



Strategies for *ab initio* Biomolecular Force Field Development


David S. Cerutti

Biomolecular vs. Small Molecule Force Fields





	Small Molecule Force Field	Biomolecular Force Field
Goal	Model a diverse chemical space of trillions of compounds	Model a limited space of up to a hundred common monomers
Chemistry	<ul style="list-style-type: none">• Double bonds, strained rings, other moieties• May take significant energy to synthesize• Low concentrations, tight binding desirable• Toxicity is common, metabolism uncertain	<ul style="list-style-type: none">• Building blocks are common to all of biology• Derived from familiar metabolic pathways• Produced by organisms in significant quantities• Ingested and recycled by metabolism
Key Properties	<ul style="list-style-type: none">• Hydration free energies• Binding free energies• Correct rotational profiles of critical bonds	<ul style="list-style-type: none">• Secondary and tertiary structure of polymers• Hydrogen-bonding propensities of common backbone and select side chains• Hydration characteristics
Training Strategies	<ul style="list-style-type: none">• Parameter libraries and interpolation• Training set archives of quantum data	<ul style="list-style-type: none">• Improve selected parameters based on previous successes
Validation Strategies	<ul style="list-style-type: none">• Fleets of TI or alchemical binding free energy calculations, windows in the 1-10ns timescale	<ul style="list-style-type: none">• 1000ns timescale simulations, replica exchange to study structural equilibria• NMR <i>J</i>-coupling, spin relaxation constants

The AMBER Protein Force Fields


1995

 W. Cornell


1999




 W. Cornell

 J. Wang


2006

 V. Hornak

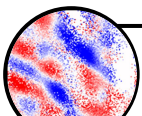
2014

 J. Maier




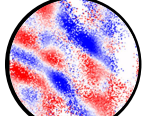

 Duan






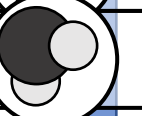
 Best



2019

 C. Tian

 L.P. Wang

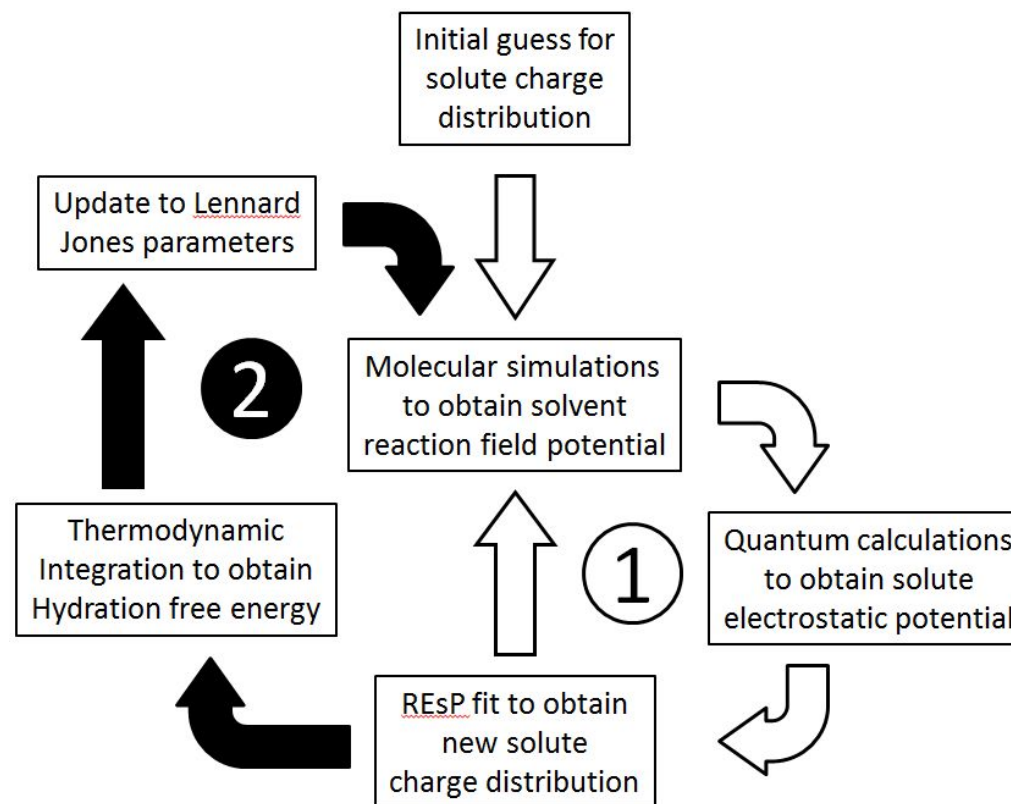
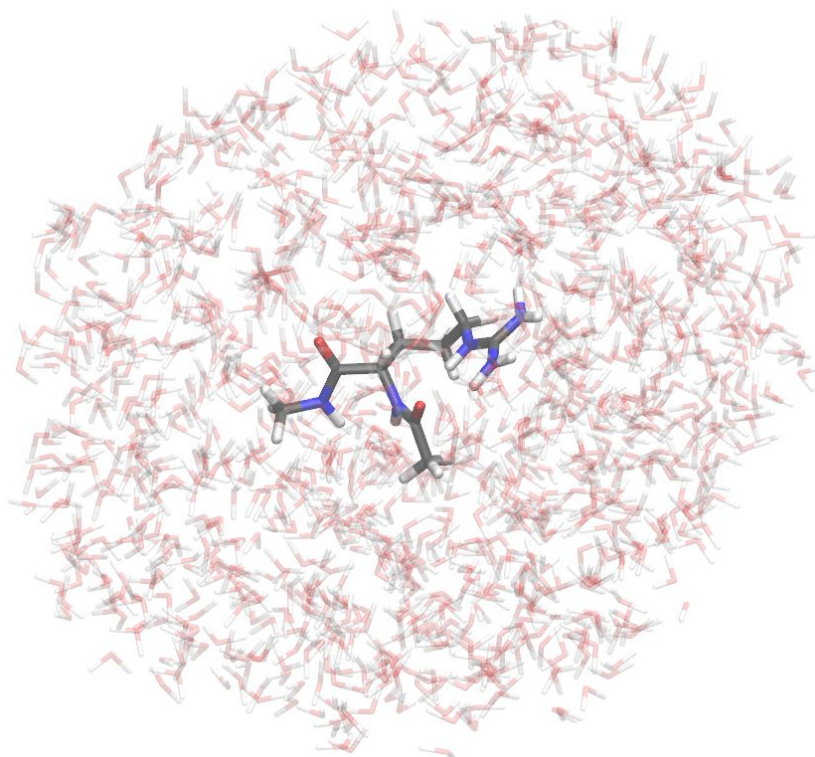
 New Charge Model
 Torsion
Refit
 Lennard-Jones Refit
 CMAP Inclusion



 Cerutti


 Debiec


 Bogetti

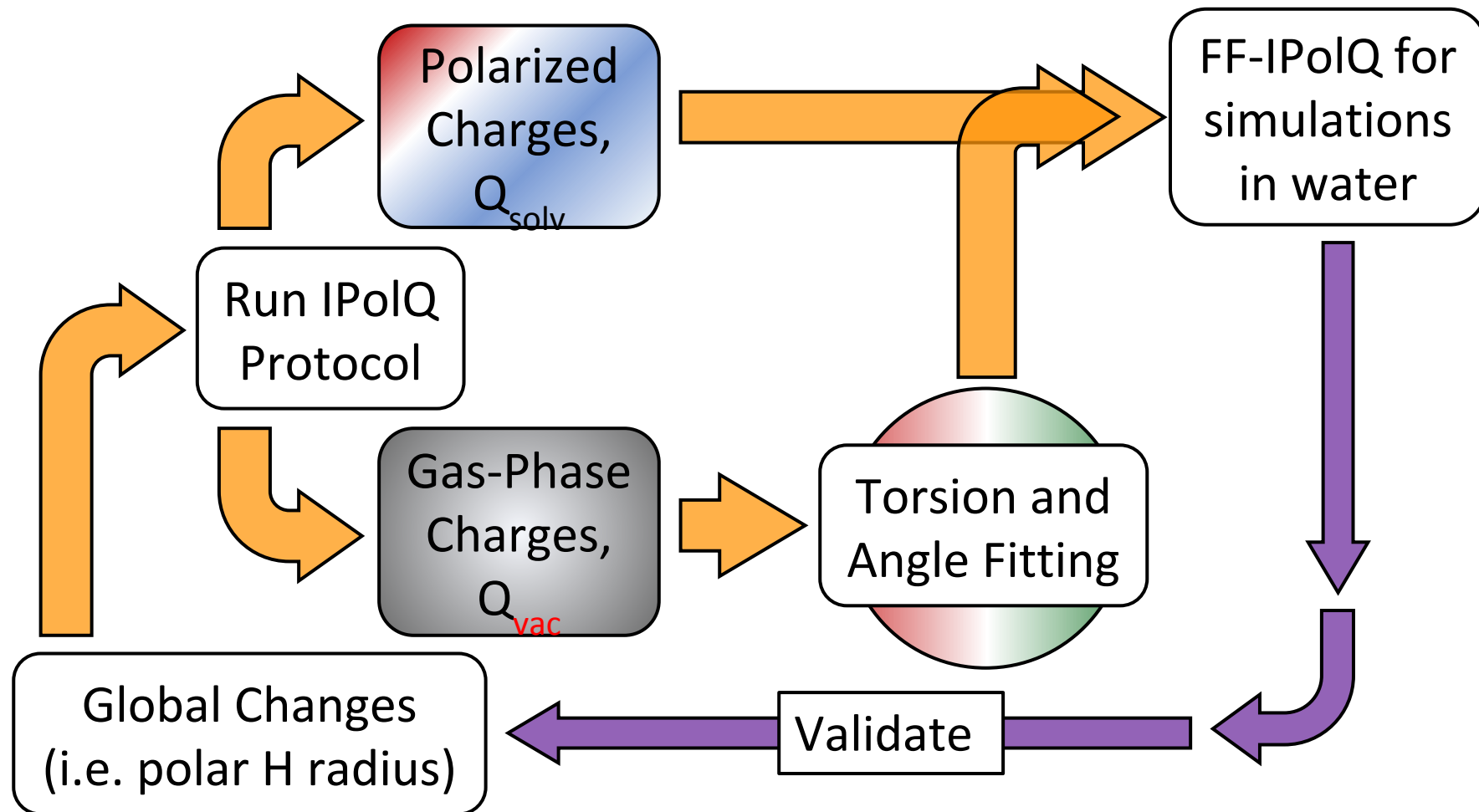
The IPolQ Charge Model and Related Force Field

- The target electrostatic potential is an average of MP2 / cc-pvTZ calculations:
 - The molecular conformation in vacuum, and...
 - In a reaction field due to a bath of (now, SPC-E/b) water
- Two charge sets emerge: one for simulations in water, the other for fitting parameters with gas phase data..



