Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement

Andrew Leaver-Fay*, ***Matthew J. O’Meara****, Mike Tyka, Ron Jacak, Yifan Song, Elizabeth H. Kellogg, James Thompson, Ian W. Davis, Roland A. Pache, Sergey Lyskov, Jeffrey J. Gray, Tanja Kortemme,, Jane S. Richardson, James J. Havranek, Jack Snoeyink, David Baker, Brian Kuhlman

Score Function Consensus

New Terms

- dun10
- orbitals
- H-patch
- cart_bonded/mm_*
- pH
- hackelec
- gb/pb
- geometric solvation
- env_dep reference weights

Re-parametrizations

- score12’
- -correct
- sp2 hbond
- lennard jones cutoff/radii
- softrep/hardrep
- P(aa|pp)
- pair

How to Demonstrate Improvement

- Explain problem/solution
- Run scientific benchmarks

Outline

- Two tools
  - Features Analysis
  - OptE
- Example Modification
  - Score12bicubic
  - Dun10
- Scientific Benchmarks

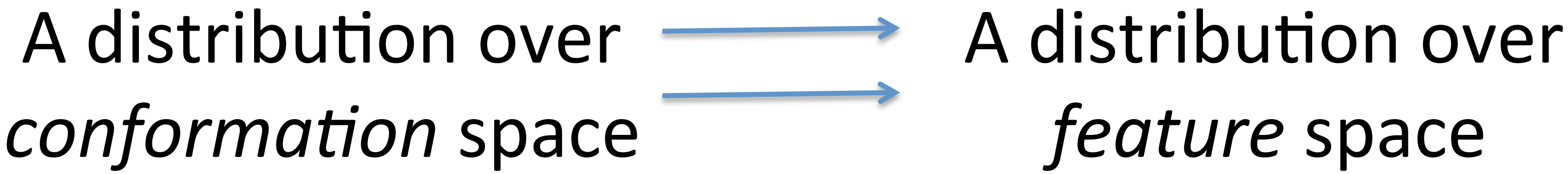
Boltzmann Distribution

$$P(C) = \frac{1}{Z} e^{\frac{-E(C)}{kT}}$$

What should P(C) be?

Feature Analysis

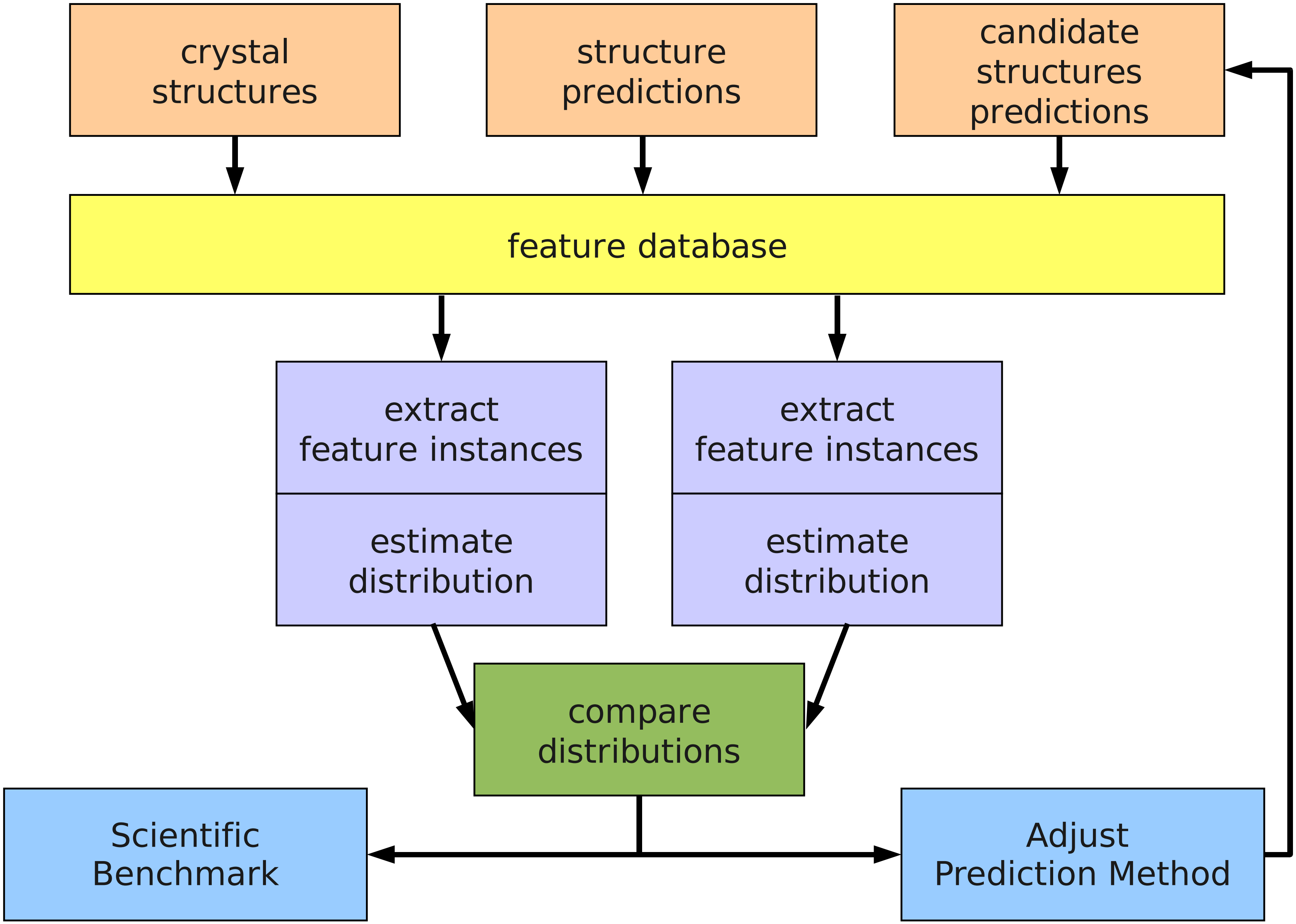
Feature: a geometric observable of a molecular conformation



