51**, 1621–1629.
420 C. L. Jenkins and R. T. Raines, *Nat. Prod. Rep.*, 2002, **19**, 49–59.
421 C. L. Jenkins, L. E. Bretscher, I. A. Guzei and R. T. Raines, *J. Am. Chem. Soc.*, 2003, **125**, 6422–6427.
422 K. Kar, S. Ibrar, V. Nanda, T. M. Getz, S. P. Kunapuli and B. Brodsky, *Biochemistry*, 2009, **48**, 7959–7968.
423 M. V. Katti, R. Sami-Subbu, P. K. Ranjekar and V. S. Gupta, *Protein Sci.*, 2000, **9**, 1203–1209.
424 Z. Yu, B. An, J. A. M. Ramshaw and B. Brodsky, *J. Struct. Biol.*, 2014, **186**, 451–461.
425 K. T. Walker, R. Nan, D. W. Wright, J. Gor, A. C. Bishop, G. I. Makhatadze, B. Brodsky and S. J. Perkins, *Biochem. J.*, 2017, **474**, 2203–2217.
426 M. Sun, S. Chen, S. M. Adams, J. B. Florer, H. Liu, W. W. Y. Kao, R. J. Wenstrup and D. E. Birk, *J. Cell Sci.*, 2011, **124**, 4096–4105.
427 M. D. Shoulders, J. A. Hodges and R. T. Raines, *J. Am. Chem. Soc.*, 2006, **128**, 8112–8113.
428 M. D. Shoulders and R. T. Raines, *Annu. Rev. Biochem.*, 2009, 78, 929–958.
429 Y. Qiu, A. Mekkat, H. Yu, S. Yigit, S. Hamaia, R. W. Farndale, D. L. Kaplan, Y. S. Lin and B. Brodsky, *J. Struct. Biol.*, 2018, **203**, 255–262.
430 A. V. Persikov, J. A. M. Ramshaw and B. Brodsky, *Biopolym. - Pept. Sci. Sect.*, 2000, **55**, 436–450.
431 A. V. Persikov, Y. Xu and B. Brodsky, *Protein Sci.*, 2004, **13**, 893–902.
432 S. D. Bolboacă and J. L, *Bull. USAMV-CN*, 2007, **64**, 311–316.
433 L. E. Bretscher, C. L. Jenkins, K. M. Taylor, M. L. DeRider and R. T. Raines, *J. Am. Chem. Soc.*, 2001, **123**, 777–778.
434 B. Brodsky, G. Thiagarajan, B. Madhan and K. Kar, *Biopolymers*, 2008, **89**, 345–353.
435 B. Brodsky and A. V Persikov, *Adv. Protein Chem.*, 2005, **70**, 301–339.
436 J. A. Fallas, J. Dong, Y. J. Tao and J. D. Hartgerink, *J. Biol. Chem.*, 2012, **287**, 8039–8047.
437 A. L. Fidler, S. P. Boudko, A. Rokas and B. G. Hudson, *J. Cell Sci.*, 2018, **131**, 203950.
438 H. -P Germann and E. Heidemann, *Biopolymers*, 1988, **27**, 157–163.
439 K. Beck, V. C. Chan, N. Shenoy, A. Kirkpatrick, J. A. M. Ramshaw and B. Brodsky, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 4273–4278.
440 J. Pennington, R. Socher and C. D. Manning, in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2014, pp. 1532–1543.
441 C. Mirabello and B. Wallner, *bioRxiv*, 2018, 394437.
442 L. Qin, G. Dong and J. Peng, in *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 708–715.

443 X. Li and D. Fourches, *J. Chem. Inf. Model.*, 2021, **61**, 1560–1569.
444 S. Katoch, S. S. Chauhan and V. Kumar, *Multimed. Tools Appl.*, 2021, **80**, 8091–8126.
445 T. V. Burjanadze and E. L. Kisiriya, *Biopolymers*, 1982, **21**, 1695–1701.
446 S. Sakakibara, K. Inouye, K. Shudo, Y. Kishida, Y. Kobayashi and D. J. Prockop, *BBA - Protein Struct.*, 1973, **303**, 198–202.
447 F. W. Kotch, I. A. Guzei and R. T. Raines, *J. Am. Chem. Soc.*, 2008, **130**, 2952–2953.
448 R. T. Raines, *Protein Sci.*, 2006, **15**, 1219–1225.
449 S. Xu, M. Gu, K. Wu and G. Li, *J. Phys. Chem. B*, 2019, **123**, 7754–7763.
450 A. V Persikov, J. A. M. Ramshaw and B. Brodsky, *J. Biol. Chem.*, 2005, **280**, 19343–19349.
451 A. Snellman, H. Tu, T. Väisänen, A. P. Kvist, P. Huhtala and T. Pihlajaniemi, *EMBO J.*, 2000, **19**, 5051–5059.
452 S. K. Areida, D. P. Reinhardt, P. K. Müller, P. P. Fietzek, J. Köwitzi, M. P. Marinkovich and H. Notbohm, *J. Biol. Chem.*, 2001, **276**, 1594–1601.