Integrating IPolQ Charges with Bonded Terms

- The central challenge: deriving angle and torsion parameters with gas-phase quantum energies for use with polarized charge distributions.



The ff15ipq Force Field

ff14ipq

IPolQ Charge Set
Feb. 28th, 2013[†]

Basic Least-Squares
Torsion Fitting

ff13 α : 28,000dp
Jan. 2014

Coupled Charge
Derivation

Generational
Optimization

ff14ipq: 65,000dp
Sept. 18th, 2014[†]

Increase Polar H
Radius to 1.5Å,
Remove Pair-
Specific Lennard
Jones Terms

Preliminary ff15ipq

ff15ipq, V1
June 4th, 2015

Angle Refinements
N-C α -C, C α -C-N, C-N-C α

Reduce Polar H
Radius to 1.3Å

ff15ipq, V2
June 21st, 2015

Angle Refinements
C α -N-H, C α -C=O

Replace TIP4P-Ew
with SPC/Eb

ff15ipq, V3
July 29th, 2015

Additional Angle
Refinements

ff15ipq, V4
Sept. 12th, 2015

ff15ipq

The Benefits of Angle Optimization

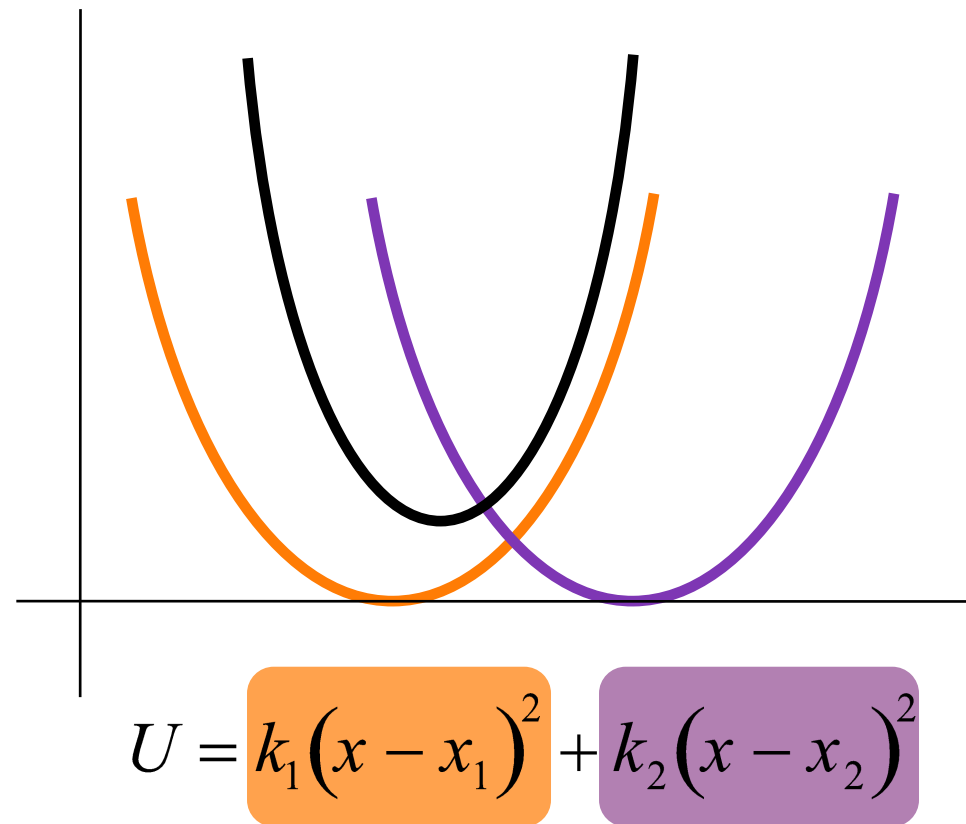
- Ala(5) J-couplings: a concise backbone diagnostic

Model	Original	DFT-1	DFT-2	K.L. Larsen
ff14ipq	1.3*	2.6	1.5	1.4
ff15ipq-V1	1.5	2.5	1.5	1.5
ff15ipq-V2	0.7	2.0	0.8	0.7
ff15ipq-V3	0.5	2.7	1.0	0.6
ff15ipq	0.5	2.8	1.1	0.7

*Mean χ^2 values are known to within 0.1Hz^2 or less

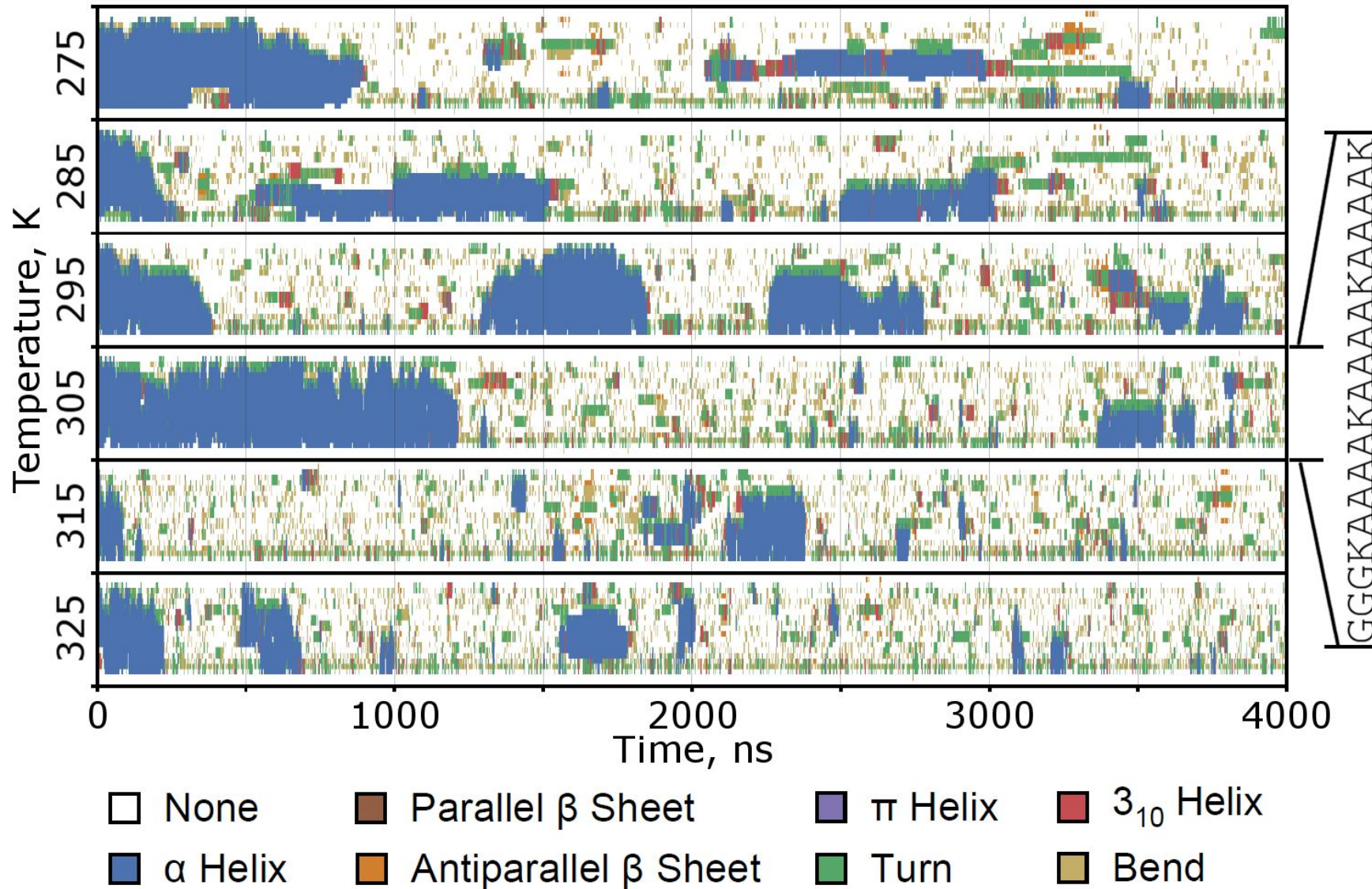
- ff15ipq covers the ff14ipq training set with equal or better accuracy and can predict the energies of new, strained conformations.
- Angle fitting appears to improve secondary structure stability.

Simultaneous fitting of both equilibria and force constants in harmonic terms:



