Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference

Justin L. MacCallum, Alberto Perez, and Ken A. Dill

Contributed April 13, 2015 to PNAS
Presented by Ben Lengerich on November 2, 2015

Author Information

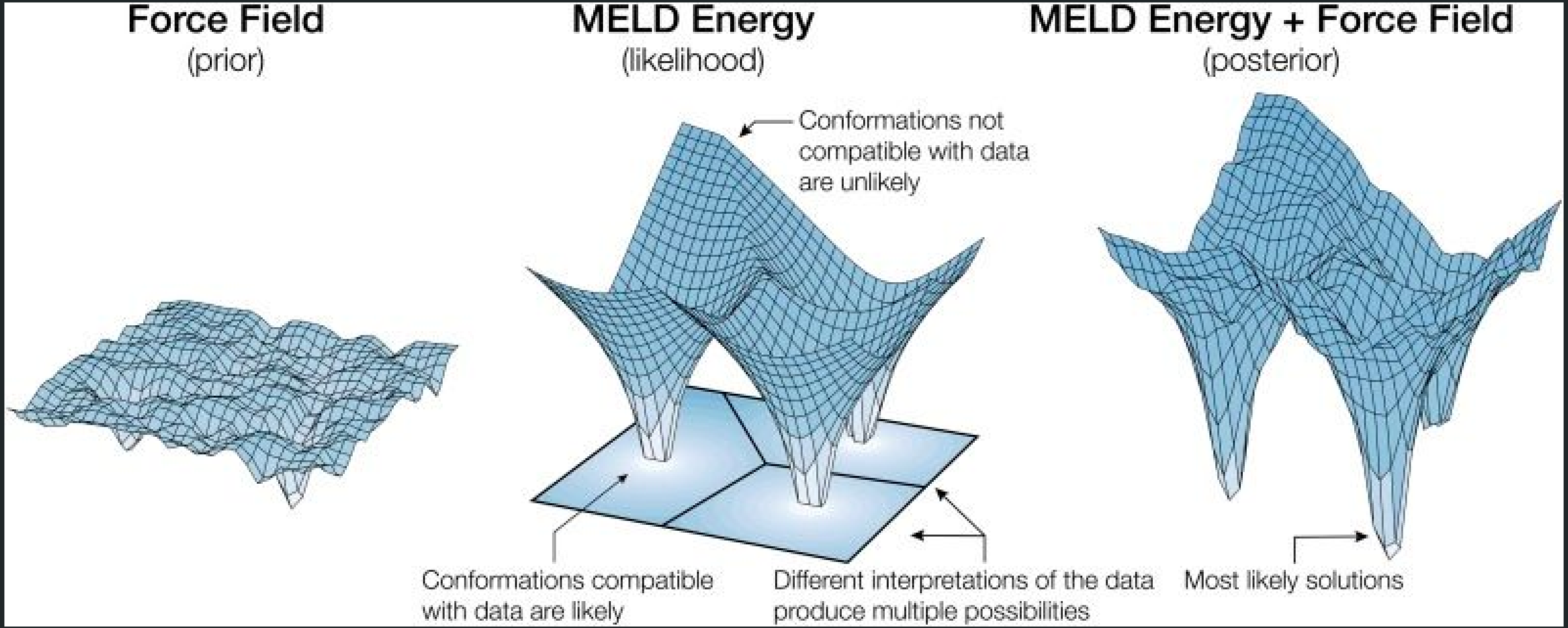
- Justin MacCallum
    - Assistant Professor, Department of Chemistry, University of Calgary
- Alberto Perez
    - Research Assistant Professor, Laufer Center for Physical and Quantitative Biology, Stony Brook University
- Ken A. Dill
    - Director of Laufer Center for Physical and Quantitative Biology, Stony Brook University

Motivation

- Most protein structures are not known at atomic detail
- We would like to be able to determine these structures from experimental data
  - MD is computationally infeasible for the necessary time scale
- However, experimental data can be unreliable
  - Uncertain - Evolution-based predictions of residue-residue contacts
  - Sparse - Solid-state NMR experiments
  - Ambiguous - spin-label EPR experiments
  - Homogeneity Bias