A distribution over *conformation* space

A distribution over *feature* space

Feature-Based Workflows



crystal structures

structure predictions

candidate structures predictions

feature database

extract feature instances

estimate distribution

extract feature instances

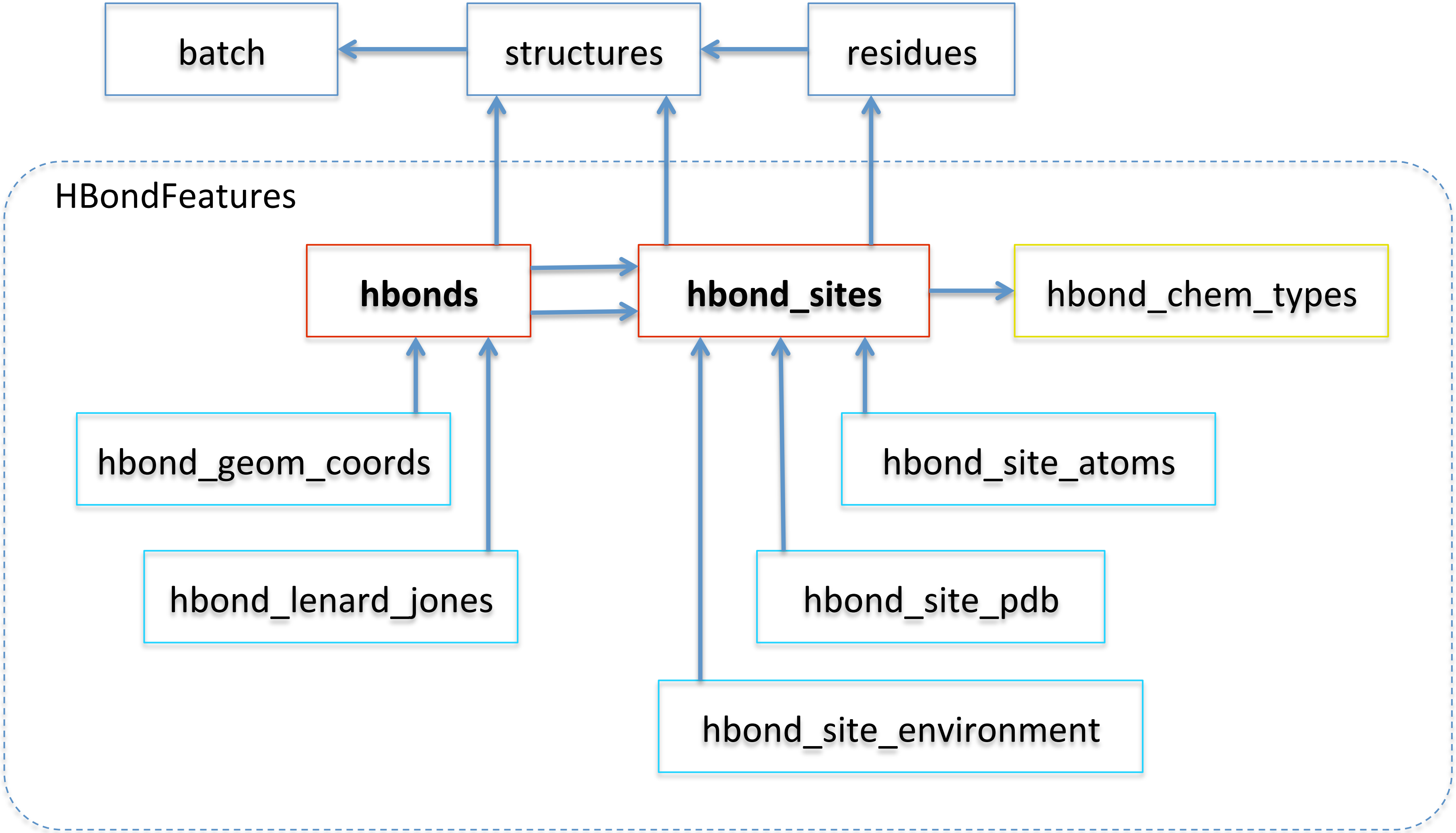
estimate distribution

compare distributions

Scientific Benchmark

Adjust Prediction Method

Hbond Features Reporter



Feature Reporters => “Schema Parts”

Meta

- Protocol
- Batch
- JobData
- PoseComments

Experimental Data

- PdbData
- PdbHeaderData
- DDG
- NMR
- DensityMap
- MultiSequenceAlignment
- HomologyAlignment

Chemical

- AtomType
- ResidueType

One Body

- Residue
- ResidueConformation
- ProteinResidueConformation
- ProteinBackboneTorsionAngle
- ResidueBurial
- ResidueSecondaryStructure
- GeometricSolvation
- BetaTurn
- RotamerBoltzmannWeight
- ResidueStrideSecondaryStructure
- HelixCapping
- BondGeometry
- ResidueLazaridisKarplusSolvation
- ResidueGeneralizedBornSolvation
- ResiduePoissonBoltzmannSolvation
- Pka
- ResidueCentroids

Two Body

- Pair
- AtomAtomPair
- AtomInResidueAtomInResiduePair
- ProteinBackboneAtomAtomPair
- HBond
- Orbital
- SaltBridge
- LoopAnchor
- DFIREPair
- ChargeCharge

Multi Structure

- ProteinRMSD
- ResidueRecovery
- ResiduePairRecovery
- ResidueClusterRecovery
- Cluster

