Finalized Lecture Summary on Core Packing

What this lecture is about

The lecture sits at the intersection of biophysics and computational structural biology and asks a deceptively simple question: how is the inside of a folded protein actually filled with atoms, and how well can a simple physical model predict the side-chain conformations that fill it? The central claim advanced is that a carefully calibrated hard-sphere, steric-repulsion model with explicit hydrogens can explain a large share of protein-core geometry — particularly the side-chain conformations of buried hydrophobic residues — without invoking a multi-term knowledge-based force field.

Key vocabulary

Dihedral angle: the angle between two planes formed by four atoms; used to describe backbone (φ, ψ) and side-chain (χ) rotations. Ramachandran plot: a 2D probability density of φ vs ψ that shows the regions populated by α-helices and β-sheets and flags steric clashes. Rotamers: discrete preferred side-chain conformations at specific χ values. Hard-sphere interaction: atoms treated as non-overlapping spheres with a contact distance equal to the sum of their radii. Voronoi tessellation: a partition of space such that every point lies in the cell of its nearest atom; used here to compute local packing fractions. SASA (solvent-accessible surface area): measures which atoms are solvent-exposed; atoms with low or zero SASA are buried and form the core. Cavity / void: interior empty space not occupied by atoms. Explicit-hydrogen model: places hydrogens explicitly rather than inflating heavy-atom radii to account for them.

Backbone geometry: Ramachandran plots

Each peptide unit has dihedral angles along the backbone that act as the protein's degrees of freedom. The Ramachandran plot — φ on the horizontal axis, ψ on the vertical — is a probability density that flags steric clashes and reveals the regions populated by right- and left-handed α-helices and β-sheets. The shape of those allowed regions depends on the atomic radii used: if atoms were treated as points, the entire diagram would be open. This is the first hint of a theme that will recur throughout the lecture: what counts as "allowed" is in part a representation choice.

Side chains and rotamers

Side chains range from one χ angle (smallest) to up to five (largest). Empirically, χ values cluster at three preferred orientations rather than spreading over a continuum. For a side chain with n χ angles, the number of common rotamers is roughly 3^n; isoleucine (n = 2) therefore has 9. Source 1 paraphrases this as "the square of its number of dihedral angles," but the underlying rule is exponential, not quadratic — the isoleucine example happens to be right only because 3^2 equals 9 for n = 2. The lecture uses leucine χ scans and Ile56 in PDB 2NWD to show how, for buried residues, the surrounding protein environment is enough to single out the experimentally observed rotamer.

The benchmark task: side-chain repacking

Side-chain recovery is the standard benchmark: strip the side chains from a known structure and put them back using only a scoring rule, then check how often you recover what was there. In the minimal hard-sphere version, the model holds bond lengths, bond angles, and backbone dihedrals fixed and uses a purely repulsive non-bonded potential with Boltzmann weighting:

$$P(\chi) \propto \exp[\, -V(\chi) / (kT)\,], \quad \text{with } V(\chi) = 0$$

when atoms do not overlap and steeply repulsive when they do, and $\sigma_{ij} = (\sigma_i + \sigma_j) / 2$. With this rule, recovery is high for buried hydrophobic side chains. The take-home is striking: in the hydrophobic core, excluded volume is doing most of the work — a deeply minimalist result that the rest of the lecture builds on.

Energy functions: where does the rest of the physics earn its keep?

More general protein-design pipelines combine many terms: stereochemistry potentials that enforce equilibrium bond lengths and angles (very important), van der Waals attractive/repulsive interactions (useful), hydrogen bonding (generally good), electrostatics and desolvation (more important on the surface), and disulfide-bond energies (not very useful). Read against the hard-sphere result, the implication is that in the core most of these terms are nearly redundant; on the surface, where directional interactions and solvation matter, they earn their keep.

Hard spheres in practice: atom radii and their sensitivity

The hard-sphere assumption sets the energy to zero beyond the contact distance and to a steeply repulsive value below it; Boltzmann reweighting then converts the potential to a probability. The choice of atomic radius matters, and experimental ranges only set a ballpark. The practical recipe is to sweep radii within reasonable bounds and pick the values that reproduce the observed rotamer distributions. The packing analysis pipeline: PDB coordinates plus atom radii → define the core via burial / SASA / residue criteria → quantify with Voronoi volumes, cavities, and rotamers → interpret for stability, model realism, and designability.