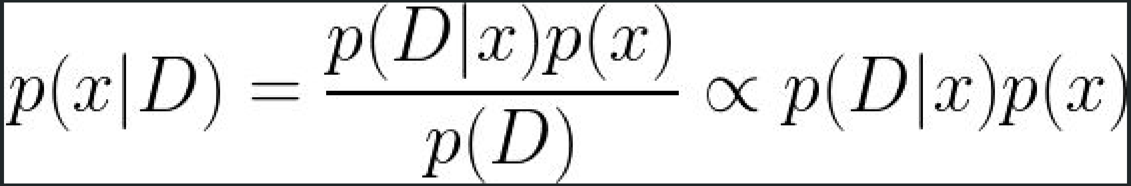
Methods

Modeling Employing Limited Data (MELD)


Force Field
(prior)
MELD Energy
(likelihood)
MELD Energy + Force Field
(posterior)
Conformations not compatible with data are unlikely
Conformations compatible with data are likely
Different interpretations of the data produce multiple possibilities
Most likely solutions

Modeling Employing Limited Data (MELD)

- MELD is a Bayesian framework which combines:
    - 3N-dimensional vector of atomic coordinates x
    - experimental data D



$$p(x|D) = \frac{p(D|x)p(x)}{p(D)} \propto p(D|x)p(x)$$

The prior probability is a Boltzmann distribution combined with a generalized-Born implicit solvation model

$$p(x) \propto exp[-\beta E_{amber}(x)]$$

where $E_{amber}(x)$ is the energy of the conformation estimated by the AMBER force field and β is a temperature parameter

This should be the final distribution as well, we are really just using the experimental data to limit our search space rather than changing the space

The likelihood function measures how well the structure agrees with the experimental restraints

For each piece of data Di, the likelihood function is

$$p(D_i|x) \propto exp[-\beta E_i^{restraint}(x)]$$

$E_i^{restraint}(x)$ is calculated by turning the experimental data into restraints (distances, torsion angles, etc.) and calculating how well the putative structure agrees with the restraints

This is identical to standard restrained MD

Spurious restraints are corrected by considering only the n restraints with lowest energy

Given n, the number of correct restraints, the likelihood function is

$$p(D|x) = \prod_{i=1}^{n} p(D_i|x) \propto \prod_{i=1}^{n} exp[-\beta E_i^{restraint}(x)]$$

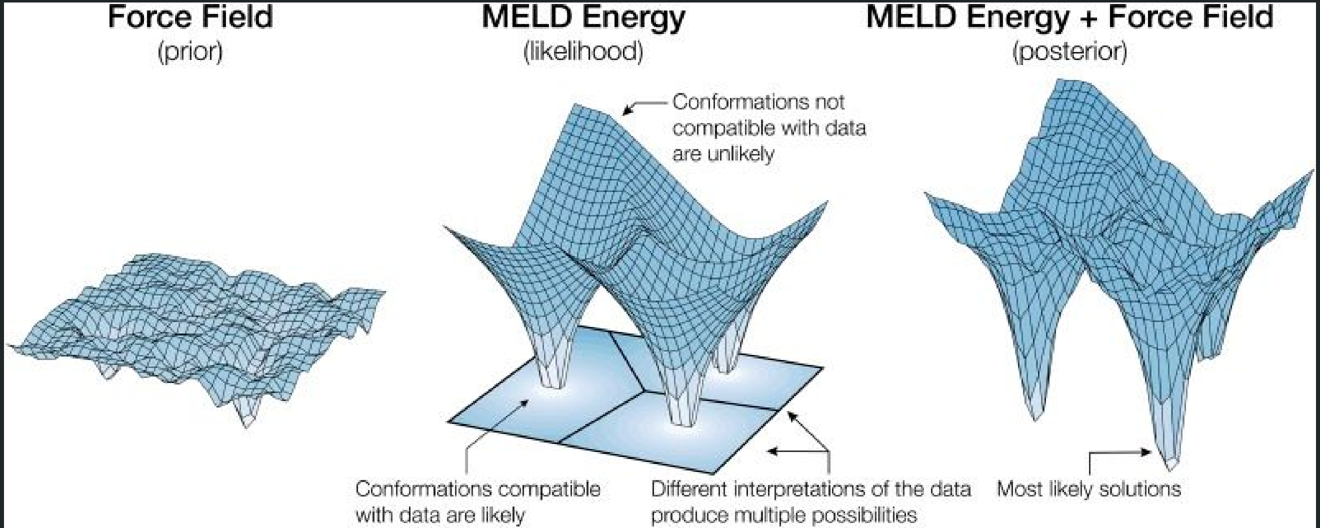
where the restraints are sorted by energy such that