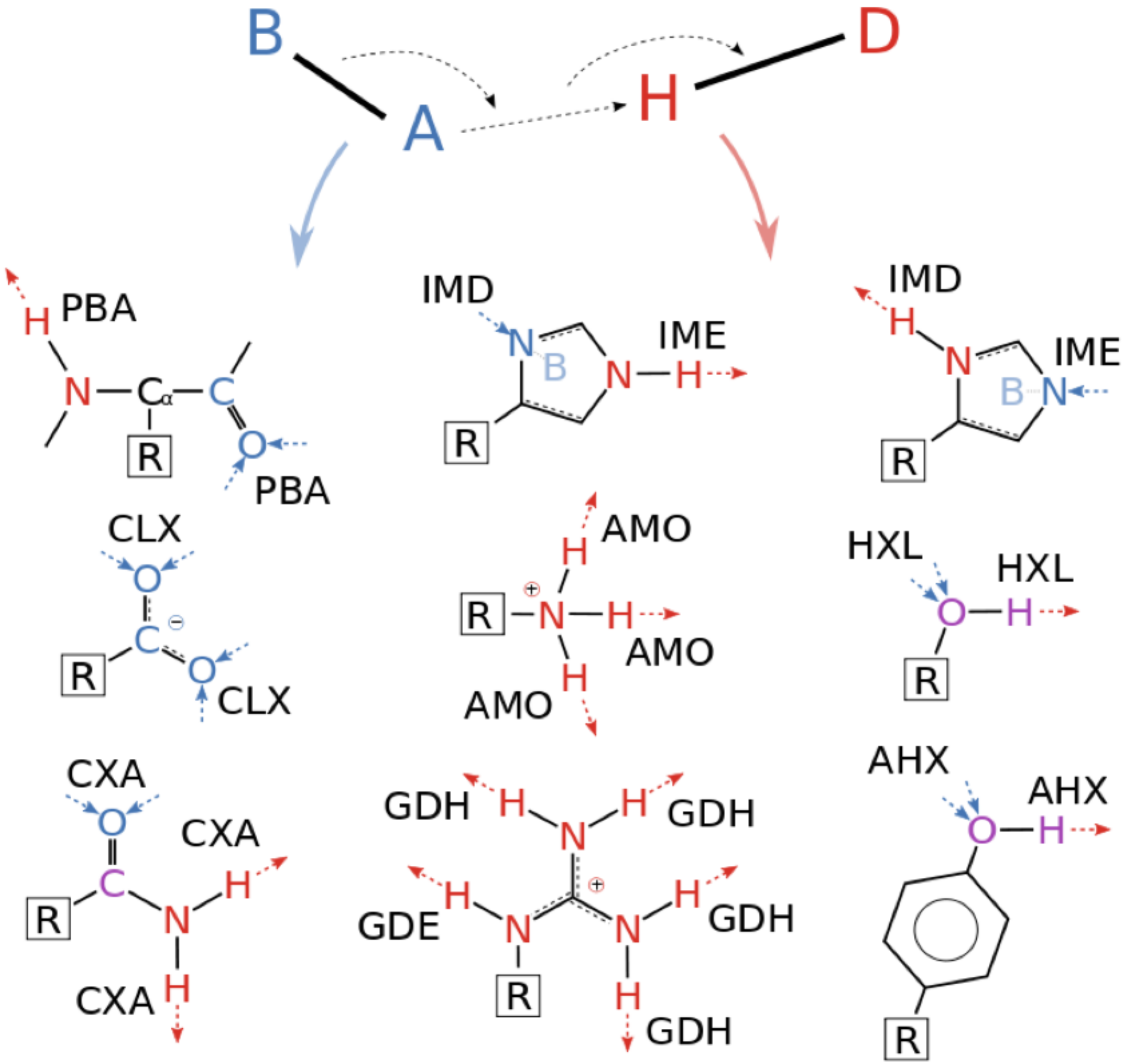
Multi Body

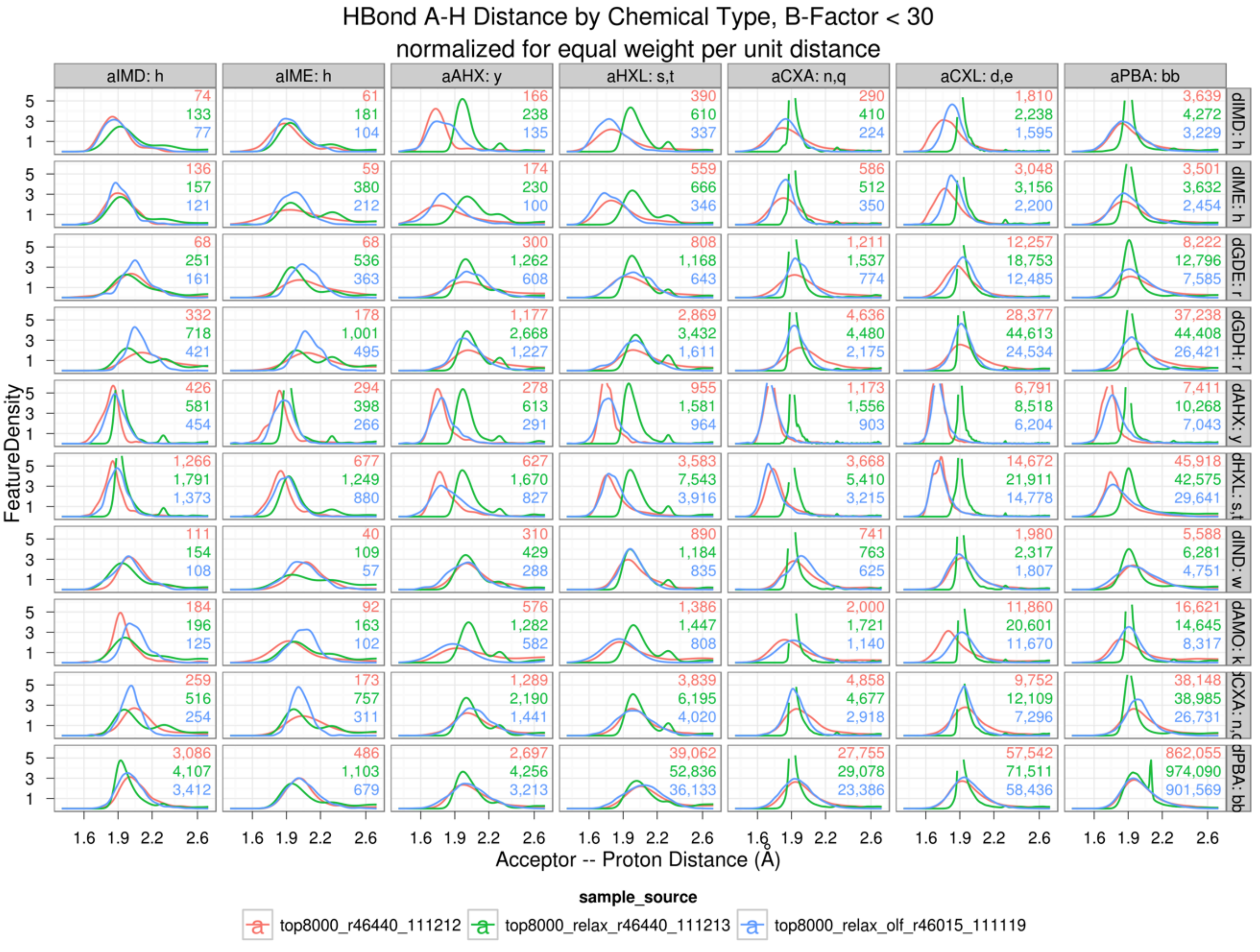
- Structure
- PoseConformation
- RadiusOfGyration
- SecondaryStructure
- HydrophbicPatch
- Cavity
- GraphMotif
- SequenceMotif
- Rigidity
- VoronoiPacking
- InterfaceAnalysis