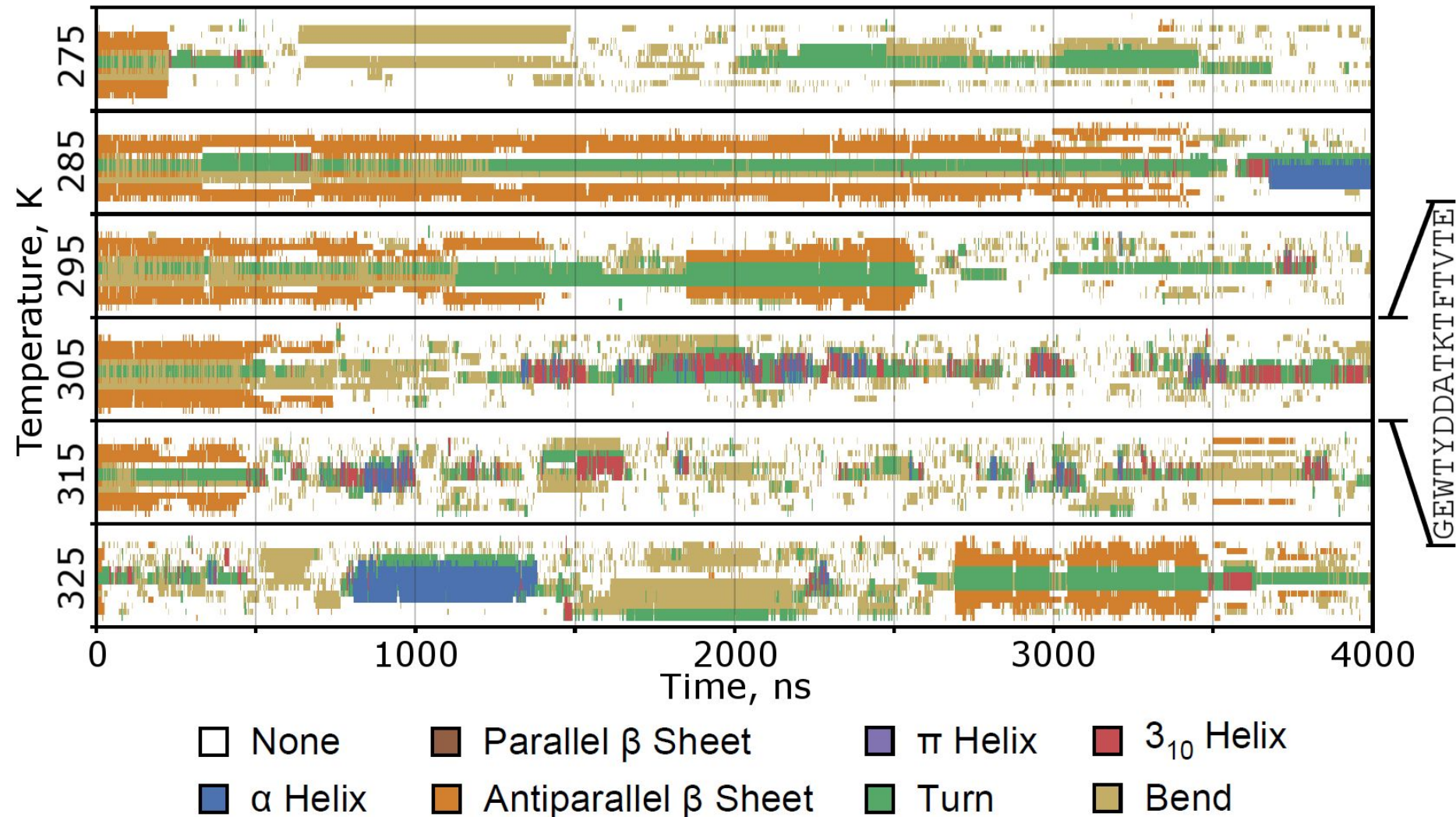
α -Helical Propensity in K19

- The **helix** is marginally less stable than the 40% target at 277K.



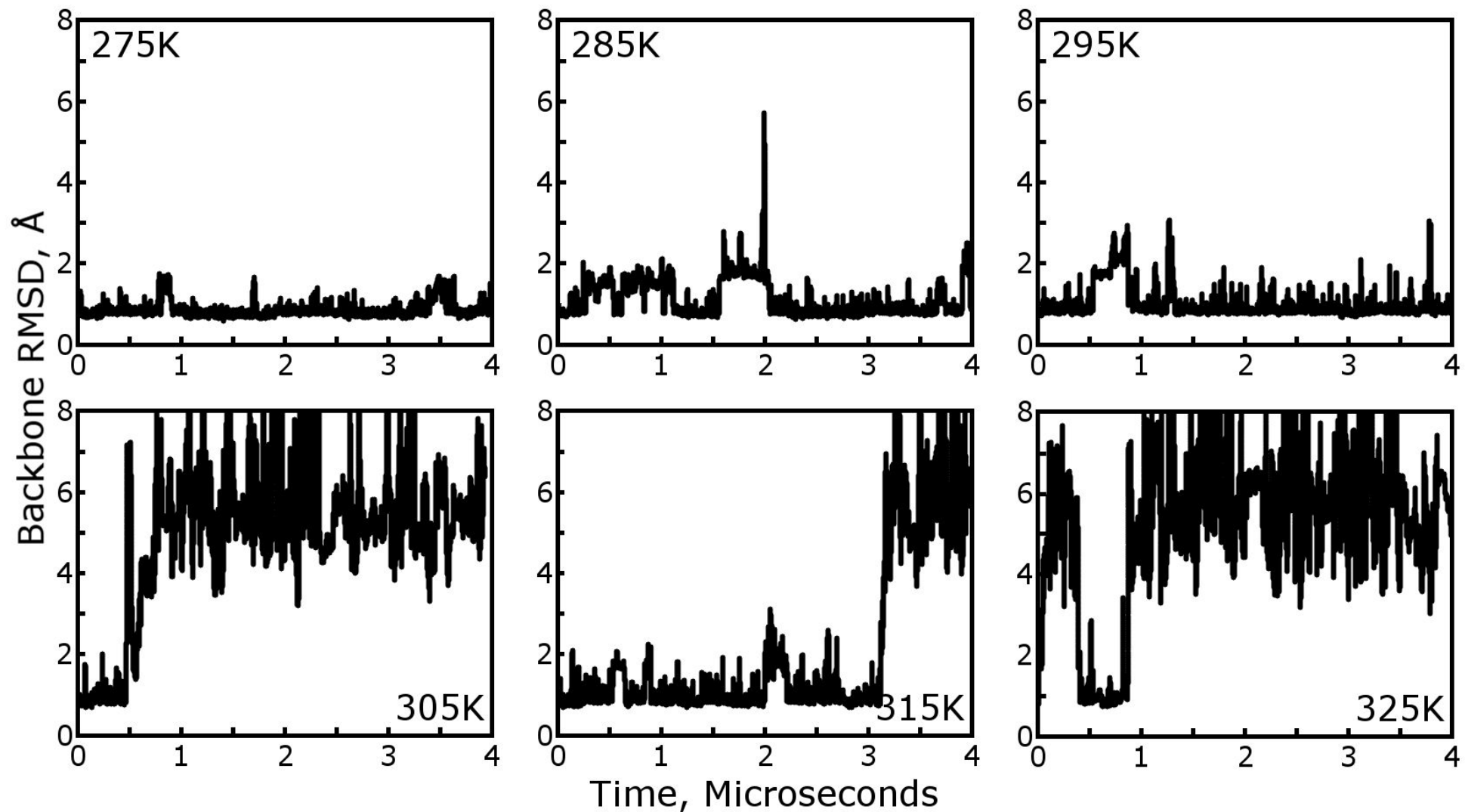
β -Sheet Propensity in GB1 Hairpin

- The **hairpin** is expected to be 50% folded at 295K. Convergence requires enhanced sampling methods.



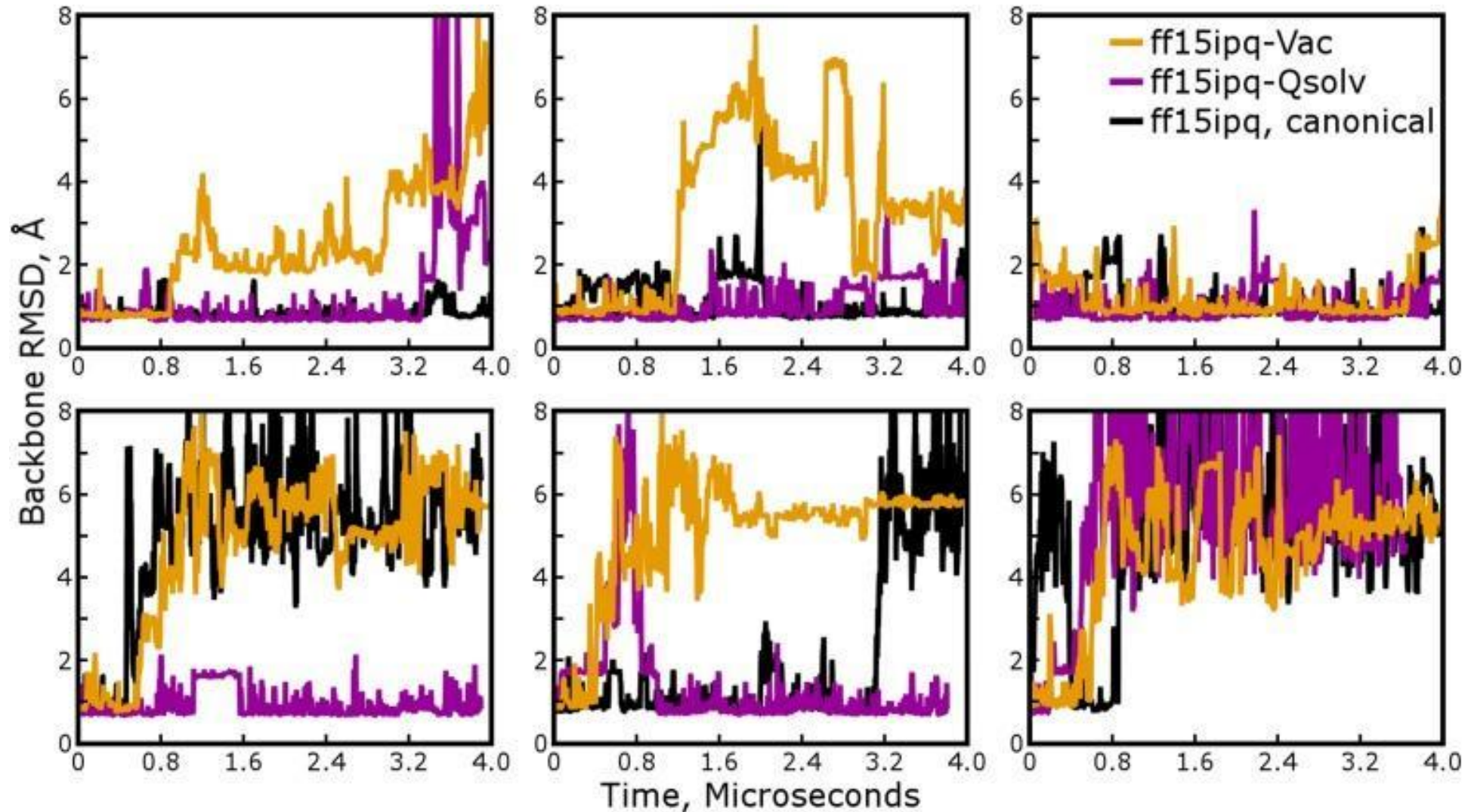
Trp-Cage Folding

- The **hairpin** is expected to be 50% folded at 295K. Convergence requires enhanced sampling methods.



