# The use of coarse-grained polymers to model structural features of folded proteins

## Objectives:

- Explain why all-atom MD is powerful but still too expensive and assumption-heavy for many protein-folding questions.
- Build a hierarchy of coarse-grained protein models, from a collapsed random walk to side-chain resolved models.
- Compare models against the structural signatures of folded proteins: $R_g(n)$, fraction core, packing fraction, and structure factor S(q).
- Identify when coarse-grained modeling is the right tool, and when atomistic refinement is still needed.

## Key Concepts and Definitions:

- **Protein backbone and side chains.** Proteins are polymers of amino acids; the backbone is common to all residues, while side chains control most of the residue-specific geometry and chemistry.
- **All-atom MD.** Standard molecular dynamics tracks every atom with high resolution by solving Newton's equations of motion at very small time steps. Time scales are limited (too slow to study some protein folding) and the force field contains many modeling assumptions.
- **Coarse-grained (CG) model.** A CG model combines several atoms into one interaction site, or bead. Fewer beads make simulations faster and easier to control by lowering the number of degrees of freedom, allowing researchers to study longer time scales. The Martini 3 framework is a forcefield that explicitly separates mapping/bead types from structure-bias terms, which alleviated a long-term problem in MD simulations of certain molecules interacting too strongly.
- **Radius of gyration.** For a chain of *N* beads, .

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |\mathbf{r}_i - \mathbf{r}_{\mathrm{cm}}|^2}, \qquad \mathbf{r}_{\mathrm{cm}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{r}_i.$$

$R_g$ measures how spread out a section of a protein chain is. To calculate the subchain version, $R_g(n)$, you choose a segment length n, measure the spread of every continuous protein segment with that length, and then average the values. By repeating this for many different values of n, you can see how protein compactness changes from short segments to longer segments. Folded proteins show two distinct slope patterns on a log-log plot: short subchains behave one way, while longer subchains behave another way because they are constrained by the overall folded structure. This two-regime scaling is different from what is expected for a simple polymer and can act as a fingerprint of protein-like structure.

- **rSASA.** Relative solvent accessible surface area is a normalized burial/exposure measure of a dipeptide,

$$\mathrm{rSASA}_i = \frac{\mathrm{SASA}_i}{\mathrm{SASA}_i^{\mathrm{ref}}},$$

  often interpreted with Gly–X–Gly reference values or residue-specific maxima. 0 = completely buried, 1 = completely solvent exposed.
- **Core residue and packing fraction:** Core residues are buried residues identified from low rSASA; for each residue $\mu$, the packing fraction is typically defined as

$$\phi_\mu = \frac{v_\mu}{V_\mu^{\mathrm{Vor}}}, \qquad \langle\phi\rangle = \langle\phi_\mu\rangle_{\mu\in\mathrm{core}}.$$

  X-ray protein cores are observed near $\langle\phi\rangle \approx 0.55$.
- **Fraction core.** The fraction of amino acids in a protein that are buried in the core, calculated as $f_c = N_{core} / N$. Typically 8-10% in real proteins.
- **Structure factor.** It is defined by,

$$S(\mathbf{q}) = \frac{1}{N}\sum_{k=1}^{N}\sum_{l=1}^{N} e^{i\mathbf{q}\cdot(\mathbf{r}_k-\mathbf{r}_l)}.$$

  and describes how the protein's mass is distributed across different length scales. At small q, S(q) encodes overall size through a Guinier-like relation, complementing $R_g$.
- **Implicit solvent:** An approach where the effect of water on the protein is captured through effective potentials rather than by actually simulating water molecules. This is much faster and is how the CG simulations in this lecture were run.
- **Effective potential/force field.** The energy function used to move and score the CG system. It often includes bonded terms, which preserve local geometry, and nonbonded terms, which approximate contacts, packing, and electrostatics.

**Main content:**

**A.** The problem with all atom MD simulations…

- Proteins are polymers made of up to 20 different amino acids connected by peptide bonds. Each amino acid has a side chain that gives it its chemical character. The way a protein folds into a 3D shape is what determines its biological function.
- The standard computational tool for studying proteins is molecular dynamics simulations. Molecular dynamics (MD) works by tracking every atom and calculating how it moves at femtosecond time steps. This provides atomic-level detail, but is slow. To simulate even 100 nanoseconds of a 50,000 atom system takes roughly a day on a computing cluster. Protein folding, on the other hand, happens on timescales of seconds or longer for most proteins.

- Even the supercomputer Anton 2 built specifically for MD simulations cannot get close to one second of simulation time. So, most of the timescale range where biologically interesting things happen is just out of reach. On top of the time problem, MD force fields involve assumptions of bond interactions, hydrogen bonding corrections, and non-bonding potentials. It is not clear that these approximations hold up on long timescales since the force fields are only validated on short ones.
- So MD has two problems: it is slow, and it makes many (potentially inaccurate) assumptions.