$$E_1^{restraint} \leq E_2^{restraint} \leq ... \leq E_N^{restraint}$$

with N the number of total restraints.

Restraints are re-sorted at every timestep.

So the enforced restraints are different for different conformations, leading to a multi-funneled energy landscape.

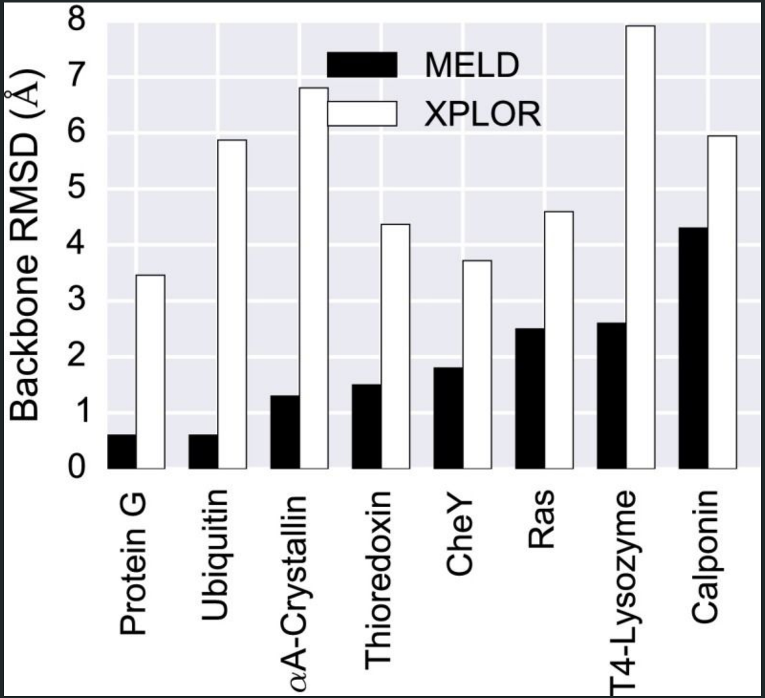

Force Field
(prior)
MELD Energy
(likelihood)
MELD Energy + Force Field
(posterior)
Conformations not compatible with data are unlikely
Conformations compatible with data are likely
Different interpretations of the data produce multiple possibilities
Most likely solutions

MELD is computationally tractable due to GPU acceleration

- Uses GPU-accelerated OpenMM library
- Avoids kinetic traps through Hamiltonian and temperature replica exchange MD