Energy Function

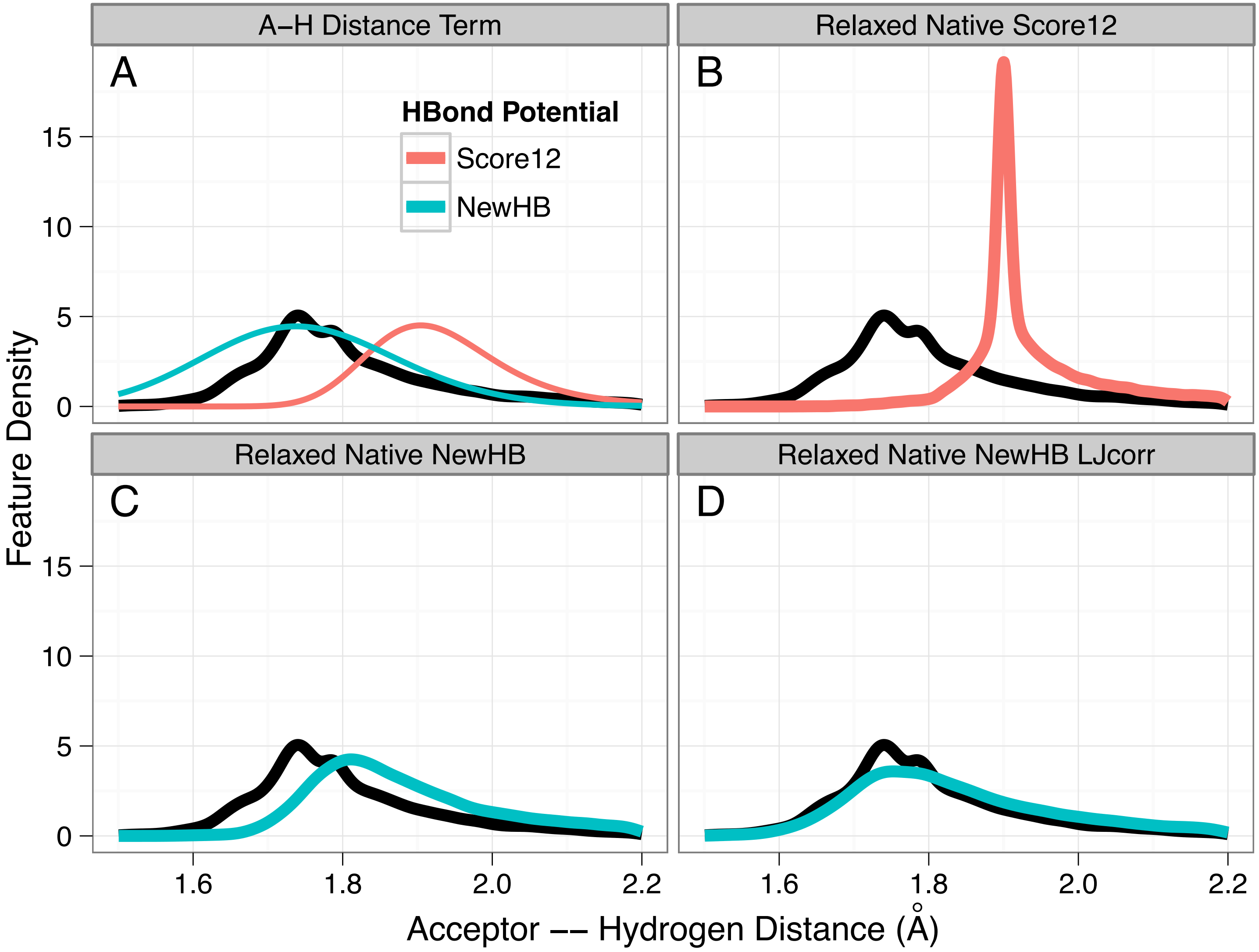
- ScoreFunction
- ScoreType
- StructureScores
- ResidueScores
- HBondParameters
- <EnergyTerm>Parameters

Refined Hydrogen Bond Model




HBond A-H Distance by Chemical Type, B-Factor < 30
normalized for equal weight per unit distance
aIMD: h
aIME: h
aAHX: y
aHXL: s,t
aCXA: n,q
aCXL: d,e
aPBA: bb
dIMD: h
dIME: h
dGDE: r
dGDH: r
dAHX: y
dHXL: s,t
dIND: w
dAMO: k
dCXA: n,q
dPBA: bb
74 133 77
61 181 104
166 238 135
390 610 337
290 410 224
1,810 2,238 1,595
3,639 4,272 3,229
136 157 121
59 380 212
174 230 100
559 666 346
586 512 350
3,048 3,156 2,200
3,501 3,632 2,454
68 251 161
68 536 363
300 1,262 608
808 1,168 643
1,211 1,537 774
12,257 18,753 12,485
8,222 12,796 7,585
332 718 421
178 1,001 495
1,177 2,668 1,227
2,869 3,432 1,611
4,636 4,480 2,175
28,377 44,613 24,534
37,238 44,408 26,421
426 581 454
294 398 266
278 613 291
955 1,581 964
1,173 1,556 903
6,791 8,518 6,204
7,411 10,268 7,043
1,266 1,791 1,373
677 1,249 880
627 1,670 827
3,583 7,543 3,916
3,668 5,410 3,215
14,672 21,911 14,778
45,918 42,575 29,641
111 154 108
40 109 57
310 429 288
890 1,184 835
741 763 625
1,980 2,317 1,807
5,588 6,281 4,751
184 196 125
92 163 102
576 1,282 582
1,386 1,447 808
2,000 1,721 1,140
11,860 20,601 11,670
16,621 14,645 8,317
259 516 254
173 757 311
1,289 2,190 1,441
3,839 6,195 4,020
4,858 4,677 2,918
9,752 12,109 7,296
38,148 38,985 26,731
3,086 4,107 3,412
486 1,103 679
2,697 4,256 3,213
39,062 52,836 36,133
27,755 29,078 23,386
57,542 71,511 58,436
862,055 974,090 901,569
FeatureDensity
1
3
5
1.6
1.9
2.2
2.6
Acceptor -- Proton Distance (Å)
sample_source
top8000_r46440_111212
top8000_relax_r46440_111213
top8000_relax_olf_r46015_111119

HBond Potential



OptE Overview

- Encode Scientific Benchmarks as a "Loss Function"
  - Seq.-Profile Recovery, Rot. Recovery, ddG, etc.

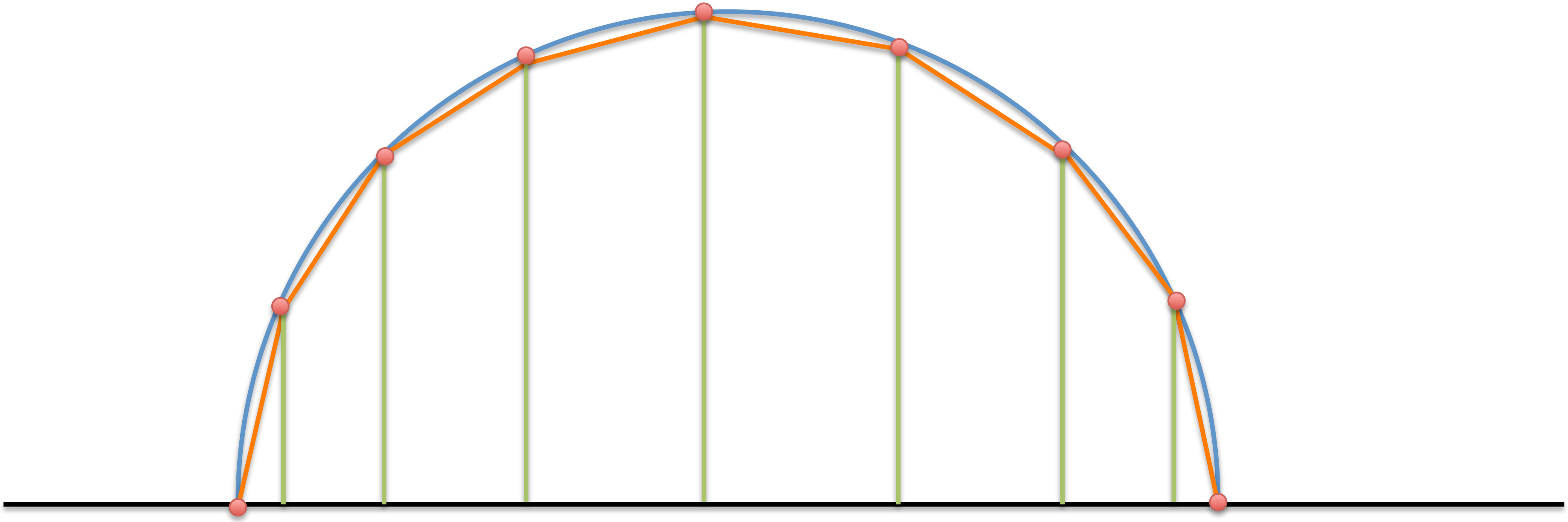
- Optimize weights to minimize loss
  - Estimate Partition Function
  - Optimize Weights
  - Repeat

