Experimental datasets for benchmarking protein force fields [Article v0.1]

Firstname Middlename Surname[1*], **Firstname Middlename Familyname**[1,2†§], **Firstname Initials Surname**[2†¶], **Firstname Surname**[2*]

[1]Institution 1; [2]Institution 2

Abstract 250 word limit

This LiveCoMS document is maintained online on GitHub at https://github.com/openforcefield/protein-benchmark-data; to provide feedback, suggestions, or help improve it, please visit the GitHub repository and participate via the issue tracker.

This version dated June 23, 2021

***For correspondence:**
email1@example.com (FMS); email2@example.com (FS)

[†]These authors contributed equally to this work
[‡]These authors also contributed equally to this work

Present address: [§]Department, Institute, Country; [¶]Department, Institute, Country

1 Introduction

- Background
  - Role of molecular dynamics in understanding protein structure and function and in drug design
  - Brief history of protein force fields
- Gap in field
  - Force fields are trained against different empirical targets and are expected to describe some behaviors well and others poorly
  - Force fields for proteins often borrow parameters from more general force fields that aim to describe broader chemistry
  - Need for a comprehensive collection of experimental datasets that interrogates a wide range of physical properties of proteins
- Goals of current review
  - Description of available datasets and not prescription of how comparisons should be made
  - Focus on peptides and globular proteins without ligands or cofactors to narrow scope
- Target audience
  - Researchers involved in developing or assessing protein force fields
  - Assume familiarity with molecular dynamics techniques, force field terms, and basics of protein structure

- Review format
  - Explanation of Perpetual Review format
  - Instructions for community involvement
- Outline of review sections

2 Goals of benchmark datasets

- Target observables
  - Target experimental observables instead of structural models or quantum chemistry data
- Accessibility
  - Identify datasets that are accessible without paywalls or restrictive licenses
- Multiple scales
  - Identify observables that interrogate physical properties at different length and time scales
  - Goal is to assess force fields rather than train parameters, so computational cost can be high
  - System size should range from small—small enough to sample an ensemble exhaustively—to medium—large enough to exhibit stable folding behaviors
- Discriminatory power
  - Identify systems that can discriminate between force fields
  - For example, most protein force fields can describe lysozyme well

3 Room-temperature (RT) crystallography

- Advantages of RT crystals
  - RT crystals are higher quality and exhibit lower mosaicity than low-temperature crystals
  - Proteins in RT crystals fluctuate more than those in low temperature crystals
  - Observables are accessible in public databases in a common format
- Observables
  - Electron density
    - Electron density is independent of a structural model unless molecular replacement was used to solve phases
    - Electron density from solvent molecules can be included
    - Comparing simulations to experiments
      - Quality metrics for structural models, e.g. R-factors or correlation coefficients, are likely too sensitive to meaningfully discriminate between force fields
      - Differences can be visualized by an $F_O - F_C$ map
      - A quantitative metric is a comparison between a structural model refined against simulated electron density and a structural model refined against experimental density, e.g. an RMSD
  - Reflections
    - Raw reflections are totally independent of a structural model
    - Reflections are available in PDB entries
    - Non-Bragg peaks from diffuse scattering inform on large-scale fluctuations
  - Debye-Waller (B) factors
    - B factors are available in PDB entries
    - B factors inform on local flexibility
    - A drawback is that B factors may reflect disorder in the crystal lattice rather than flexibility of the crystallized molecules
  - Populations of alternative conformations
    - Although alternative conformations rely on a structural model, this low resolution metric may discriminate between force fields that perform similarly on other observables
- Running crystal simulations
  - Simulations of single unit cells are less expensive but may miss fluctuations that are

important for some observables
  - Simulation of supercells are more realistic but may fail to maintain the correct symmetry
  - May need to include co-solvents in mother liquor

- Systems
  - Criteria/desiderata
    - High resolution (<= 1.2 Å) crystals to ensure high quality target data and identify tautomers and protonation states
    - Protonation state can be determined unambiguously by neutron diffraction
    - Aim for diversity in secondary structure
    - Systems for which data from multiple crystals with different symmetry are available are useful
  - Systems
    - David Case
    - Julian Chen
    - James Fraser
    - Daniel Keedy
    - Michael Wall