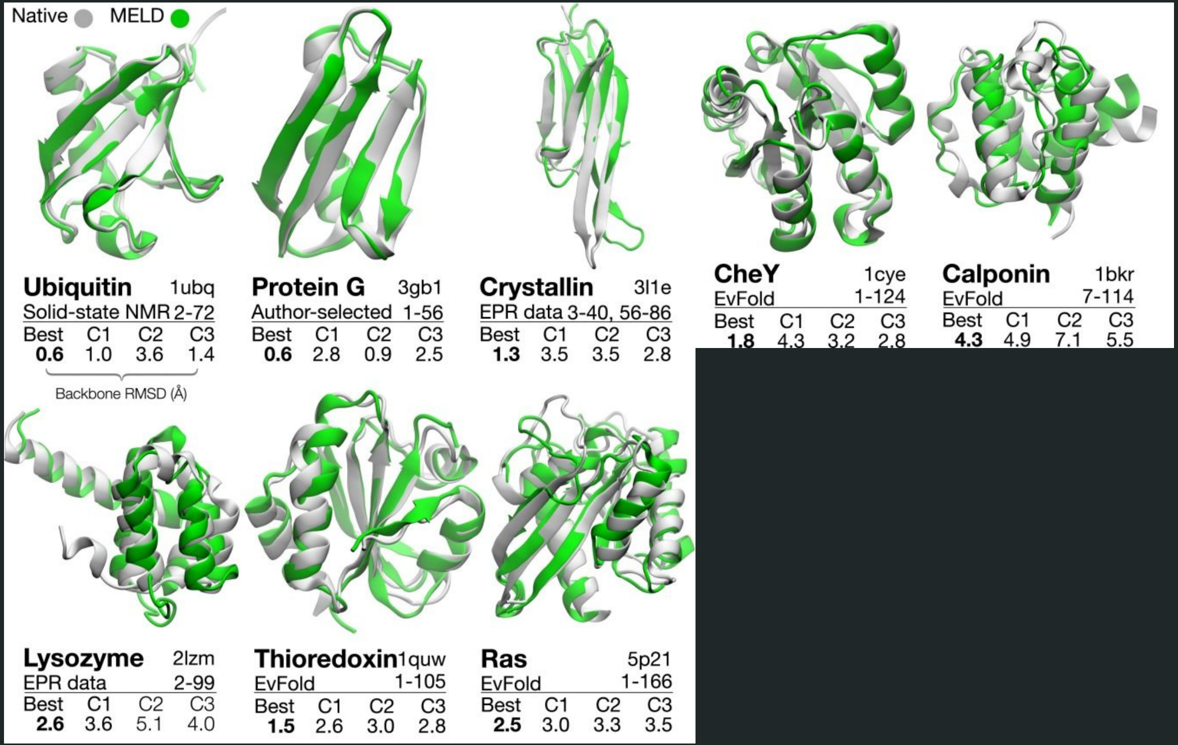
Results

MELD samples native-like structures well



MELD samples more accurate structures than X-PLOR-NIH for all test cases in this study. Each bar represents the single best structure produced for that target by each method.