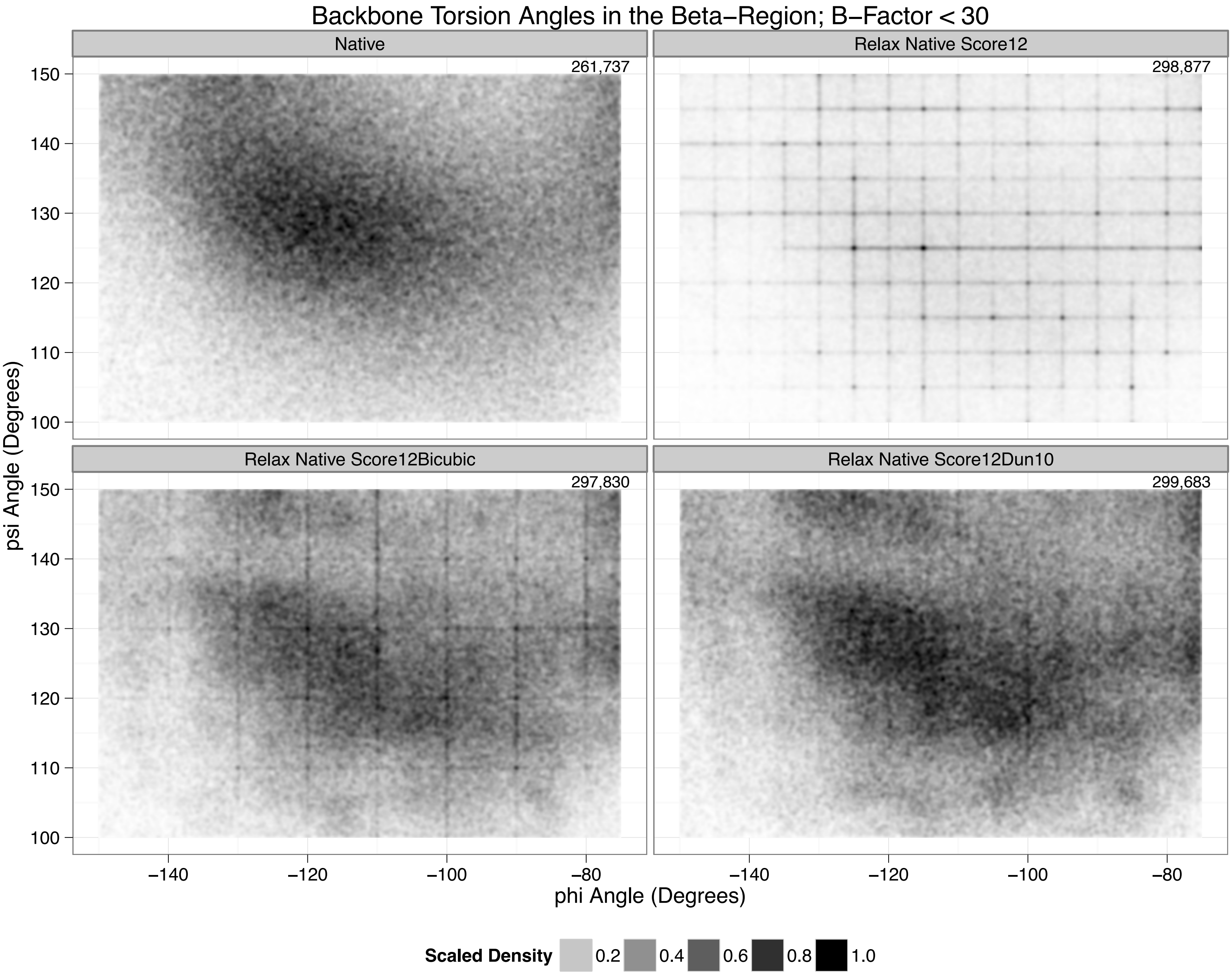
- Jim, Andrew, Ron, Liz, Yifan, Mike, James, Ian, more…

OptE Capabilities

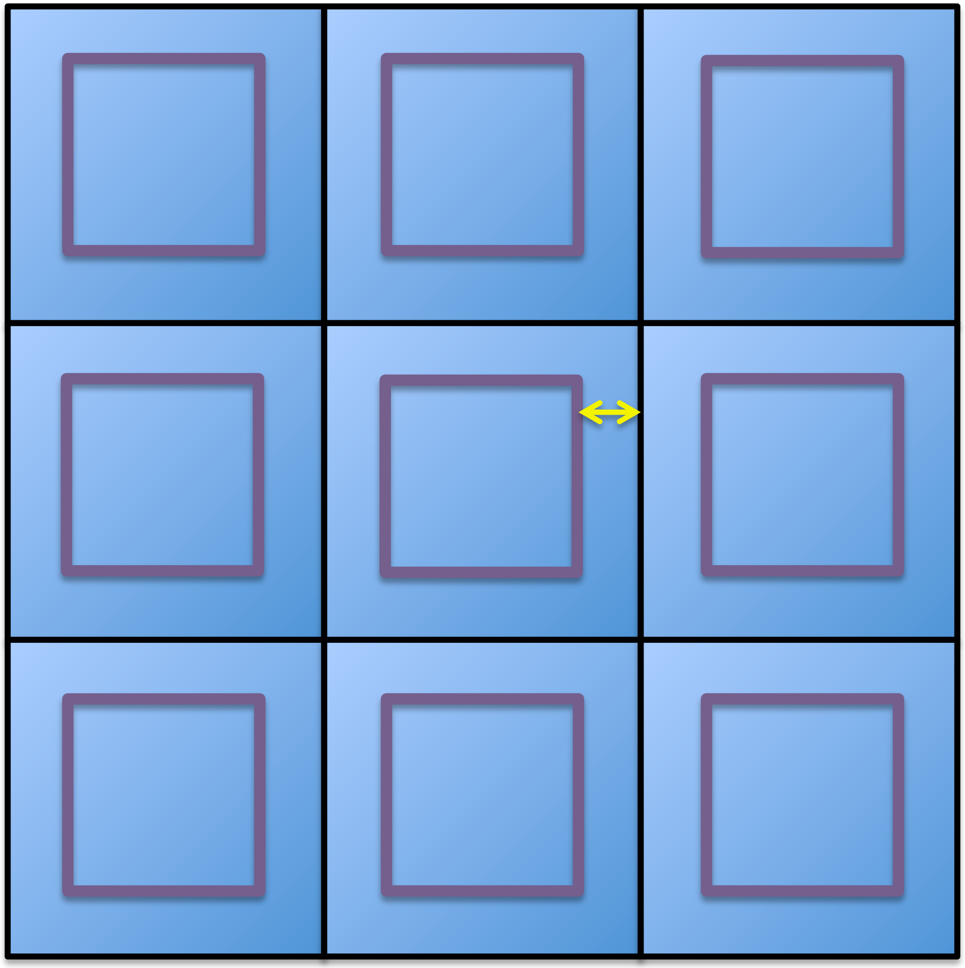
- Re-fit reference weights
  - Sequence-profile-recovery
- Test targeted hypotheses with a few weights
  - Hackeleck
  - CH-bond potential

