Statistical Potentials in Rosetta

Roland Dunbrack

Institute for Cancer Research
Fox Chase Cancer Center
Philadelphia PA 19111

Roland.Dunbrack@fccc.edu
http://dunbrack.fccc.edu

OUR NEW JUSTICE

Confirmed 68-31

Statistical Potentials in Rosetta

- Ramachandran term (coil residues) $p(\phi,\psi|Res)$
- Rotamer term (all residues) $p(r|\phi,\psi,Res)$
- Design term (all residues) $p(Res|\phi,\psi)$
- Hydrogen bond, other terms...

Important to be smooth, differentiable (ϕ,ψ) $d\log p/d\phi$

Based on accurate input data

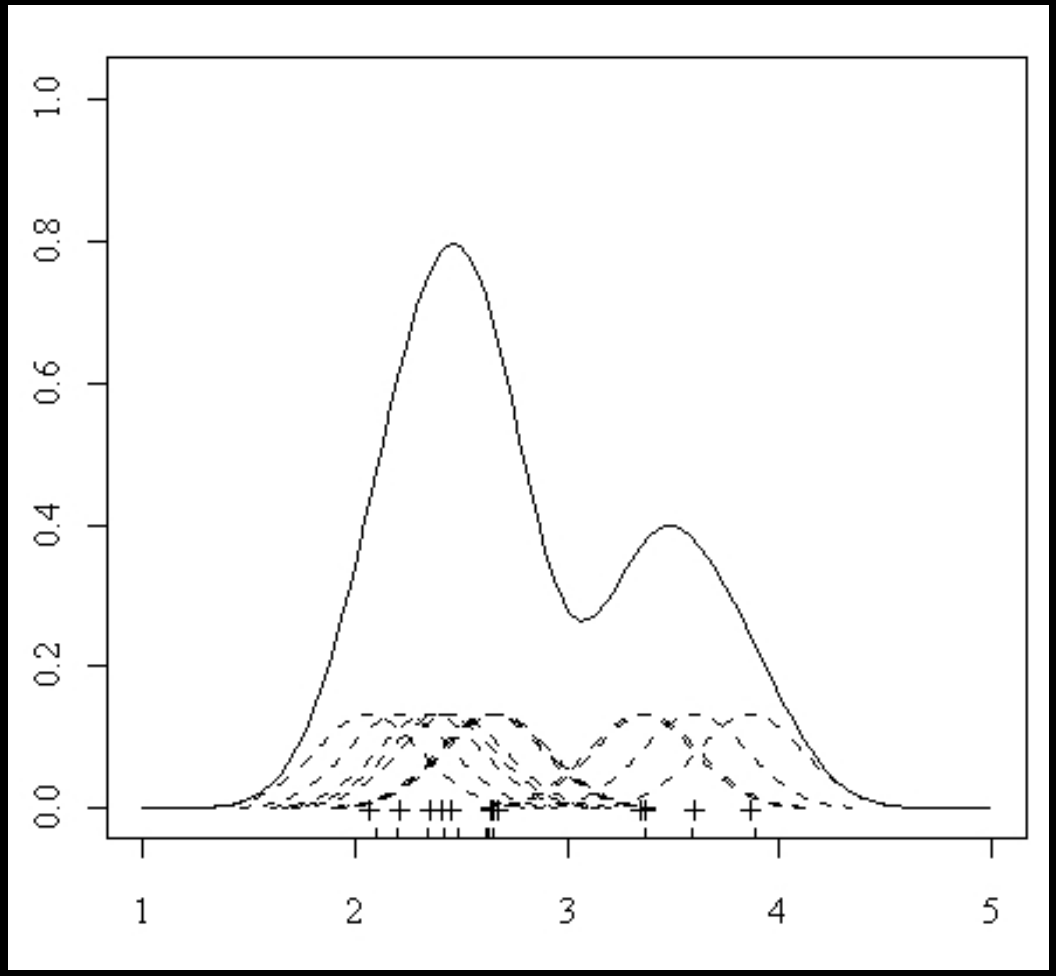
What input data sets? (all sec. str, coil, turns?)

Non-parametric statistics
Kernel density estimates

Assign a function (kernel) to every data point to spread out or smooth the data. At each "query point" (e.g. on a regular grid), add the value of all the kernels.