453 H. P. Bachinger, P. Bruckner, R. Timpl and J. Engel, *The Role of Cis-Trans Isomerization of Peptide Bonds in the Coil + Triple Helix Conversion of Collagen*, 1978.
454 H. P. Bachinger, P. Bruckner, R. Timpl, D. J. Prockop and J. Engel, *Folding Mechanism of the Triple Helix in Type-I11 Collagen and Type-I11 pN-Collagen Role of Disulfide Bridges and Peptide Bond Isomerization*, 1980, vol. 106.
455 A. V. Buevich, Q. Dai, X. Liu, B. Brodsky and J. Baum, *Biochemistry,* 2000, **15**, 4299-4308.
456 K. J. Doege and J. H. Fessler, *J. Biol. Chem.*, 1986, **261**, 8924–8935.
457 S. H. McLaughlin and N. J. Bulleid, *Matrix Biol.*, 1998, **16**, 369–377.
458 A. V. Buevich, T. Silva, B. Brodsky and J. Baum, *J. Biol. Chem.*, 2004, **279**, 46890–5.
459 C. M. Stultz, *Protein Sci.*, 2006, **15**, 2166–77.
460 S. Keten and M. J. Buehler, *Phys. Rev. Lett.*, 2008, **100**, 1–4.
461 S. Keten and M. J. Buehler, *Nano Lett.*, 2008, **8**, 743–748.
462 T. Ackbarow, X. Chen, S. Keten and M. J. Buehler, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 16410–16415.
463 A. A. Culbert, M. P. Lowe, M. Atkinson, P. H. Byers, G. A. Wallis and K. E. Kadler, *Biochem. J.*, 1995, **311**, 815–820.
464 D. L. Bodian, B. Madhan, B. Brodsky and T. E. Klein, *Biochemistry*, 2008, **47**, 5424–5432.
465 A. V Buevich, T. Silva, B. Brodsky and J. Baum, *J. Biol. Chem.*, 2004, **279**, 46890–5.
466 H. Cheng, S. Rashid, Z. Yu, A. Yoshizumi, E. Hwang and B. Brodsky, *J. Biol. Chem.*, 2011, **286**, 2041–2046.
467 D. R. Walker, S. A. H. Hulgan, C. M. Peterson, I. C. Li, K. J. Gonzalez and J. D. Hartgerink, *Nat. Chem.*, 2021, **13**, 260–269.
468 K. N. Brown and J. Cagan, *Artif. Intell. Eng. Des. Anal. Manuf. AIEDAM*, 1997, **11**, 219–235.
469 M. D. Shoulders and R. T. Raines, *Annu. Rev. Biochem.*, 2009, 78, 929–958.
470 M. W. H. Kirkness and N. R. Forde, *Biophys. J.*, 2018, **114**, 570–576.
471 A. Terzi, E. Storelli, S. Bettini, T. Sibillano, D. Altamura, L. Salvatore, M. Madaghiele, A. Romano, D. Siliqi, M. Ladisa, L. De Caro, A. Quattrini, L. Valli, A. Sannino and C. Giannini, *Sci. Rep.*, 2018, **8**, 1429.
472 J. Xie, M. Bao, S. M. C. Bruekers and W. T. S. Huck, *ACS Appl. Mater. Interfaces*, 2017, **9**, 19630–19637.
473 T. Novak, B. Seelbinder, C. M. Twitchell, S. L. Voytik-Harbin and C. P. Neu, *Adv. Funct. Mater.*, 2016, **26**, 5427–5436.
474 L. S. Wang, J. E. Chung, P. Pui-Yik Chan and M. Kurisawa, *Biomaterials*, 2010, **31**, 1148–1157.
475 N. Keshwani, S. Banerjee, B. Brodsky and G. I. Makhatadze, *Biophys. J.*, 2013, **105**, 1681–1688.
476 M. J. Buehler, F. F. Abraham and H. Gao, *Nature*, 2003, **426**, 141–146.
477 V. Putignano, A. Rosato, L. Banci and C. Andreini, *Nucleic Acids Res.*, 2018, **46**, D459–D464.
478 S. W. Cranford, A. Tarakanova, N. M. Pugno and M. J. Buehler, *Nature*, 2012, 482, 72–76.
479 S. Keten, Z. Xu, B. Ihle and M. J. Buehler, *Nat. Mater.*, 2010, **9**, 359–367.
480 I. Su, Z. Qin, T. Saraceno, A. Krell, R. Mühlethaler, A. Bisshop and M. J. Buehler, *J. R. Soc. Interface*, 2018, **15**, 20180193.