**B.** Using coarse-grained models is faster and tunable

- Coarse-grained modeling is presented as a way to reduce cost while asking a more focused question: which structural ingredients are actually necessary to recover the hallmark features of folded proteins?
- Coarse-graining addresses the bottleneck of a large number of particles by replacing groups of atoms with beads that capture the slow, collective degrees of freedom that dominate many processes.
- In practice, this looks like collapsing each amino acid down to one or a few spheres instead of keeping all 25 atoms per residue. This alone reduces the number of particles by roughly a factor of ten. Because the simulation is three dimensional, this means the speed increase scales roughly as $10^3$. On top of that, using implicit solvent further increases speed, by reducing the need to model each water molecule. Further, CG models allow for tunability; the researcher sets which interactions to include in the model.

**Technical deep dive: simulation logic**

The CG workflow can be summarized as

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} U(\{\mathbf{r}\}),$$

followed by time integration under those forces. In a standard explicit integrator, one representative update is

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{\Delta t^2}{2m_i}\mathbf{F}_i(t),$$

with velocities updated consistently from the same forces. The lecture's protocol uses this idea to randomize a starting configuration and then collapse the chain under simple forces before comparing structural observables.

**C.** Four metrics for comparing models to real proteins

1. Subpolymer radius of gyration ($R_g(n)$). For simple polymers, $R_g(n) \sim N^{\nu}$ with $\nu = 1$ for a fully extended chain, $\nu = 1/2$ for a random walk, and $\nu = 1/3$ for a collapsed polymer. Folded proteins also behave as compact objects overall, but their internal scaling $R_g(n)$ versus subchain length *n* shows two regimes rather than one: a larger exponent at small *n* (< 20-25 residues) and a smaller exponent at large n. This kink is essentially a fingerprint of folded protein structure and no simple polymer model reproduces it on its own.

2. Fraction core. This is the fraction of amino acids buried in the protein core, where the core is defined by a low rSASA value. In real proteins this is about 8 to 10% regardless of protein size. Even for proteins with 1,500 amino acids, only about 10% of them are in the core. This is a benchmark any CG model needs to hit.
3. Core packing fraction. This is measured using Voronoi tessellation and tells us how tightly the core amino acids are packed together. In X-ray crystal structures this value consistently sits around 0.55 with a very tight distribution. This suggests that all folded proteins pack their cores in roughly the same way, and it is a useful universal test for any model.

**Technical deep dive: rSASA and packing fraction**

The clean way to think about the workflow is:

1. compute $\mathrm{SASA}_i$ for each residue;
2. normalize to $\mathrm{rSASA}_i = \mathrm{SASA}_i / \mathrm{SASA}_i^{\mathrm{ref}}$;
3. mark low-rSASA residues as core;
4. compute $\phi_\mu = v_\mu / V_\mu^{\mathrm{Vor}}$ for each core residue;
5. average over core residues.

4. Structure factor *S(q)*.
    a. Also improves with side chain complexity.

D. Six coarse grained models in order of complexity

1. Collapsed random walk.
    a. Each amino acid is one backbone sphere placed at the Cα position
    b. Forces
        i. Harmonic bond potential: a repulsive potential to stop spheres from overlapping
        ii. Central attractive force that makes the chain collapse inward to mimic folding
    c. Does not reproduce correctly and gets fraction core and packing fraction wrong.
2. Bond Angle Dihedral Angle (BADA) model.
    a. Adds bond angle and dihedral angle potentials, built from the real distributions of these angles in protein crystal structures.
    b. The bond angle distribution has a big peak near 90 degrees from alpha-helices and a secondary peak near 120 degrees from beta-sheets.
        i. Researchers smoothed this distribution rather than copying it into the potential to avoid biasing toward a secondary structure type
3. Freely-Jointed Side Chain (FJSC) model.
    a. Adds a single side chain sphere to each backbone bead.
        i. The sphere size is drawn from a distribution specific to that amino acid type, since different amino acids have very different side chain sizes.
        ii. The side chain can rotate freely around the backbone bond.

    b. Adding even just one side chain sphere per amino acid makes a big difference to the $R_g(n)$ curve and makes it look protein-like.
4. InSeq model.
    a. Variant of FJSC model that uses the actual amino acid sequence from each protein in the dataset rather than randomly assigning identities, which makes the size distributions more realistic.
5. The Multi Particle Side Chain (MPSC) model.
    a. Instead of one sphere per side chain, each amino acid gets multiple spheres arranged to match the actual shape of that side chain.
    b. This captures the fact that side chains are not round blobs but have real geometry.
        i. For example, tryptophan gets five spheres and tyrosine gets three.
        ii. The spheres are placed to maximize their overlap with the real atomic side chain.
6. The modified Multi Particle Side Chain (modMPSC) model.
    a. Includes two spherical beads for the side chains of Leucine and Valine
    b. MPSC and modMPSC models are the only ones whose average fraction core actually falls in the 8 to 10% range seen in real proteins.