MELD chooses correct structures



Ubiquitin 1ubq
Solid-state NMR 2-72

Best	C1	C2	C3
0.6	1.0	3.6	1.4

Protein G 3gb1
Author-selected 1-56

Best	C1	C2	C3
0.6	2.8	0.9	2.5

Crystallin 3l1e
EPR data 3-40, 56-86

Best	C1	C2	C3
1.3	3.5	3.5	2.8

CheY 1cye
EvFold 1-124

Best	C1	C2	C3
1.8	4.3	3.2	2.8

Calponin 1bkr
EvFold 7-114

Best	C1	C2	C3
4.3	4.9	7.1	5.5

Lysozyme 2lzm
EPR data 2-99

Best	C1	C2	C3
2.6	3.6	5.1	4.0

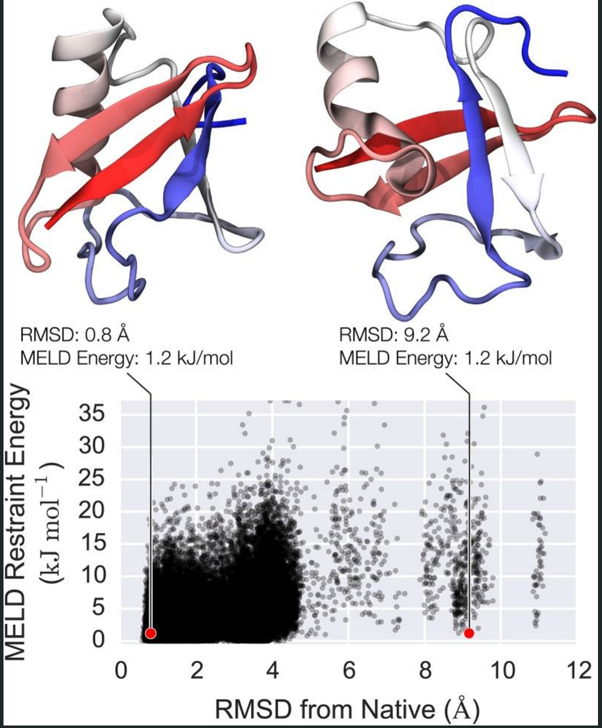
Thioredoxin 1quw
EvFold 1-105

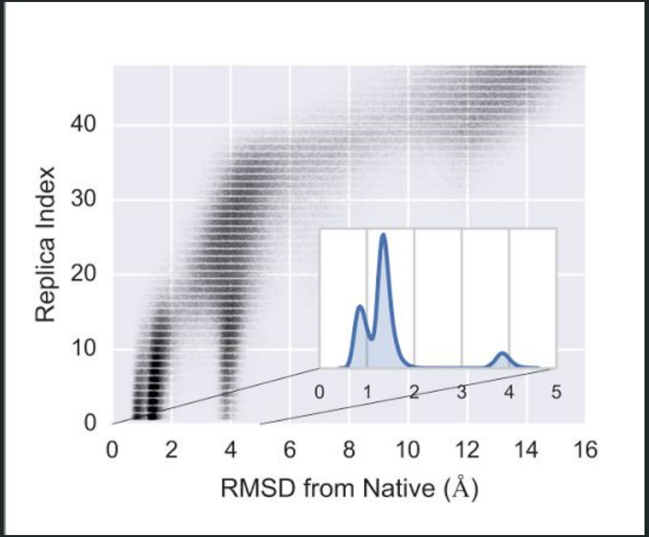
Best	C1	C2	C3
1.5	2.6	3.0	2.8

Ras 5p21
EvFold 1-166

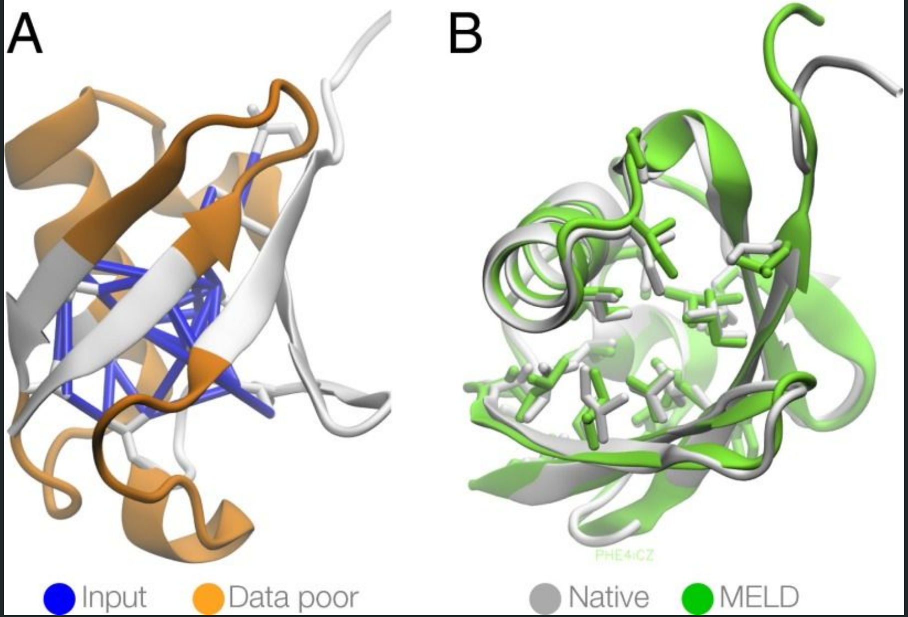
Best	C1	C2	C3
2.5	3.0	3.3	3.5

This is interesting because the experimental data does not uniquely define the structure





MELD handles sparse information well



Structure determination of ubiquitin using MELD with sparse solid-state NMR data and Talos+ secondary structure predictions. (*A*) The input restraints overlaid on the crystal structure. Data-poor regions longer than 10 residues are shown in orange. (*B*) Overlay of native and MELD prediction showing the remarkable agreement in the prediction of side-chain conformations.