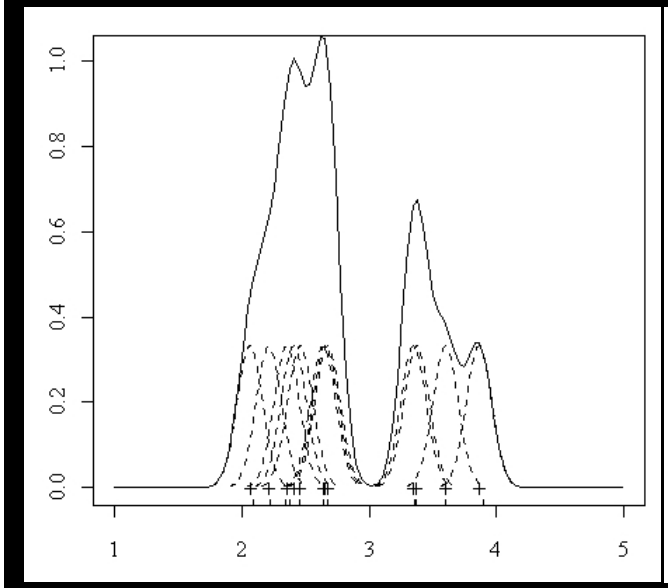
$$\hat{p}(x) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