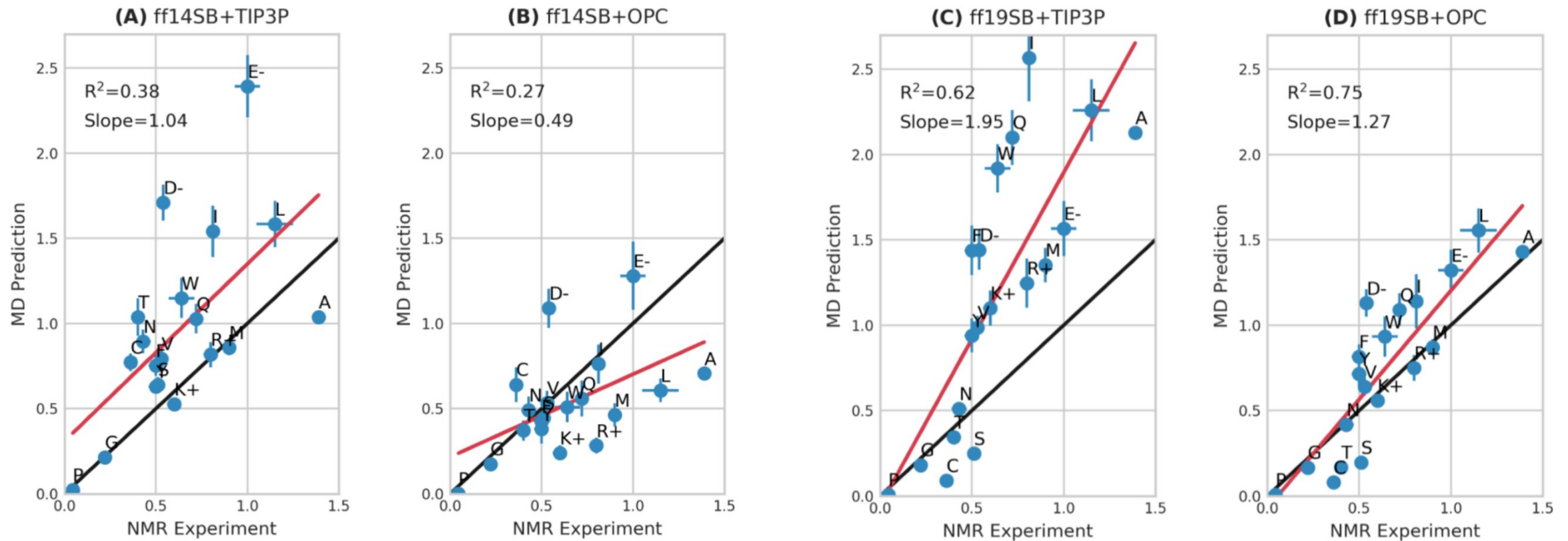
Trp-Cage Folding with Alternative Force Fields

- The unorthodox strategy behind ff15ipq proved to be better than alternatives, although the charge polarization itself had the largest effect in a battery of tests.



A CMAP-Based Force Field: ff19SB

- Following the logic of accounting for polarization effects in the bonded term fitting, Tian et al. fitted CMAP and torsion potentials to DFT calculations in implicit solvent.

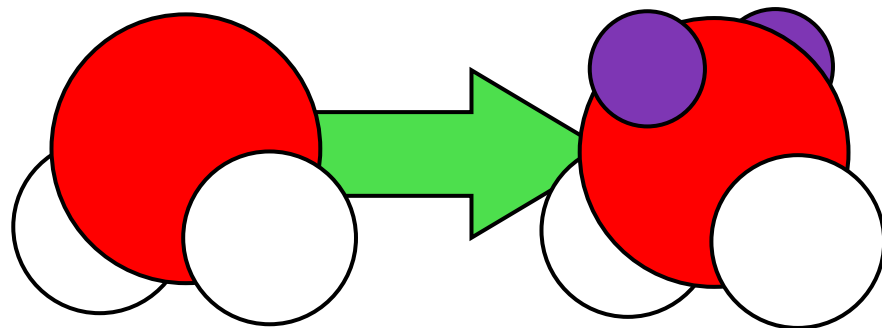


ff14SB Helical Propensities by Residue

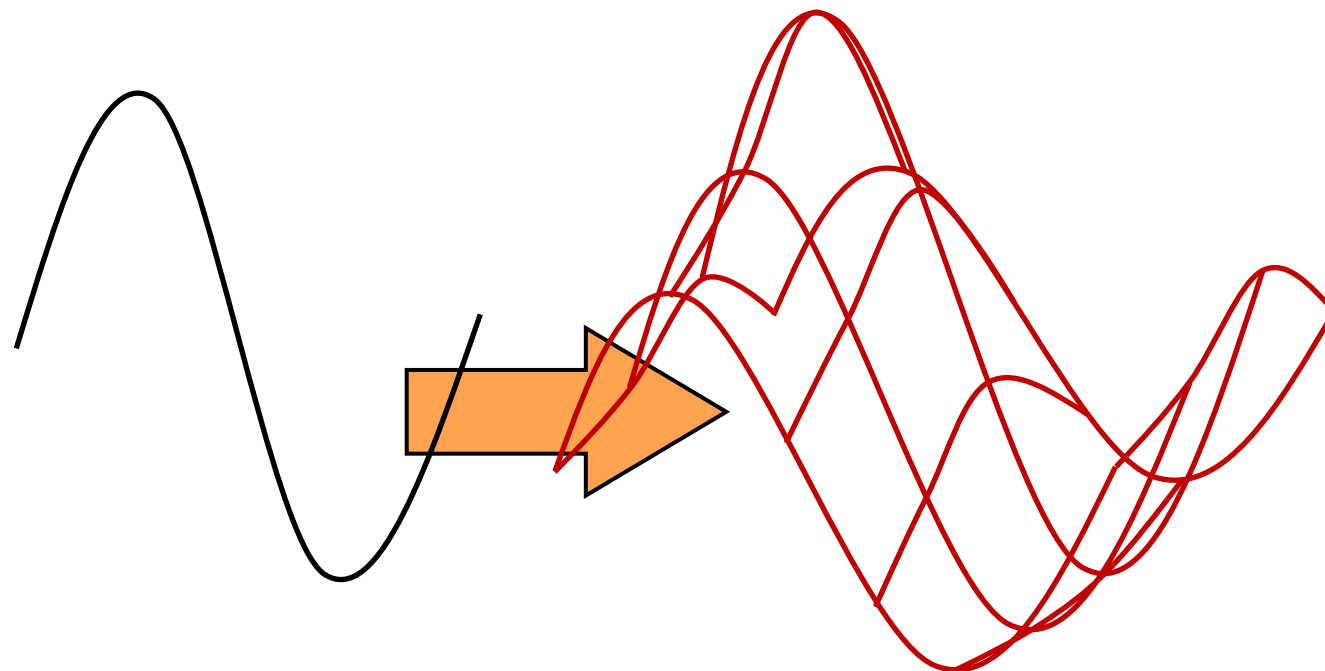
ff19SB Helical Propensities by Residue

Options for Improving Biopolymer Force Fields

- Elaborate on the complexity:



Additional Monoplexes

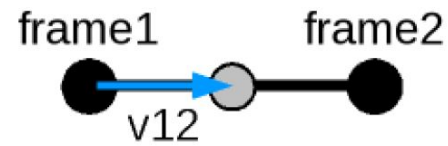


Tabulated potentials for cross-terms

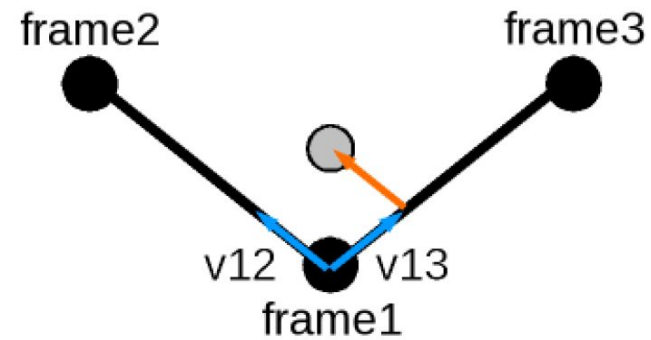
- Improve the fitting process:
 - Incorporate solvation effects in torsion drives
 - Mine additional, degenerate solutions for each parameter set fitted to QM data, pare them down with experimental data.

An Orthogonal Basis of Six Virtual Site Frames

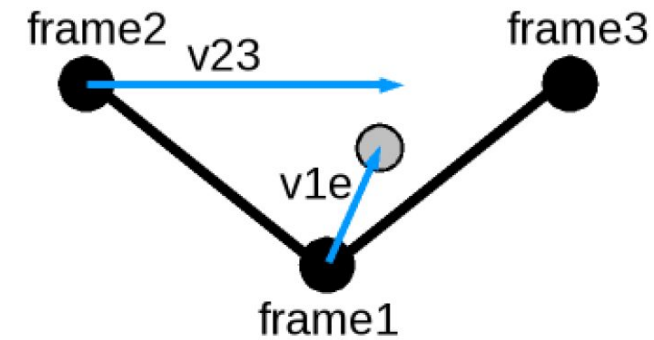
Style 1



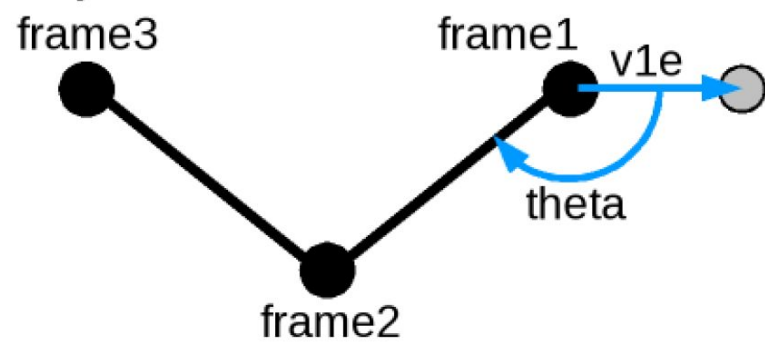
Style 2



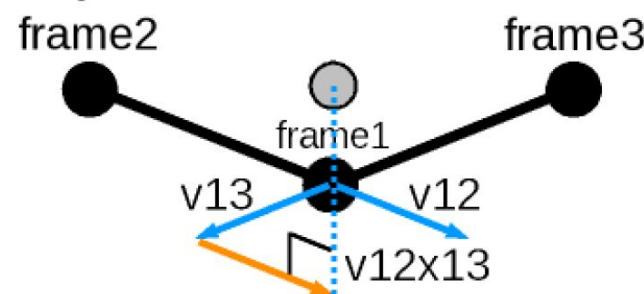
Style 3



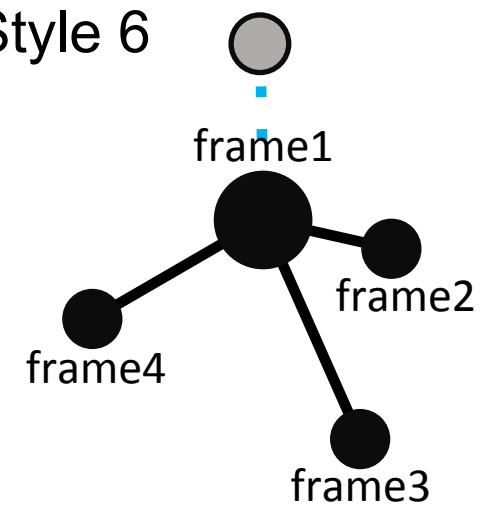
Style 4



Style 5

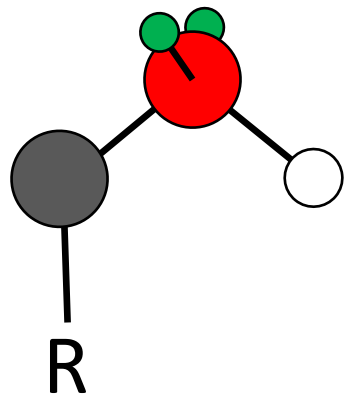


Style 6



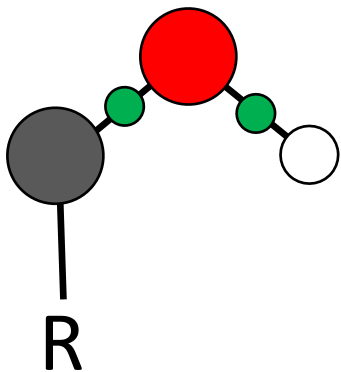
Freshman chemistry might not inform EP placement