E. The applications of coarse-grained models

- The most immediate application is studying protein folding at timescales MD cannot reach. This opens up the millisecond to second timescale range where most real protein folding happens.
- Alzheimer's disease, Parkinson's disease, and Huntington's disease all involve proteins that misfold and clump together.
    - If a CG model can reproduce how these proteins aggregate in simulation, it gives researchers a way to test ideas about what causes aggregation and how to stop it. This kind of mechanistic understanding is hard to get from experiments alone.
- Another application is antibody design and drug discovery.
    - Almost every major biologic drug currently in development is a monoclonal antibody, meaning it is a protein that binds to a specific target. CG models that can represent protein-protein interactions well could help screen large numbers of candidate antibodies computationally before expensive lab tests are run.
- CG models are useful for studying mutations in disease causing proteins, since tools like AlphaFold3 are not always accurate for single amino acid mutations and do not provide thermodynamic information about folding stability.

**Discussion and Comments:**

- There is a distinction between MARTINI 3, which validates by checking how well it folds one specific protein to a target structure, and the work presented in the lecture which instead of fitting to one protein, aims to match bulk statistical properties across thousands of proteins. This is a harder but more general test.

- In all-atom MD, the hydrophobic effect comes from water molecules rearranging around hydrophobic groups. In implicit solvent models that effect has to be built directly into the potentials– this is one of the main design choices in CG modeling.
- Toward the end, a student asked about potential applications if MD could be run for a full minute. Getting to that timescale would be heroic, and since force fields are only validated at shorter timescales, accuracy at a minute would be questionable even if it were achievable.

- Overall, a useful way to read the lecture is as a test of parsimony: the question is not whether a model can be made to fold one protein, but whether it can reproduce the statistical geometry of many real proteins with as few assumptions as possible.
- In that sense, the talk's strongest lesson is that local stereochemistry and side-chain shape are enough to reshape global observables like $R_g(n)$, core fraction, and packing fraction.

**References:**


- Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. Coarse-Grained Protein Models and Their Applications. Chemical Reviews, 2016.
- Logan, Sumner, Grigas, Shattuck, O'Hern, "Effect of stereochemical constraints on the structural properties of folded proteins" (Phys. Rev. E, 2025; preprint/abstract pages). https: //journals.aps.org/pre/abstract/10.1103/9wf9-ywhw
- Logan, Sumner, Grigas, Shattuck, O'Hern, "The effect of stereochemical constraints on the radius of gyration of folded proteins" (arXiv 2501.02424 / preprint). https://www.researchgate.net/publication/387767558_The_effect_of_stereochemical_constraints_on_the_radius_of_gyration_of_folded_proteins
- O'Hern group papers/review pages on packing in protein cores and SASA/rSASA-based core identification. https://pmc.ncbi.nlm.nih.gov/articles/PMC7415476/
- Souza et al., Martini 3: a general purpose force field for coarse-grained molecular dynamics (Nature Methods, 2021). https://cgmartini.nl/docs/publications/entries/2021/Souza2021_Martini3.html
- Martini Force Field Initiative tutorial on Martini 3 protein models and structure-bias layers. https://cgmartini.nl/docs/tutorials/Martini3/ProteinsI/
- Logan JA, Sumner J, Grigas AT, Shattuck MD, O'Hern CS. Effect of stereochemical constraints on the structural properties of folded proteins.
- Physical Review E. 2025 Nov;112(5):054405. Wang Y, Csanyi G, Ortner C. Many-body coarse-grained molecular dynamics with the atomic cluster expansion. arXiv preprint arXiv:2502.04661. 2025 Feb 7.
- Voth G. Systematic Coarse-graining of Molecular Dynamics Simulations. InAPS March Meeting Abstracts 2015 Mar (Vol. 2015, pp. D19-004).


**Suggested Reading:**

- Joshi SY, Deshmukh SA. A review of advancements in coarse-grained molecular dynamics simulations. Molecular Simulation. 2021 Jul 24;47(10-11):786-803.
- Pak AJ, Voth GA. Advances in coarse-grained modeling of macromolecular complexes. Current opinion in structural biology. 2018 Oct 1;52:119-26.
- Frenkel D, Smit B. Understanding molecular simulation: from algorithms to applications. elsevier; 2023 Jul 13.