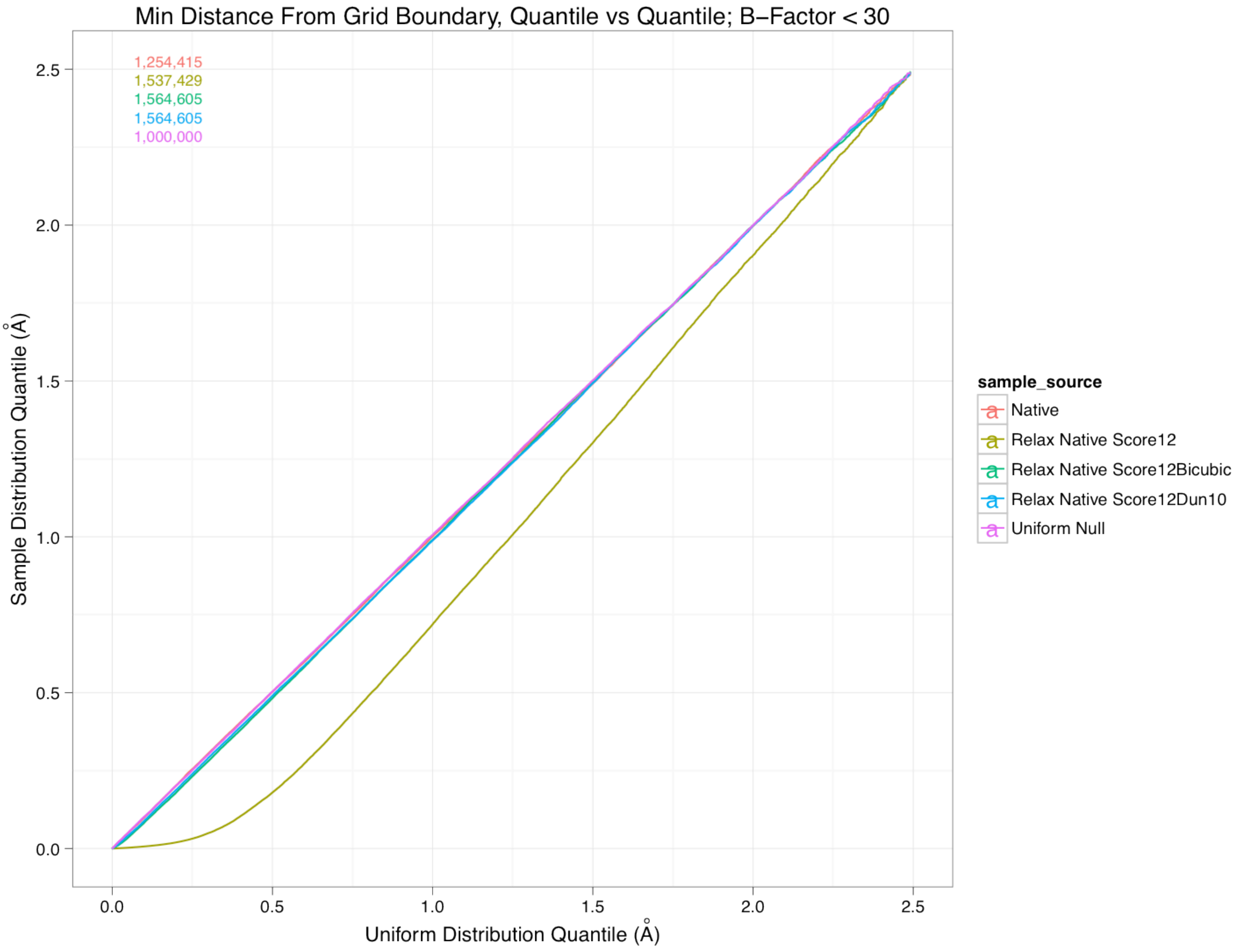
Linear vs Spline Interpolation




Backbone Torsion Angles in the Beta−Region; B−Factor < 30
Native
261,737
Relax Native Score12
298,877
Relax Native Score12Bicubic
297,830
Relax Native Score12Dun10
299,683
psi Angle (Degrees)
150
140
130
120
110
100
phi Angle (Degrees)
−140
−120
−100
−80
Scaled Density
0.2
0.4
0.6
0.8
1.0