- The Lewis structure lone pairs are not the best places to locate EPs, in any cases that I have yet examined. Take two simple side chains with lone pairs:

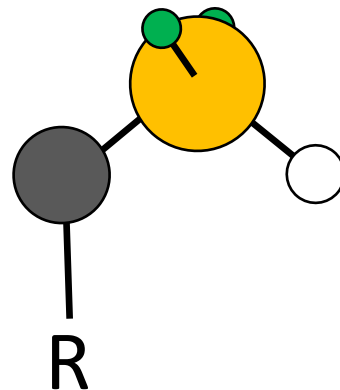


No effect on
electrostatic
potential fit

[Serine]

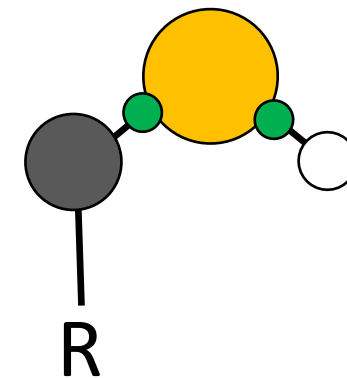


Profound
effect on
electrostatic
potential fit



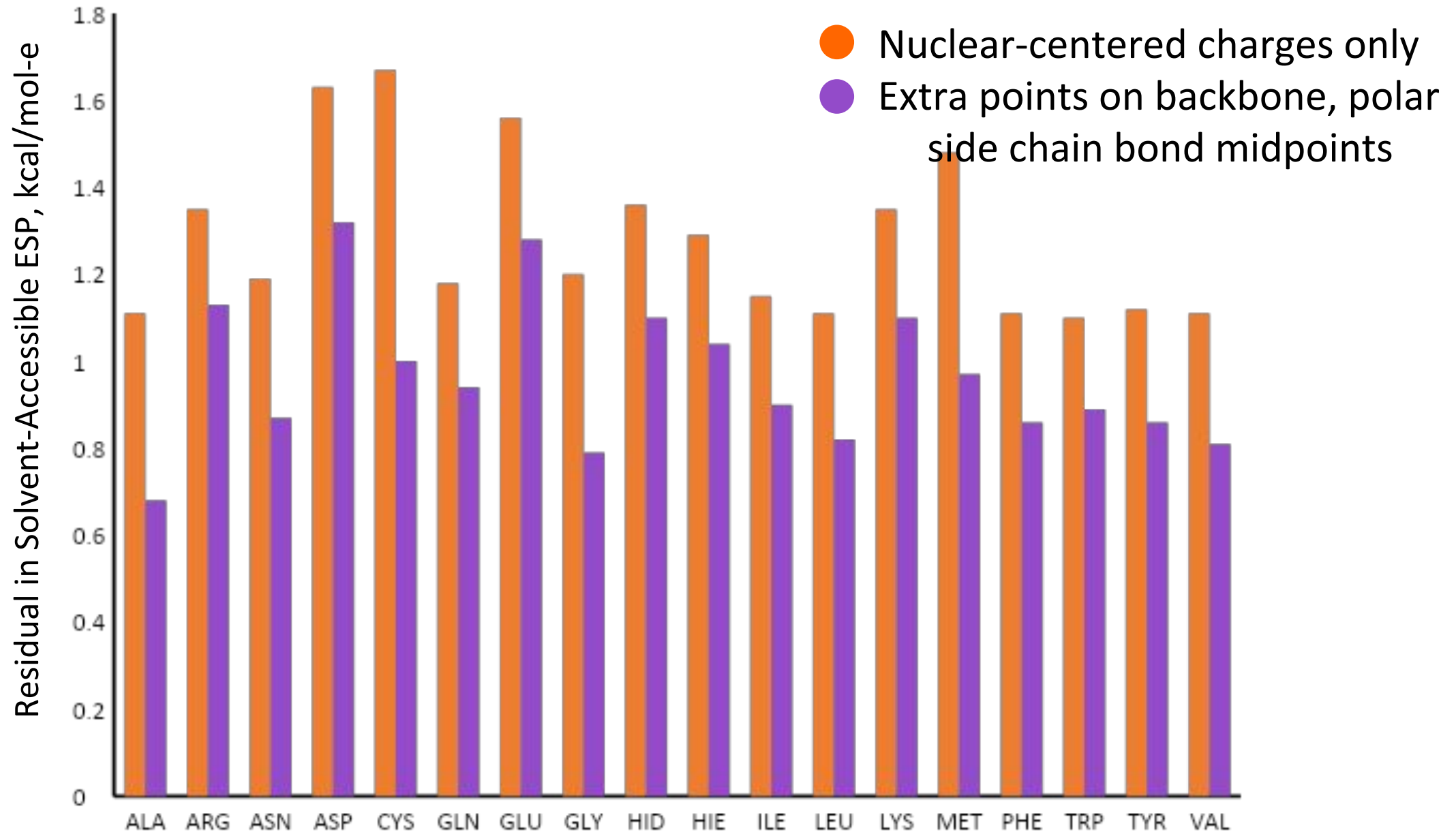
Beneficial to
electrostatic
potential fit

[Cysteine]



Still better!

Many EPs Can have a Moderate Effect on Accuracy

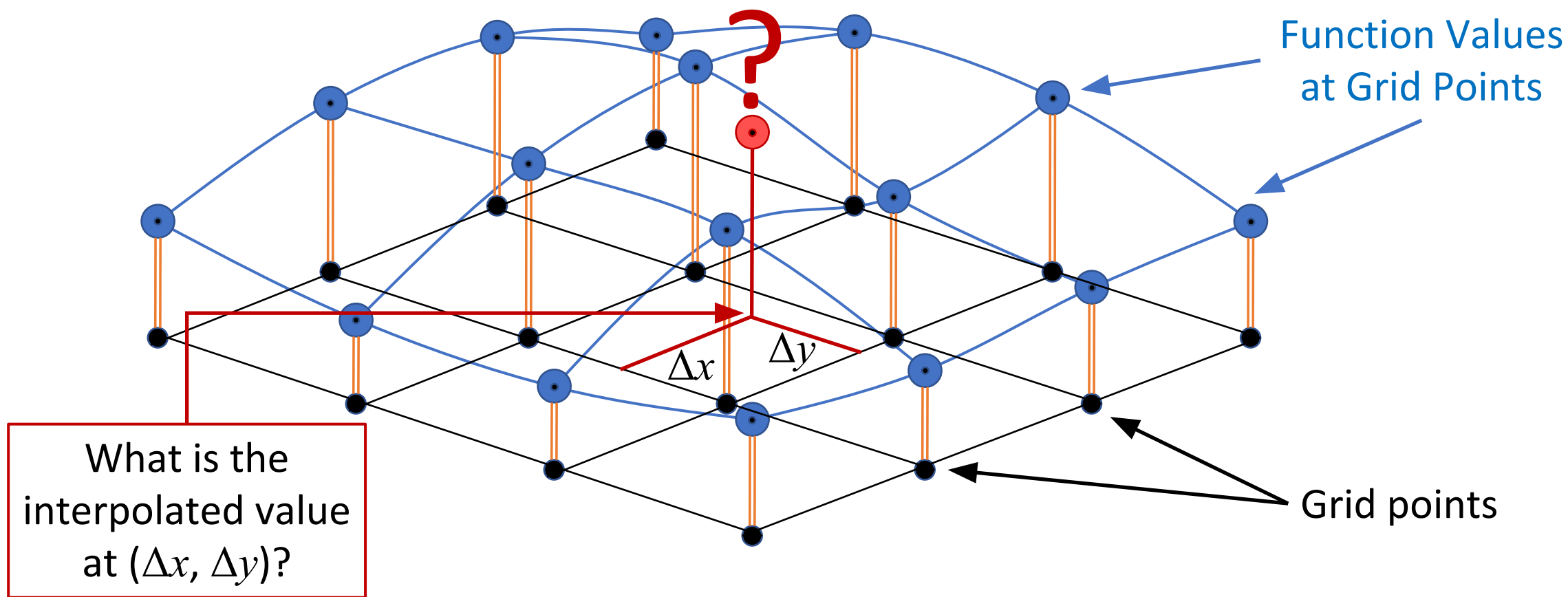


CMAP Fitting: Surfaces with Bicubic

Splines

The key is to recognize the grid points as the unique, independent variables.

- Seek a linear expression for everything else based on those values.