Packing fraction: 0.56, not 0.74

The packing fraction φ is the proportion of space occupied by atoms. For ordered identical spheres, the densest 3D arrangement (face-centred cubic) reaches $\varphi \approx 0.74$. Older protein-core estimates of $\varphi \approx 0.70$–0.74 came from extended-atom representations. Newer explicit-hydrogen analyses, with radii calibrated against observed side-chain distributions, instead give

$$\varphi_{core} \approx 0.56,$$

a value also confirmed by an independent jamming simulation in which amino-acid-shaped particles with realistic stereochemistry are isotropically compressed until they jam. This shift is one of the headline results of the lecture, and it has the structure of a representation correction rather than an empirical surprise: the older and newer numbers do not actually disagree about the data, they disagree about what "core" and "atom" mean.

Voronoi tessellation: how φ is actually computed

In 2D, a Voronoi partition draws perpendicular bisectors between point pairs so that each polygon contains all points closer to its centre than to any other. Local packing fraction is then disc area / polygon area; averaging gives the global value. The 3D extension uses an agglomeration of overlapping spheres. Formally, the residue-level packing fraction is

$$\varphi_r = (\Sigma_i V_i) / (\Sigma_i V_i^v)$$

where V_i is the occupied volume of atom i and V_i^v is its Voronoi cell volume. Applying the same ratio over all core atoms gives the core-level fraction φ_{core}. The point of this coarse-graining is that one interpretable scalar can then be compared across different structures, models, or design candidates.

Physical interpretation: jammed, not crystalline

If protein cores really packed near 0.74 like ordered spheres, they should display much stronger bond-orientational order than they actually do. The lecture proposes that the right physical analogy is instead a jammed packing of elongated, rough particles. Side-chain shape roughness lowers the random-packing limit to the observed ~0.56. This also resolves a longstanding interpretive tension: older near-crystal numbers and newer ~0.56 numbers are not in contradiction; they differ because of representation choices (extended atoms vs explicit hydrogens, definition of "core," treatment of cavities). Across generations, the consistent finding is that cores are dense and solid-like, just not crystalline.

Where the minimal model breaks down

Serine is the informative failure case: a small polar side chain whose conformation is shaped by hydrogen bonding more than by sterics, so a pure hard-sphere model under-predicts. Source 1 makes the same observation in passing — serine is hard because it hydrogen-bonds to neighbours; methionine is hard because it has three χ angles and low electron density. More broadly, any context dominated by directional or polar interactions — surfaces, binding sites, allosteric pockets — will need more than hard spheres.

Why core packing matters in practice

Packing controls (i) stability, because well-packed cores reduce voids and improve van der Waals complementarity; (ii) mutational tolerance and design, because many hydrophobic sequences can support the same fold but only if their volumes and side-chain geometries remain compatible with the core architecture; and (iii) structure validation, because under-packed regions can flag unrealistic models even when other scores look acceptable. Tools like RosettaHoles target exactly these under-packed voids that pairwise scores tend to miss.

Brief case studies

Richards (1974) turned the qualitative claim that interiors are tightly packed into measurable atomic and group volumes; packing became a structural observable rather than an intuition. Lim & Sauer's λ-repressor work (1989) showed that more than one hydrophobic repacking arrangement can support the same fold, meaning cores are constrained but not unique. Karpusas et al.'s T4 lysozyme cavity-filling mutants (1989) showed that filling voids stabilizes the protein only when it does not introduce strain; packing is a tradeoff, not a monotonic goal. Gaines, Regan, and O'Hern's modern reanalysis with explicit hydrogens produced the $\varphi_{core} \approx 0.56$ result and described the core as a random close packing of irregular particles.

Connections to statistics / machine learning

Several pieces of the methodology map cleanly onto familiar ML/statistics ideas: side-chain recovery and rotamer prediction are supervised learning problems (ESL/ISL Ch. 2); energy minimization is conceptually parallel to regression (ESL Ch. 3, ISL Ch. 6); parameterising and constraining energy

functions is a form of regularization (ESL Ch. 5, ISL Ch. 6); and protein structure prediction lives in the high-dimensional regime of ESL Ch. 18. Worth flagging alongside this: AlphaFold and RoseTTAFold sit at the opposite end of the spectrum — they bypass explicit force fields almost entirely and learn the geometry directly from sequence and coevolution. The hard-sphere result and the deep-learning result are doing different jobs: one explains why a simple physical principle is enough for a particular structural question, the other shows that data-driven models can short-cut the physics altogether when enough sequence-structure pairs are available.