4 Nuclear magnetic resonance spectroscopy

- Advantages of NMR
  - NMR experiments are performed in the desired ensemble for most applications
  - Comparison to NMR data may reveal native state bias that is difficult to diagnose with crystal simulations
  - Many NMR observables can be related to specific FF terms
- Observables
  - Chemical shift
    - Easily accessible for many systems in BMRB
    - Directly informs on local backbone conformation for unstructured peptides and disordered proteins
    - Difficult to interpret for larger, folded proteins due to aromatic ring currents, spin diffusion, etc.
  - Scalar coupling
    - Scalar coupling values for backbone amide proton inform on local backbone conformation
    - Requires Karplus parameters, which can be derived from QM
  - Helical propensities (merge with chemical shift section?)
    - $^{13}C=O$ chemical shifts inform on helical propensities of amino acids
    - Benchmarks can target chemical shifts directly or Lifson-Roig helix extension parameters
  - Nuclear Overhauser effect (NOE) spectroscopy
    - NOEs inform on interactions between residues distant in primary sequence
    - NOE intensities are nonlinear averages that are difficult to converge, so they may serve better as ordinal (i.e. strong/medium/weak) rather than quantitative assessments
  - Residual dipolar coupling (RDC)
    - RDCs inform on large spatial motions
    - Calculating RDCs for large proteins requires computing an expensive alignment tensor
  - Spin relaxation
    - Spin relaxation rates inform on large spatial motions for folded proteins
    - Spin relaxation can discriminate between force fields that describe global conformations and those that describe only local conformations
    - There is error from zero point motion and difference between modeled and true bond lengths, but the necessary correction may be small enough to ignore
    - Spin relaxation rates will be difficult to converge for large, folded proteins
- Running NMR simulations

Table 1. Room-temperature crystallography datasets

Description	PDB ID	Experiments	Experimental references	Computational references
Endoglucanase	3X2P	X-ray diffraction Neutron diffraction		
Scorpion toxin II	1AHO	X-ray diffraction		

  - Viscosity of water model is known to affect tumbling rates and thus spin relaxation rates
- Systems
  - Kyle Beauchamp chemical shifts and scalar couplings
  - Bernie Brooks spin relaxation dataset for lipids, good for methods
  - Lillian Chong scalar couplings for protein mimetics, good for methods
  - Kresten Lindorff-Larsen chemical shift and NOEs
  - Samuli Ollila spin relaxation dataset for proteins
  - Paul Robustelli chemical shifts, NOEs, and helical propensities
  - Lars Schäfer c-Myb chemical shifts and NOEs

5 Hydrogen-deuterium exchange (HDX) experiments

- Advantages of HDX
  - HDX informs on folding of small proteins with simple tertiary structures
  - HDX discriminates between proteins with intermediate and high folding stability that have similar bulk properties or spin relaxation rates
- Observables
  - Chemical shifts or HSQC measured by NMR
  - Mass spectrometry
  - Protection factor (exchange frequency relative to unfolded state) has an ambiguous relationship to computable quantities, e.g. free energies
- Systems
  - Gabe Rocklin and Tobin Sosnick HDX dataset
  - Vincent Shaw G proteins
  - Vincent Voelz ubiquitin, BPTI, and myoglobin

6 List of potential figures

- Visualization of protein crystal supercell
- Visualization of differences in electron density with $F_O - F_C$ map
- Solution protein structure with NMR observables labeled
  - Folded tertiary structure labeled with "RDC" and "Spin relaxation"
  - Long range contact labeled with "NOE"
  - Inset of α helix labeled with "HDX" and "Helical propensity"
  - Inset of peptide backbone with "Chemical shift" and "^3J coupling" labeled
- Histograms of observables in larger datasets (perhaps borrowed from original publications
  - Distribution of spin relaxation rates in Ollila dataset
  - Distribution of HDX exchange rates in Rocklin/Sosnick dataset

7 Conclusions

- Summarize key points
- Additional type of experiments
  - Kirkwood-Buff integrals for co-solvents

Table 2. Nuclear magnetic resonance spectroscopy datasets

Description	PDB ID	Experiments	Experimental references	Computational references
c-Myb transactivation domain	1SB0	^{1}H chemical shifts		
		NOESY		
Short peptides		HDX exchange rates		

  - Paramagnetic relaxation enhancement interactions
  - Binding free energies
  - Salt bridge dissociation rates
  - Folding observables
    * Free energies
    * Kinetic rates
    * Melting temperatures
  - Small angle x-ray scattering observables
    * Radii of gyration
    * Kratky plots
    * Pairwise distribution functions

- Additional protein systems
  - Membrane proteins (Benoit Roux)
  - Proteins with ligands or cofactors
  - Protein mimetics, e.g. peptoids or β-peptides (Lillian Chong)

8 Author Contributions

(Explain the contributions of the different authors here)

For a more detailed description of author contributions, see the GitHub issue tracking and changelog at https://github.com/openforcefield/protein-benchmark-data.

9 Other Contributions

(Explain the contributions of any non-author contributors here) For a more detailed description of contributions from the community and others, see the GitHub issue tracking and changelog at https://github.com/openforcefield/protein-benchmark-data.

10 Potentially Conflicting Interests

MKG has an equity interest in and is a cofounder and scientific advisor of VeraChem.

11 Funding Information

We thank the Open Force Field Consortium and Initiative for financial and scientific support and the Molecular Sciences Software Institute (MolSSI) for its support of the Open Force Field Initiative. We also thank the National Institutes of Health (NIH) for support of this research via NIH R01GM132386. MKG acknowledges funding from National Institute of General Medical Sciences (GM061300). These findings are solely of the authors and do not necessarily represent the views of the NIH.

Author Information

ORCID:

Author 1 name: AAAA-BBBB-CCCC-DDDD
Author 2 name: EEEE-FFFF-GGGG-HHHH