$$\int_U K(u)\,du = 1$$

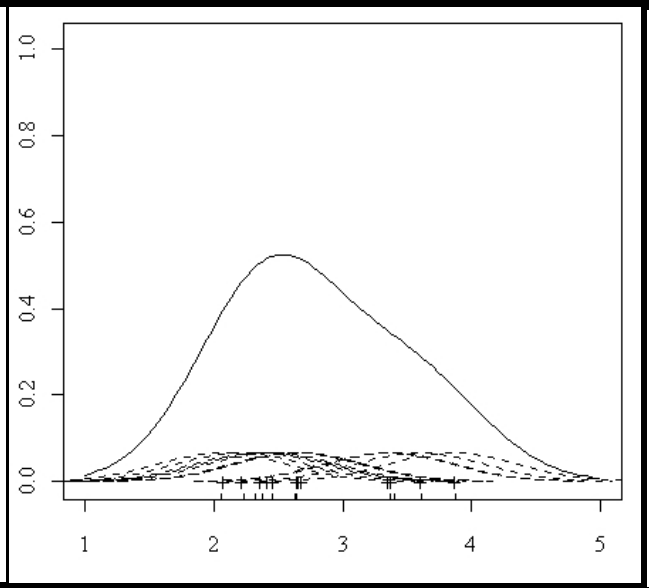


Kernel width

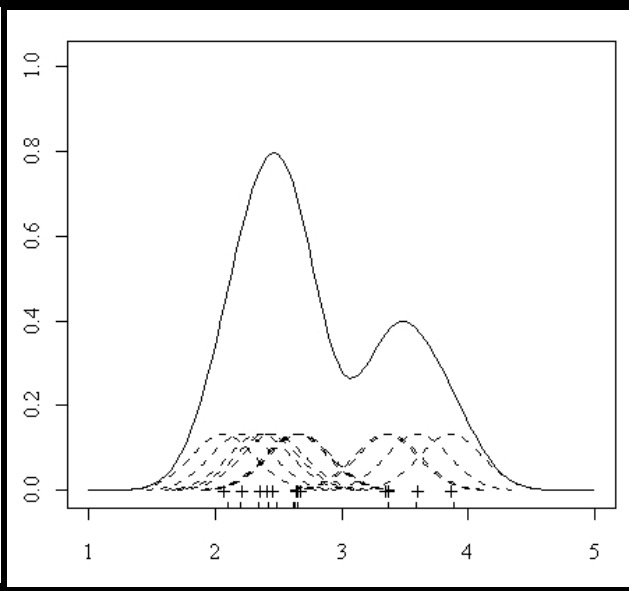
Undersmoothed



Oversmoothed

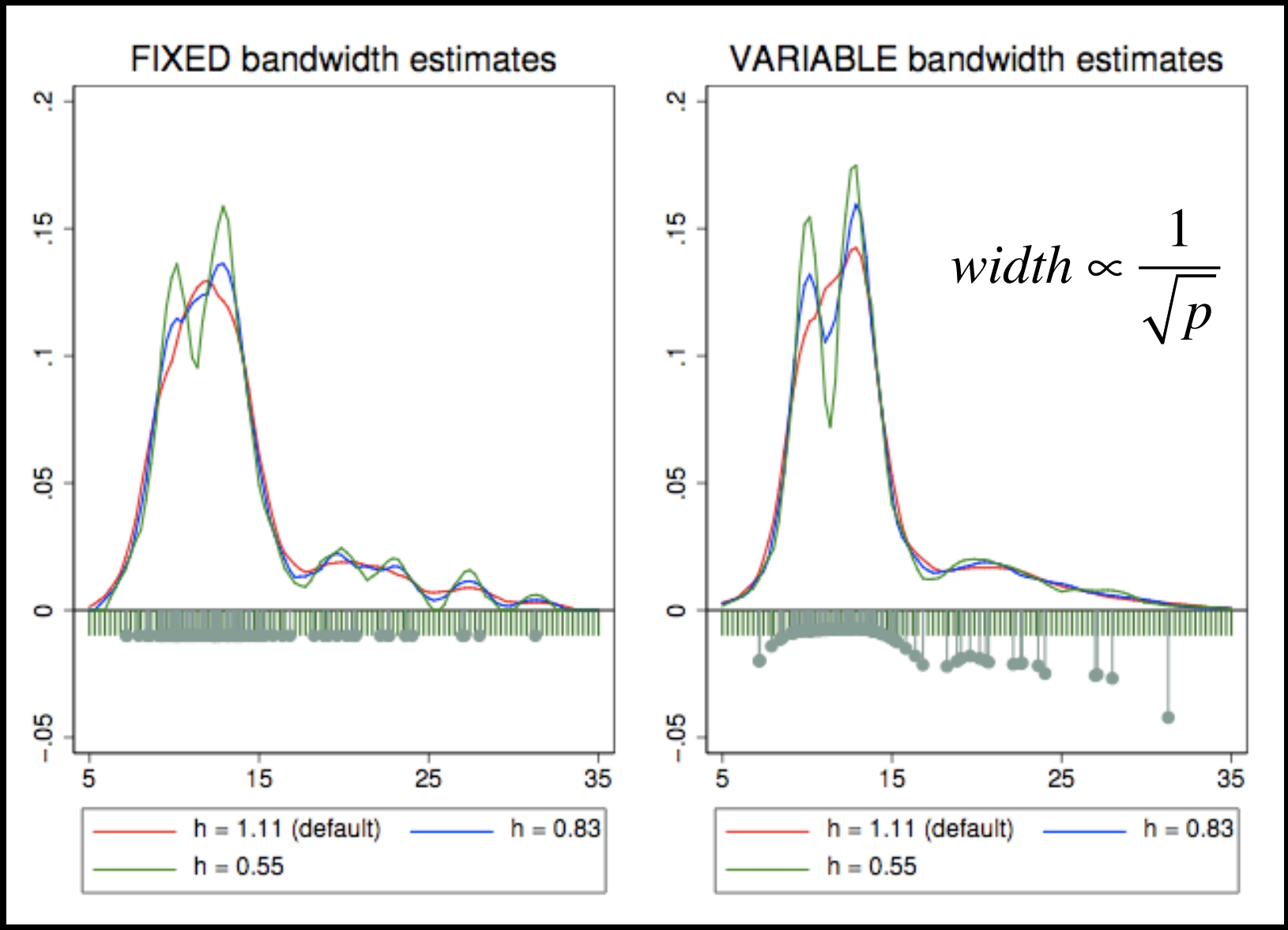


Optimally smoothed



http://www.maths.uwa.edu.au/~duongt/seminars/intro2kde/

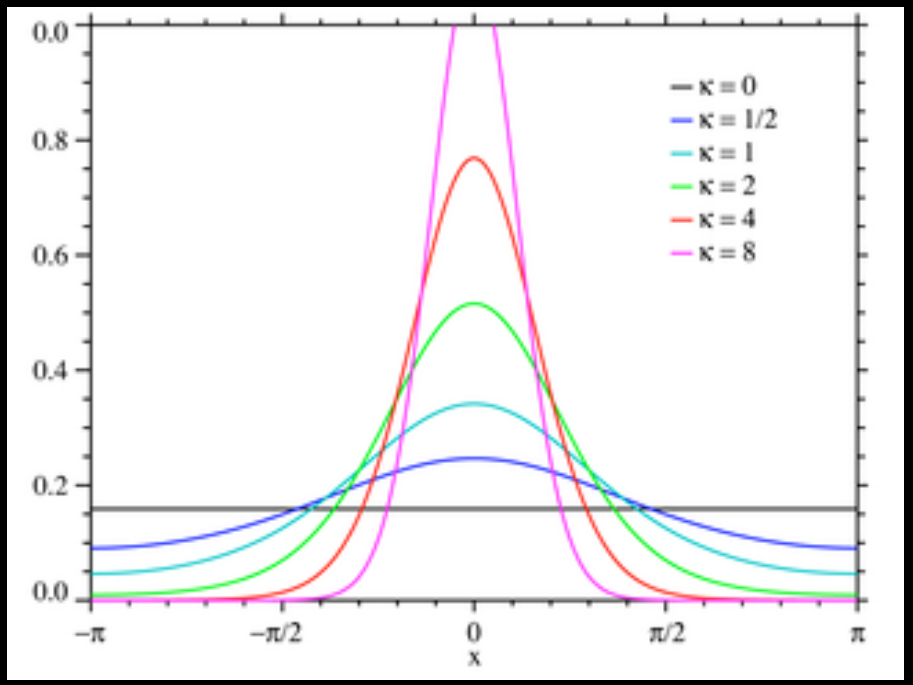
Adaptive kernel density estimates



Philippe van Kerm: http://ideas.repec.org/p/boc/usug03/15.html

Directional statistics
Von Mises Probability Distributions

$$p(\theta) = \frac{\exp\left(\kappa \cos(\theta - \theta_0)\right)}{2\pi I_0(\kappa)}$$



(wikipedia page on Von Mises distns)