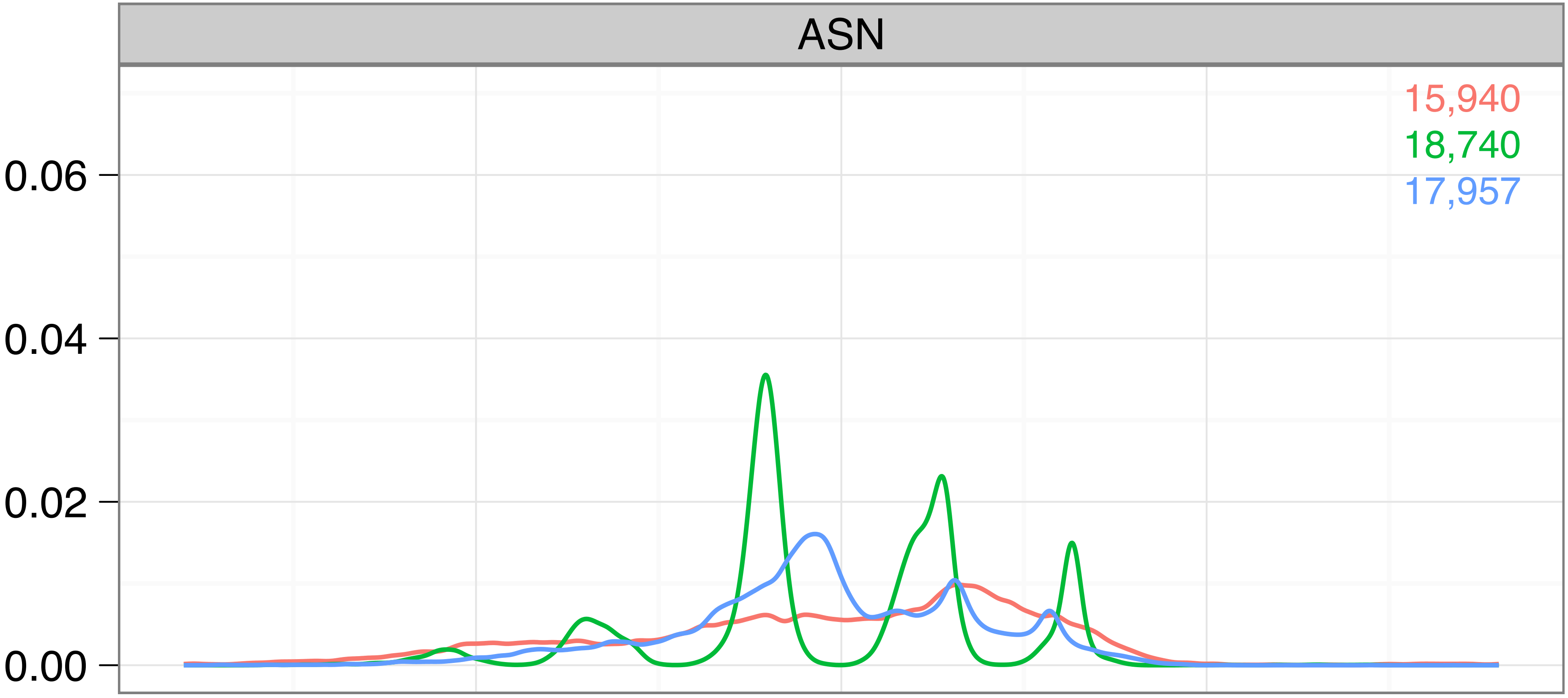
off_grid feature




Min Distance From Grid Boundary, Quantile vs Quantile; B−Factor < 30
1,254,415
1,537,429
1,564,605
1,564,605
1,000,000
Sample Distribution Quantile (Å)
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
Uniform Distribution Quantile (Å)
sample_source
Native
Relax Native Score12
Relax Native Score12Bicubic
Relax Native Score12Dun10
Uniform Null

Dun10: Semi Rotameric

$P(X_2 \mid X_1 = \text{trans})$



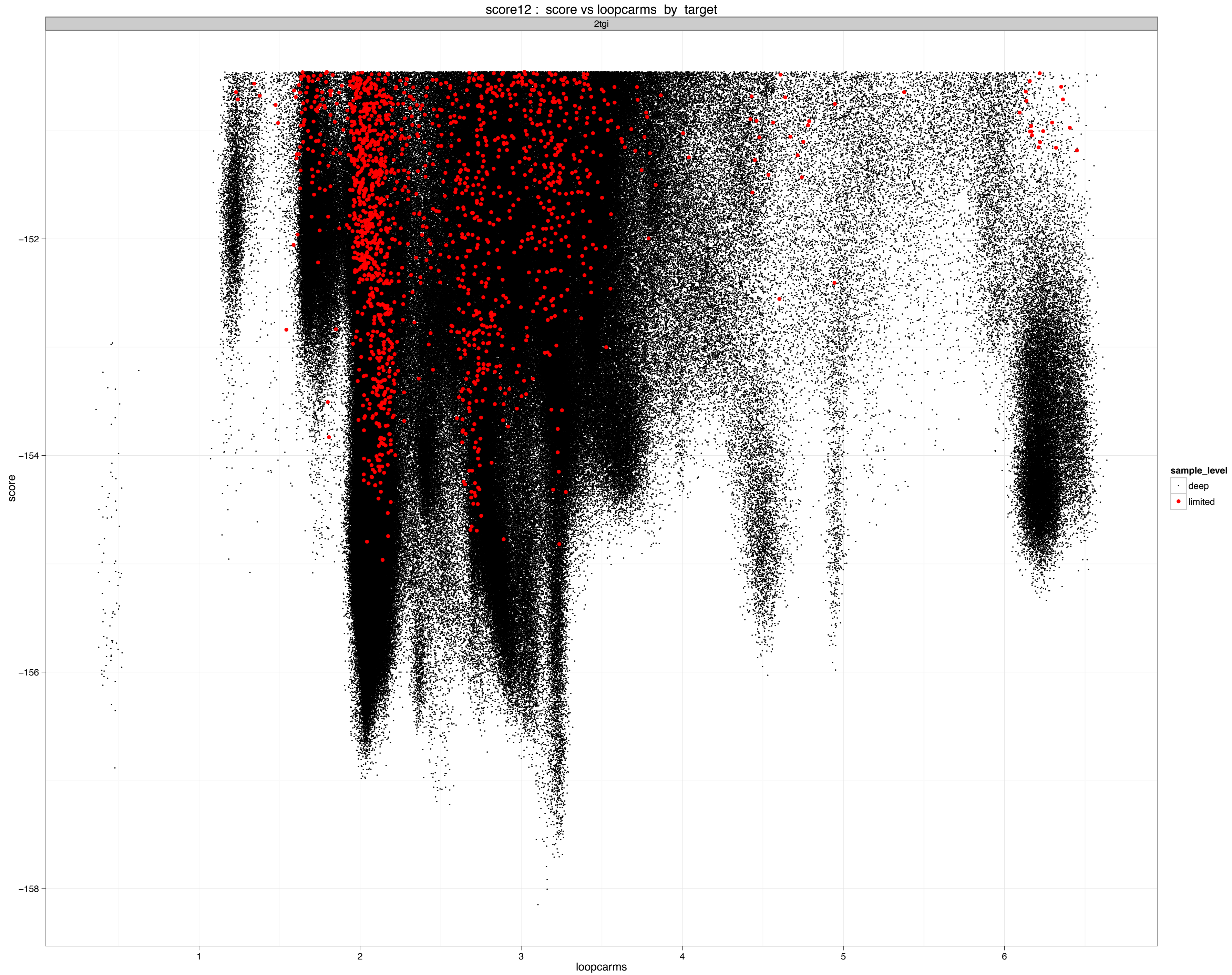
Major Scientific Benchmarks

Currently Available

- Rotamer Recovery
- Sequence Recovery
- ddG Prediction
- Loop Recovery
- Ab-relex Recovery
- Docking local-refine
- RNA benchmark(s)

Upcoming

- Fit into electron density (frank)
- LoopHash discrimination (TJ)
- Ligand docking (Rocco/Sagar/Ora)
- Single Mutant Scan (Yifan)
- Fix-interface Design (Jacob)
- Flex-Interface Design (Sarel)
- Flex-BB Design (Nobu)
- NMR recovery (Oliver)
- <Your Benchmark here>