481 B. Marco-Dufort, R. Iten and M. W. Tibbitt, *J. Am. Chem. Soc.*, 2020, **142**, 15371–15385.
482 J. Ryu, S. H. Ku, H. Lee and C. B. Park, *Adv. Funct. Mater.*, 2010, **20**, 2132–2139.
483 Y. Z. Zhou, Y. Cao, W. Liu, C. H. Chu and Q. L. Li, *ACS Appl. Mater. Interfaces*, 2012, **4**, 6901–6910.
484 A. U. Regitsky, B. Keshavarz, G. H. McKinley and N. Holten-Andersen, *Biomacromolecules*, 2017, **18**, 4067–4074.
485 C. González-Obeso, M. González-Pérez, J. F. Mano, M. Alonso and J. C. Rodríguez-Cabello, *Small*, 2020, **16**, 2005191.
486 E. Khare, C. Gonzalez-Obeso, D. L. Kaplan and M. J. Buehler, *ACS Biomater. Sci. Eng.*, 2022, **8**, 4301–4310.
487 E. Khare, C. H. Yu, C. Gonzalez Obeso, M. Milazzo, D. L. Kaplan and M. J. Buehler, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2209524119.

488 J. H. Dawson, *Probing Structure-Function Relations in Heme-Containing Oxygenases and Peroxidases*, W. Horz and W. Altenburger, 1988.
489 R. H. Holm, P. Kennepohl and E. I. Solomon, *Chem. Rev.*, 1996, **96**, 2239–2314.
490 S. J. Lippard and J. M. Berg, in *Principles of Bioinorganic Chemistry*, 1994, pp. 349–377.
491 M. Hendrich, *J. Am. Chem. Soc.*, 2001, **123**, 191–191.
492 T. J. Meyer, *Pure Appl. Chem.*, 1986, **58**, 1193–1206.
493 N. L. Rosi, J. Eckert, M. Eddaoudi, D. T. Vodak, J. Kim, M. O'Keeffe and O. M. Yaghi, *Science (80-.).*, 2003, **300**, 1127–1129.
494 H. Furukawa, N. Ko, Y. B. Go, N. Aratani, S. B. Choi, E. Choi, A. Ö. Yazaydin, R. Q. Snurr, M. O'Keeffe, J. Kim and O. M. Yaghi, *Science (80-.).*, 2010, **329**, 424–428.
495 J. Lee, O. K. Farha, J. Roberts, K. A. Scheidt, S. T. Nguyen and J. T. Hupp, *Chem. Soc. Rev.*, 2009, **38**, 1450–1459.
496 H. Furukawa, K. E. Cordova, M. O'Keefe and O. M. Yaghi, *Science (80-.).*, 2013, **341**, 1230444.
497 C. J. Pedersen, *J. Am. Chem. Soc.*, 1967, **89**, 7017–7036.
498 D. J. Cram, *Angew. Chemie Int. Ed. English*, 1988, 27, 1009–1020.
499 J. -M Lehn, *Angew. Chemie Int. Ed. English*, 1988, **27**, 89–112.
500 F. Wang, J. Zhang, X. Ding, S. Dong, M. Liu, B. Zheng, S. Li, L. Wu, Y. Yu, H. W. Gibson and F. Huang, *Angew. Chemie - Int. Ed.*, 2010, **49**, 1090–1094.
501 J. M. Lehn, A. Rigault, J. Siegel, J. Harrowfield, B. Chevrier and D. Moras, *Proc. Natl. Acad. Sci. U. S. A.*, 1987, **84**, 2565–2569.
502 R. W. Saalfrank, H. Maid and A. Scheurer, *Angew. Chemie - Int. Ed.*, 2008, 47, 8794–8824.
503 K. C. Bentz and S. M. Cohen, *Angew. Chemie - Int. Ed.*, 2018, **57**, 14992–15001.
504 F. Paquin, J. Rivnay, A. Salleo, N. Stingelin and C. Silva, *J. Mater. Chem. C*, 2015, **3**, 10715–10722.
505 M. J. Wiester, P. A. Ulmann and C. A. Mirkin, *Angew. Chemie - Int. Ed.*, 2011, 50, 114–137.
506 Y. Shimazaki, M. Takani and O. Yamauchi, *Dalt. Trans.*, 2009, 7854–7869.
507 J. Yang, M. A. Cohen Stuart and M. Kamperman, *Chem. Soc. Rev.*, 2014, **43**, 8271–8298.
508 M. D'Ischia, A. Napolitano, V. Ball, C. T. Chen and M. J. Buehler, *Acc. Chem. Res.*, 2014, **47**, 3541–3550.
509 W. C. Yount, D. M. Loveless and S. L. Craig, *J. Am. Chem. Soc.*, 2005, **127**, 14488–14496.
510 E. Evans and K. Ritchie, *Biophys. J.*, 1999, **76**, 2439–2447.
511 J. Bertaud, J. Hester, D. D. Jimenez and M. J. Buehler, *J. Phys. Condens. Matter*, 2010, **22**, 035102.
512 S. Keten and M. J. Buehler, *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 2008, **78**, 061913.
513 J. Husson and F. Pincet, *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 2008, **77**, 026108.