For large kappa, very close to normal distributions

Using Von Mises Distributions for Adaptive Kernel Density Estimates of Angular Data

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\pi I_0(\kappa/h_i)} \exp\left(\frac{\kappa}{h_i}\cos(x - x_i)\right)$$

Ramachandran PDF

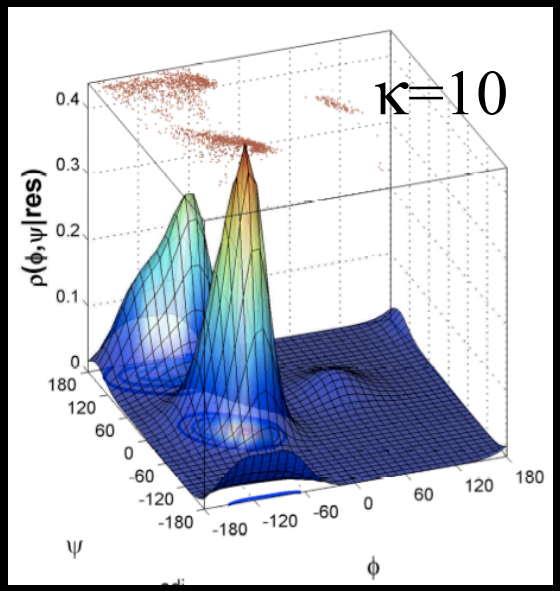
$$p(\phi,\psi|Res) = \frac{1}{4\pi^2 N}\sum_{i=1}^{N} \frac{1}{\left(I_0(\kappa/\lambda_i)\right)^2} \exp\left(\frac{\kappa}{\lambda_i}\left(\cos(\phi_i - \phi) + \cos(\psi_i - \psi)\right)\right)$$

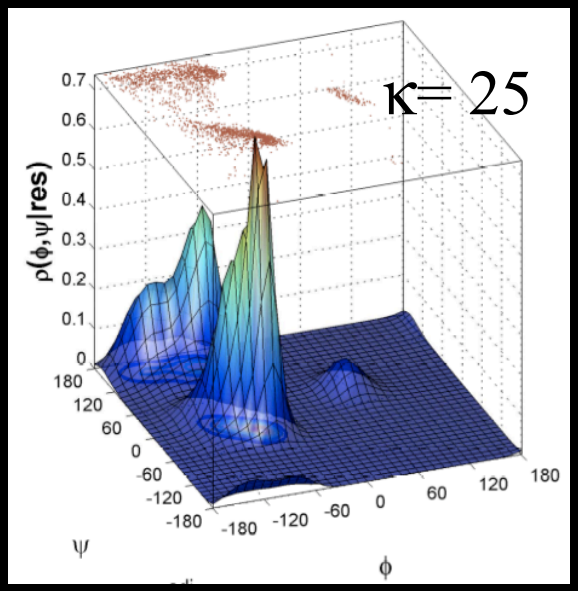
$$\lambda_l = \sqrt{\frac{\prod_i \hat{f}(\phi_i,\psi_i)}{\hat{f}(\phi_i,\psi_i)}}$$