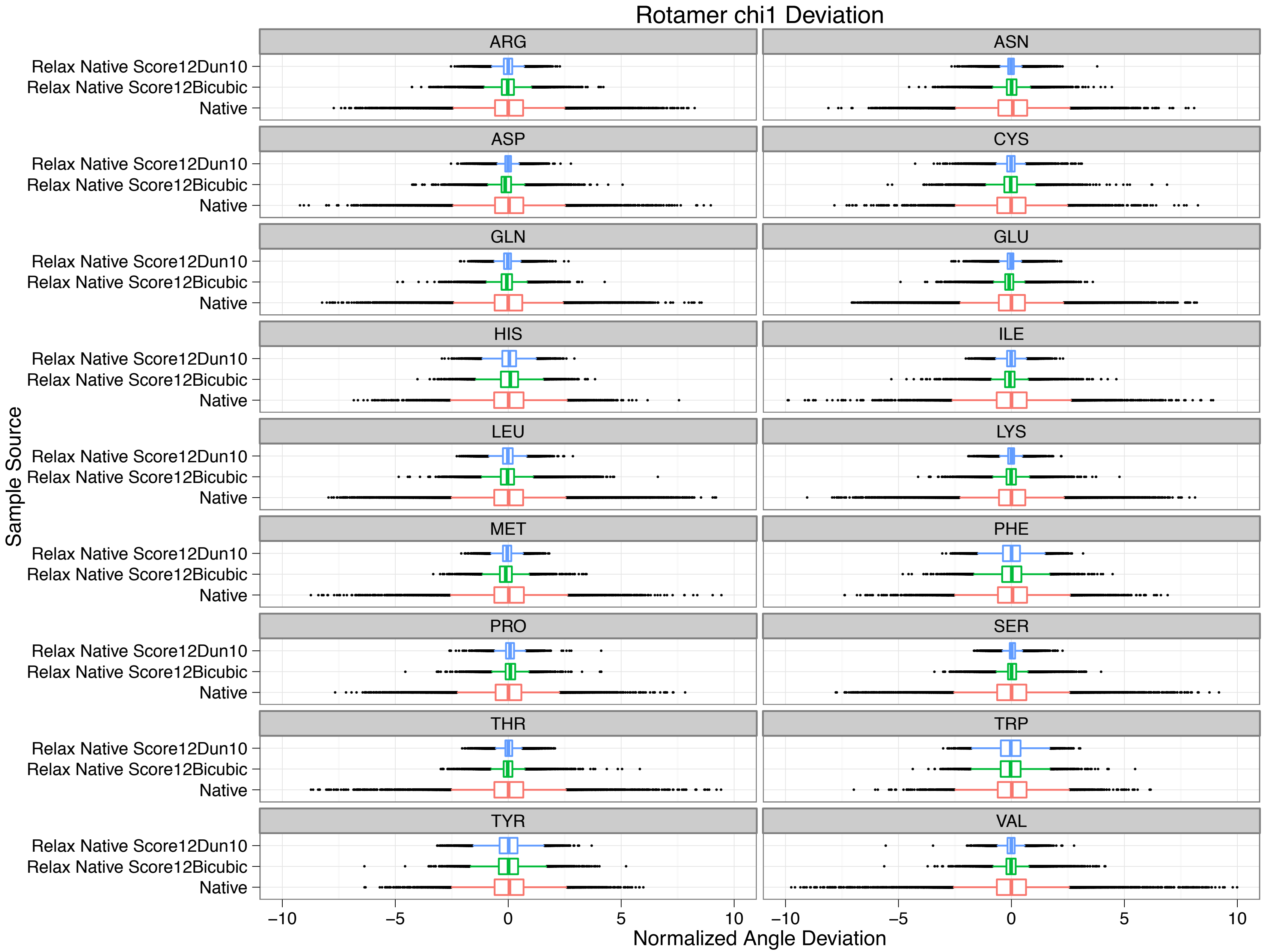

score12 : score vs loopcarms by target
2tgi
score
−152
−154
−156
−158
loopcarms
1
2
3
4
5
6
sample_level
deep
limited

Bicubic/Dun10 Results

	Rotamer Recovery Benchmark				**Seq. Rec. Bench**		**ΔΔG Bench**	**High-Res. Refinement Benchmark**			**Loop Modeling Benchmark**		
Energy Function	*pack rots (%)*	*min pack (%)*	*rot. trials (%)*	*rt-min (%)*	% Rec	KL-Div.	R-Value	#(pNat > 0.8)	Σ pNat	#(eNat < eDec)	1st Quart. (Å)	Med (Å)	3rd Quart. (Å)
Score12	66.19	69.07	71.49	73.12	32.6	0.019	0.69	67	74.6	104	0.468	0.637	1.839
Score12'	-	-	-	-	37.0	0.008	0.67	-	-		-	-	-
Score12Bicubic	66.24	67.51	71.52	73.15	37.6	0.010	0.68	68	77.9	105	0.499	0.644	1.636
Score12Dun10	67.82	70.50	72.60	74.23	37.6	0.009	0.67	60	72.0	104	0.461	0.677	1.463

Thanks

- Brian Kuhlman / Jack Snoeyink
- Andrew Leaver-Fay
- Rosetta Community


Rotamer chi1 Deviation
ARG
ASN
ASP
CYS
GLN
GLU
HIS
ILE
LEU
LYS
MET
PHE
PRO
SER
THR
TRP
TYR
VAL
Relax Native Score12Dun10
Relax Native Score12Bicubic
Native
Sample Source
−10
−5
0
5
10
Normalized Angle Deviation

Top8000 Data Set

- Assembled by Richardson Lab (at Duke)
  - March 2011 snapshot of **Protein Databank**
  - Clustered so intra-cluster homology is at most 70%
  - Filter out structures having
    - greater than 2A resolution
    - known oddities
    - non-canonical amino acids
  - Selected best average MolProbity score and resolution
  - Place hydrogen atoms using Reduce
  - 6,563 Chains

Rotamer Recovery

- 152 chains, 17,463 residues
  - Subset of Top5200 (from Richardson Lab)
  - 50-200 residues each
  - at most 70% seq. homology
  - at most 1.2A resolution
- Recovered: all angles within 20° of native
- Starting info / DOFs:
  - RotamerTrials PackRotamers
  - RTMin MinPack

Sequence Recovery

- 38 large structures (Ding & Dekholyan 2006)
- Accuracy:
  - Native AA
  - Kullback-Leibler divergence with input AA-profile
- Starting info / DOFs:
  - Fixed backbone
  - PackRotamers protocol

High Resolution Refinement

- 114 Sequences (Tyka, *et al*. 2010)
  - – 4 centroid mode data sets
    - homolog fragments / relax to low RMSD + low energy
    - 6,000 FastRelax predictions
- Accuracy:
  - – Boltzmann weighted probability of “near-native”
    - where less than 2A RMSD => near-native
  - – If 80% near-native
  - – If min near-native energy < min decoy energy

Loop Prediction

- 45 12-Residue Loops (Mandell *et al*. 2009)
  - 8,000 Kinimatic Loop Closure (KIC) predictions
- Accuracy:
  - Min Cα-RMSD over 5 lowest-energy structures

ddG Prediction

- 1210 Point Mutants (Kellogg *et al*. 2011)
  - Native crystal structures
  - Experimental ddG of folding
- Accuracy:
  - correlation coefficient
- Input info / DOFs:
  - all-atom, soft-rep repacking
  - backbone + sidechain, hard-rep minimization
  - uniform constraints