Bicubic Interpolation

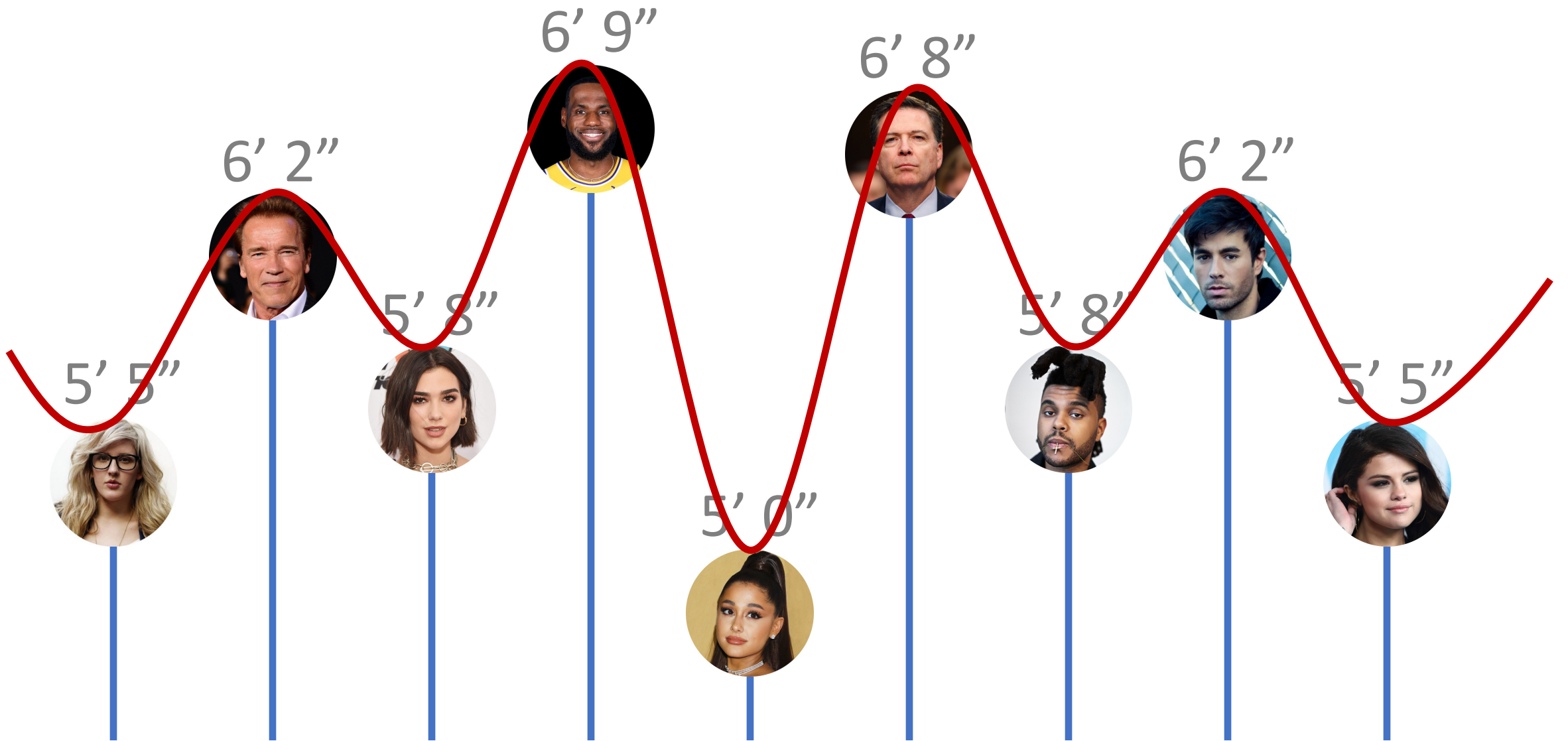
- The interpolant at any point within a grid segment of size $S_x \times S_y$ is given below.
- What is needed, then, is an expression for the derivatives in terms of grid values.

$$p(\Delta x, \Delta y) = \begin{bmatrix} 1 & \Delta x & (\Delta x)^2 & (\Delta x)^3 \end{bmatrix} \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 1 \\ \Delta y \\ (\Delta y)^2 \\ (\Delta y)^3 \end{bmatrix}, \text{ given } Q = \begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 0 & 3 & -2 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

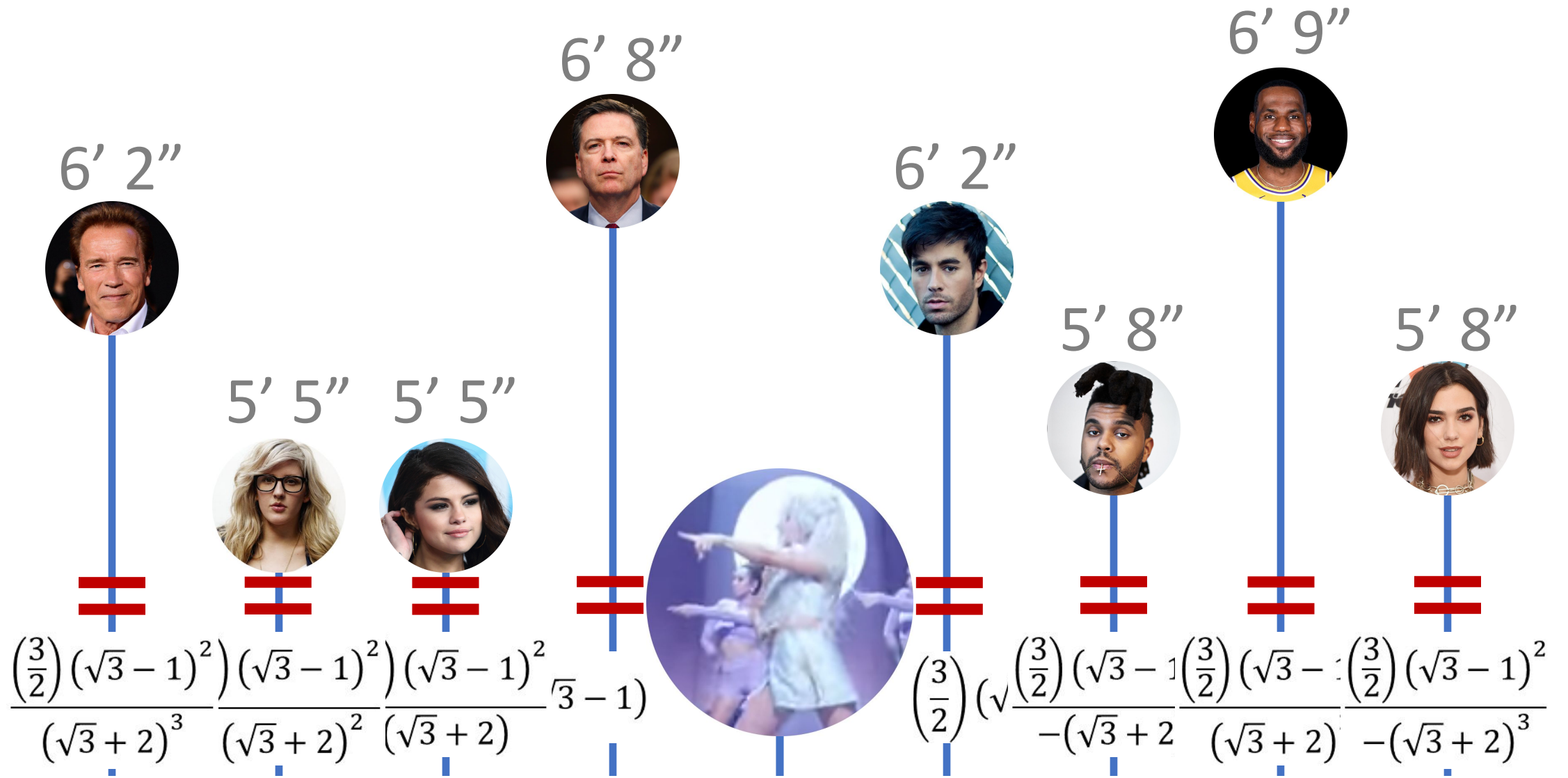
where

$$\begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{bmatrix} = Q^T \begin{bmatrix} f(0,0) & f(0,S_y) & \frac{\delta}{\delta y} f(0,0) & \frac{\delta}{\delta y} f(0,S_y) \\ f(S_x,0) & f(S_x,S_y) & \frac{\delta}{\delta y} f(S_x,0) & \frac{\delta}{\delta y} f(S_x,S_y) \\ \frac{\delta}{\delta x} f(0,0) & \frac{\delta}{\delta x} f(0,S_y) & \frac{\delta^2}{\delta x \delta y} f(0,0) & \frac{\delta^2}{\delta x \delta y} f(0,S_y) \\ \frac{\delta}{\delta x} f(S_x,0) & \frac{\delta}{\delta x} f(S_x,S_y) & \frac{\delta^2}{\delta x \delta y} f(S_x,0) & \frac{\delta^2}{\delta x \delta y} f(S_x,S_y) \end{bmatrix} Q$$

Derivatives of a Piecewise Cubic Spline



Derivatives of a Piecewise Cubic Spline



Bicubic Interpolation

- The derivative at any grid point is a weighted sum of the values at other grid points.

$$p(\Delta x, \Delta y) = \begin{bmatrix} 1 & \Delta x & (\Delta x)^2 & (\Delta x)^3 \end{bmatrix} \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 1 \\ \Delta y \\ (\Delta y)^2 \\ (\Delta y)^3 \end{bmatrix}, \text{ given } Q = \begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 0 & 3 & -2 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

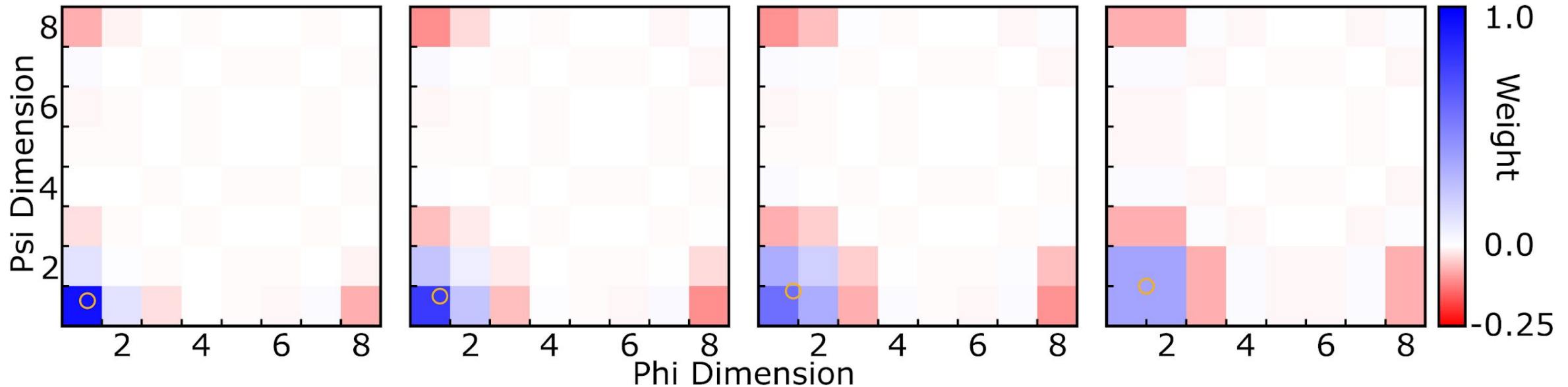
where

$$\begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{bmatrix} = Q^T \begin{bmatrix} f(0,0) & f(0,S_y) & \frac{\delta}{\delta y} f(0,0) & \frac{\delta}{\delta y} f(0,S_y) \\ f(S_x,0) & f(S_x,S_y) & \frac{\delta}{\delta y} f(S_x,0) & \frac{\delta}{\delta y} f(S_x,S_y) \\ \frac{\delta}{\delta x} f(0,0) & \frac{\delta}{\delta x} f(0,S_y) & \frac{\delta^2}{\delta x \delta y} f(0,0) & \frac{\delta^2}{\delta x \delta y} f(0,S_y) \\ \frac{\delta}{\delta x} f(S_x,0) & \frac{\delta}{\delta x} f(S_x,S_y) & \frac{\delta^2}{\delta x \delta y} f(S_x,0) & \frac{\delta^2}{\delta x \delta y} f(S_x,S_y) \end{bmatrix} Q$$

- The surface value anywhere is a linear combination of the values at the grid points!