MELD handles sparse information well

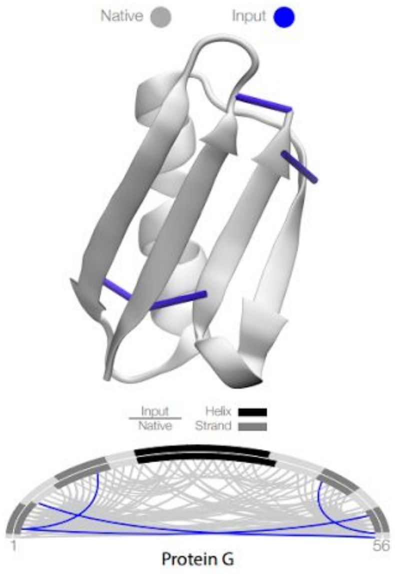
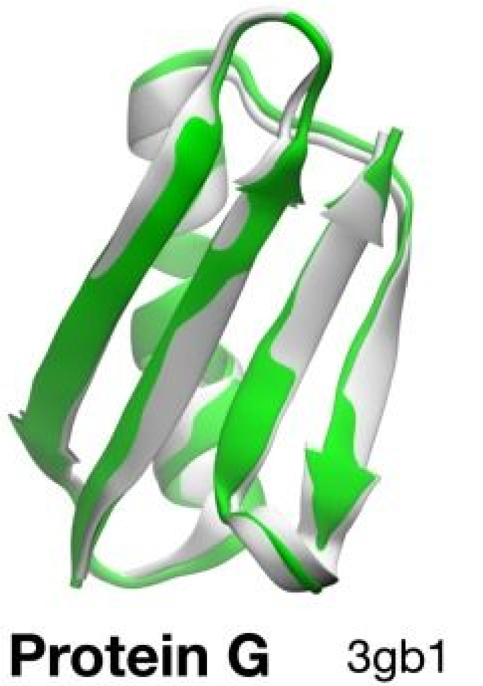


Figure S6: Data used in Protein G calculation.



MELD handles ambiguous information well

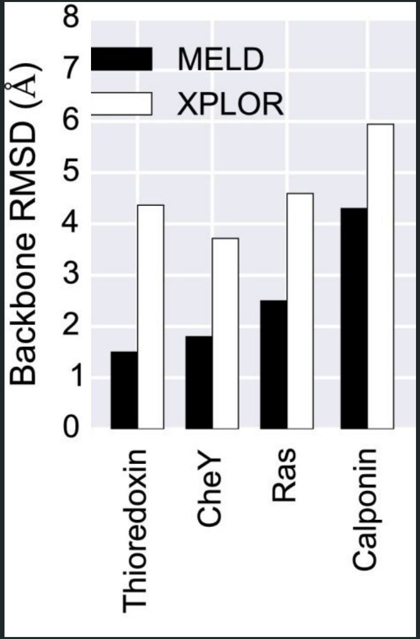
From spin-label EPR data, they obtained restraints using ROSETTA-EPR and secondary structure predictions from PSIPRED and used MELD to sample conformations for Lysozyme and Crystallin. The results outperform XPLOR and are comparable to results using ROSETTA-EPR.

	MELD	X-PLOR	ROSETTA-EPR
Lysozyme	2.6	7.9	1.8
Crystallin	1.3	6.8	4.0

MELD handles uncertain information well

Used predicted residue-residue contacts from EvFold for four targets. Restraints are predicted by co-evolution in multiple sequence alignments.

The average improvement of the most populous cluster from MELD over the lowest-energy structure for EvFold is 2.5 Å.



Conclusions

- MELD is useful for combining experimental data with atomistic modeling to determine protein structure
- Future work will focus on generalizing the method by placing priors on the parameters (active fraction, cutoff distance, etc.)
- Also will incorporate Bayesian inferences from loose insights (“hydrophobic cores”, etc.)
    - **Accelerating molecular simulations of proteins using Bayesian inference on weak information.** PNAS 2015 112 (38) 11846-11851; published ahead of print September 8, 2015,doi:10.1073/pnas.1515561112