Bicubic Interpolation

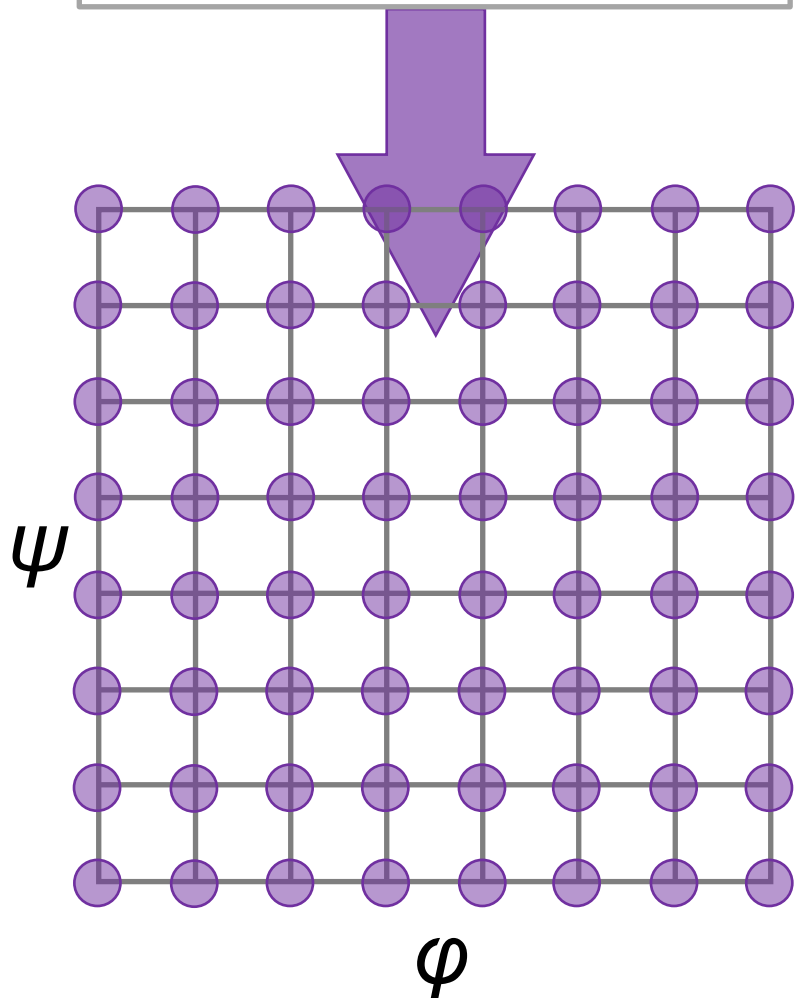
- In practice, the stencils for interpolated values have a generality to their form: high positive dependence on the nearest 1-4 points with weaker negative dependence on other near neighbors, a wavelet-like form decaying exponentially with distance.



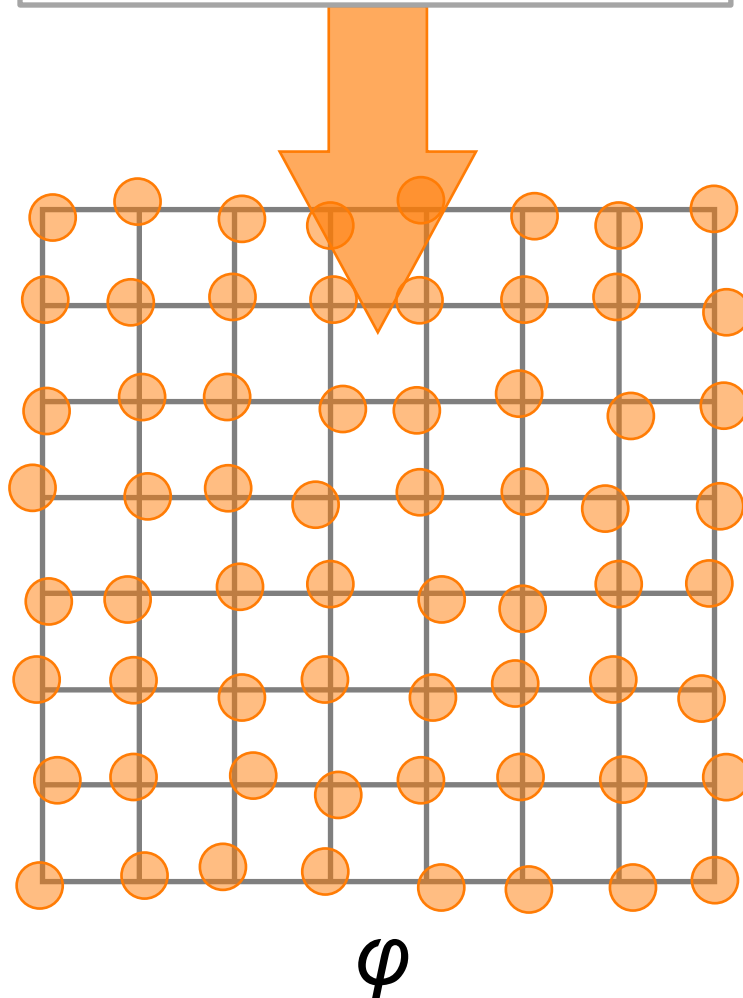
- These stencils let us construct a matrix equation with one independent variable for every grid point that can be populated with observations and the appropriate stencil values to solve a bicubic spline.

Training CMAPs and Tabulated Functional Forms

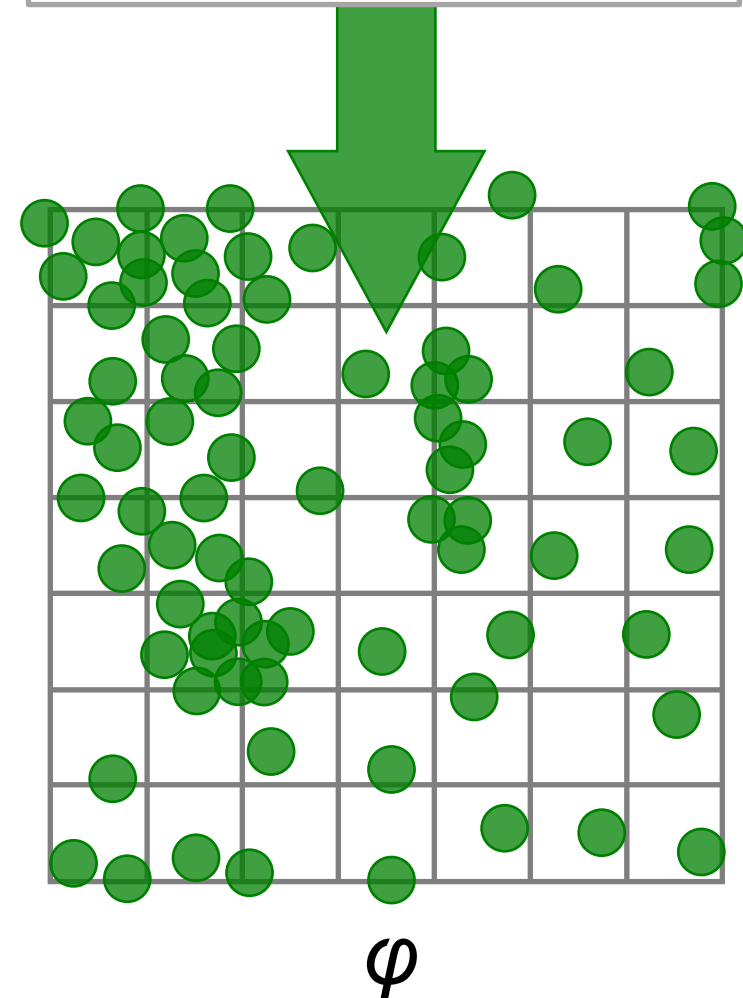
A naïve interpretation is to train the function with data on exact grid points.



However, molecules cannot be restrained to exact coordinates.

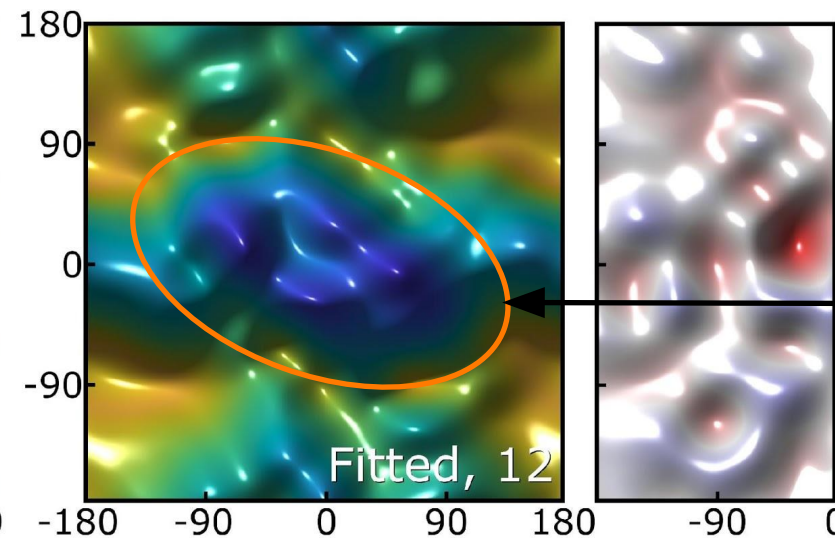
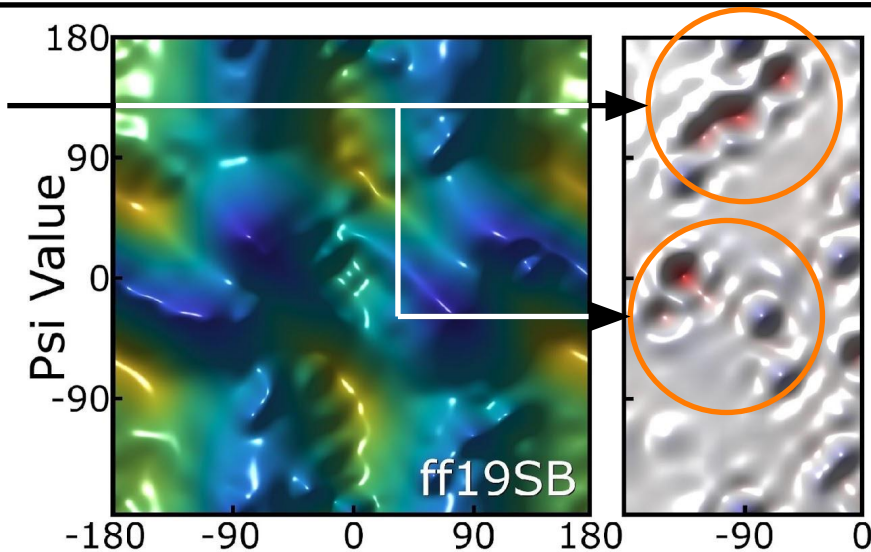


Math in the preceding slides: any data set with good coverage will do.

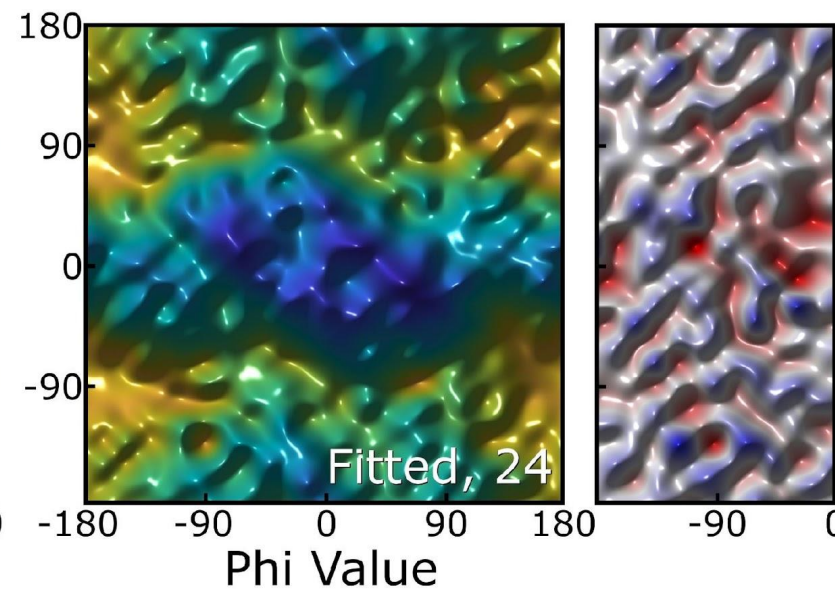
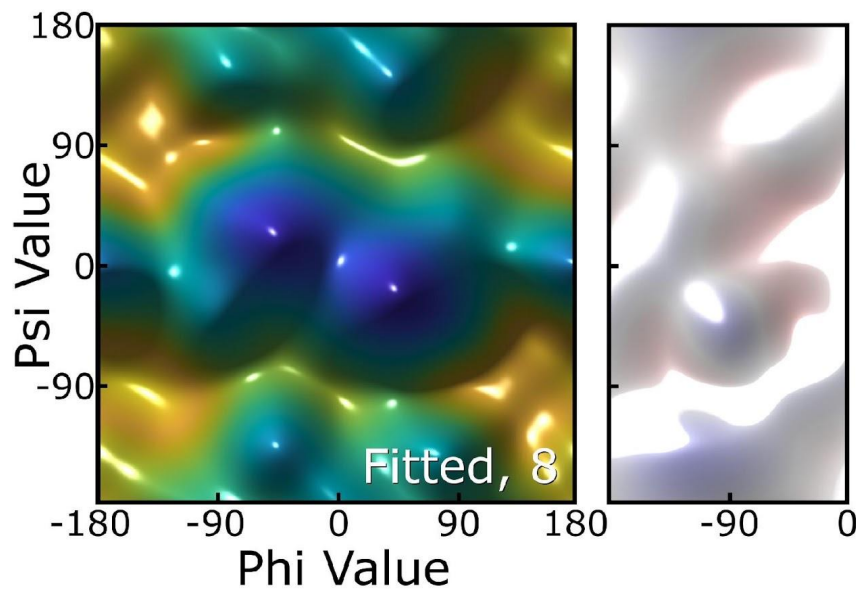


Glycine CMAPs

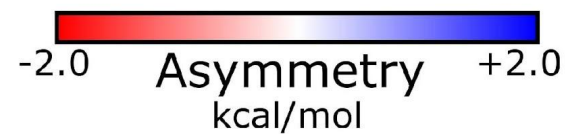
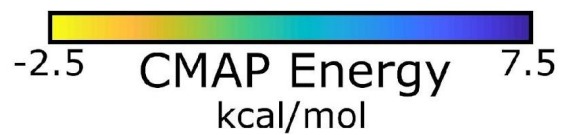
Symmetry
oversights in
ff19SB's map



High energy
region common
to all maps

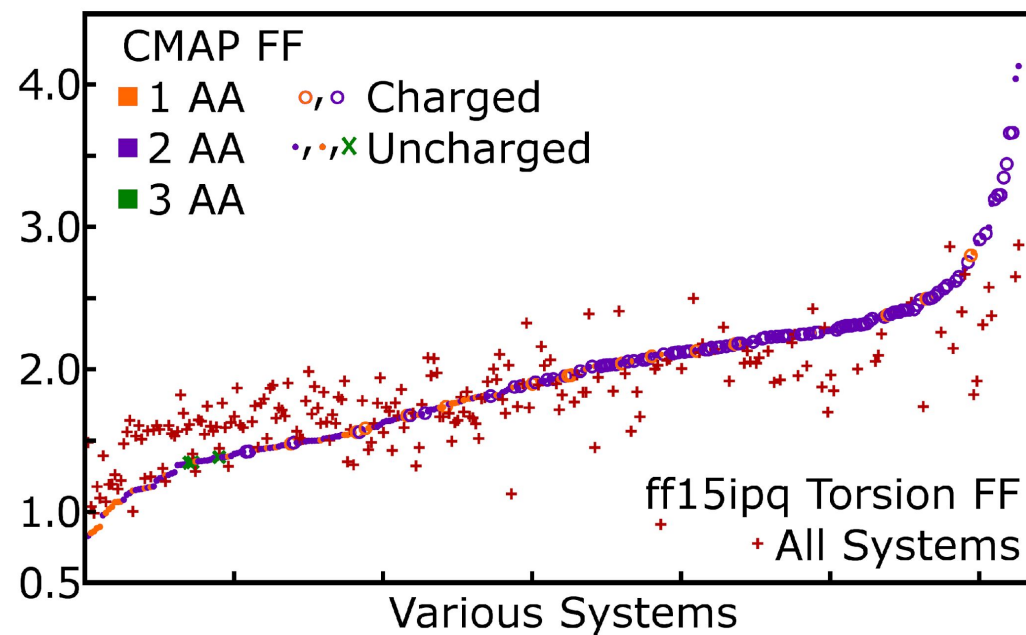
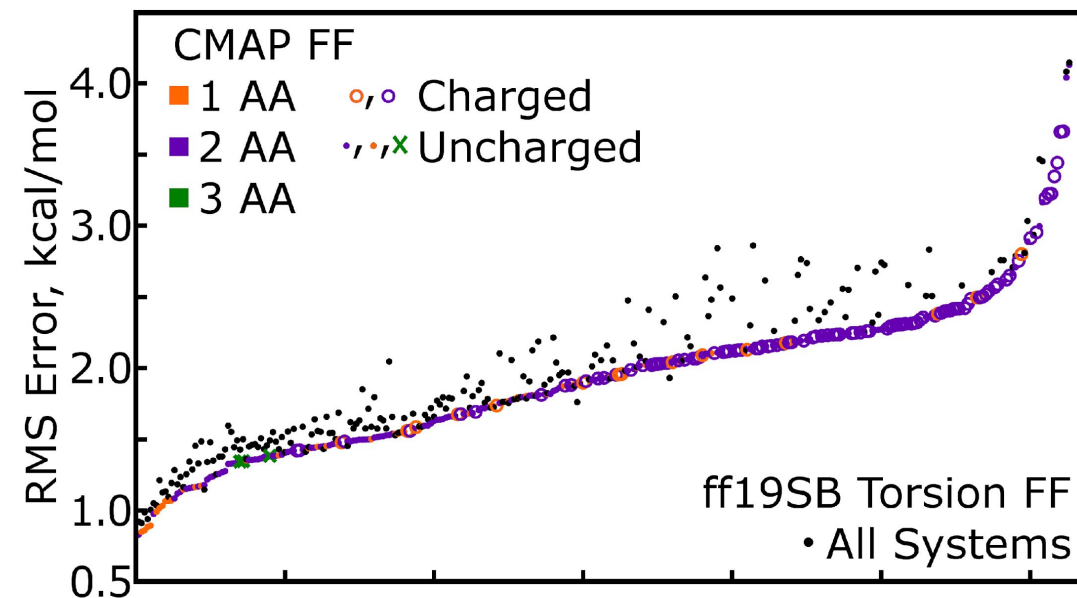


Even 6,500 data
points spread
throughout
populated regions
and some
high-energy
sampling do not
converge a
24-point map.



General Amino Acid CMAPs vs. Specific C_α Types

- When applied to amino acids in many systems, the benefits of the CMAPs are still clear, but dampened relative to individual amino acids.
- Residue-specific C_α typing, as in ff14SB and ff15ipq, may obtain better overall fits by over-fitting in poorly sampled, larger peptide structures.
- We need to understand the over-fitting problem we have, and the one we are stepping into.



Proposed Strategies in Biopolymer Force Field

- Common sets of (N, H, C, O) backbone charges for (+) charged, (-) charged, and all other amino acids, ESP fits challenged with approx. 64 conformations per residue. Additional charge sets for β 3 backbones and other common non-native residues.
 - Common sets of backbone torsion and angle parameters paired with charge sets for the above classifications, approx. 500 conformations per parameter.
 - B3LYP-(min. aug.) def2-tzvPP quantum calculations
 - Ace-Yaa-Xzz-Yaa-Nme tetrapeptides for charges, possibly bonded parameters
- Common sets of backbone charges and torsion parameters for DNA and RNA backbones, similar level of quantum theory to amino acids.
- For carbohydrates, assign common charges to each C, O, N, and H atom based on permutations of neighboring atoms and bonding structures in the ring. Torsion and angle parameters of the ring follow suit.

Future Directions for Force Field Development

- Ideally, our bonded parameters would be good enough to interpret raw quantum data without re-optimizing to relax molecular mechanics DoFs.
 - No ambiguity as to what coordinates should produce a particular energy
 - Force Balance runs much faster on big data sets
 - Train to gradients at particular atoms as well as the overall energy
- Bond :: bond angle CMAPs could achieve this level of accuracy
- Long-view, new interaction type: hydrogen bond corrections formulated as D-H :: A distance, D-H-A angle tables with cubic spline interpolation

Acknowledgement

- Rutgers, the State University of New Jersey
 - Professors David A. Case and Darrin York
 - Dr. Taisung Lee
- NVIDIA Corporation
 - Jon Lefman, Mark Berger
 - Peng Wang, Ke Li
 - Norbert Juffa (ret.)
- Ross Walker, GlaxoSmithKline
 - Charles Lin
 - Dan Mermelstein (UCSD)
- Scott Le Grand, Amazon