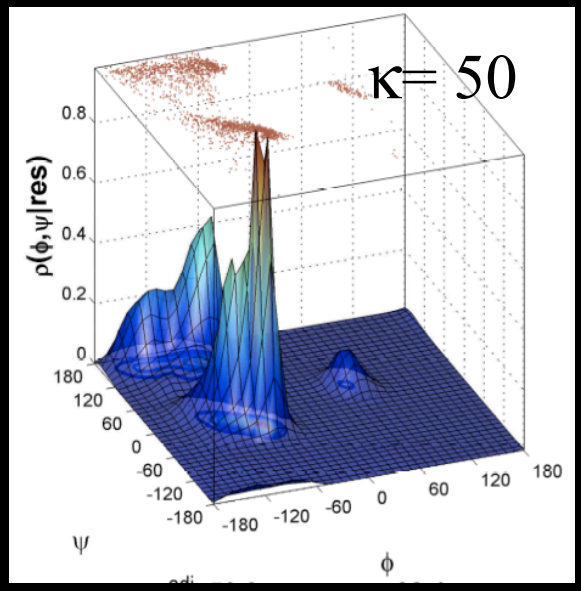
κ becomes geometric mean of kernel widths

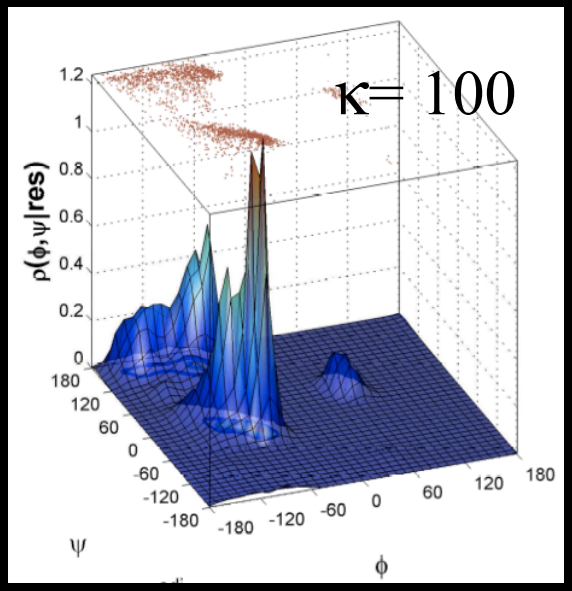
Kernel widths from pilot density estimate from non-adaptive KDE

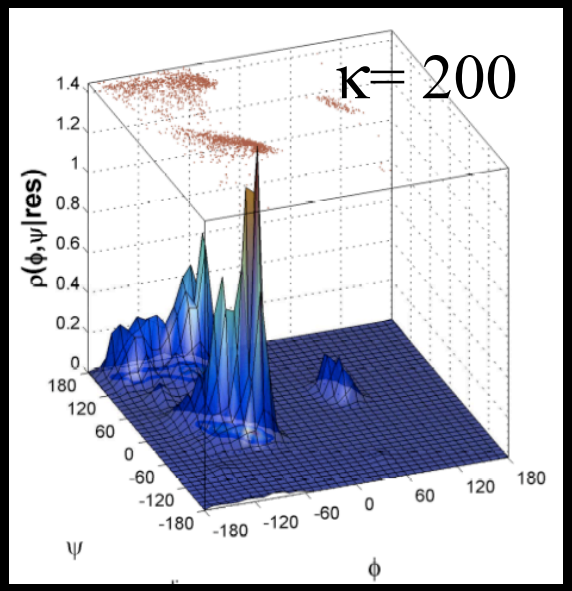
Rama potential: kernel widths











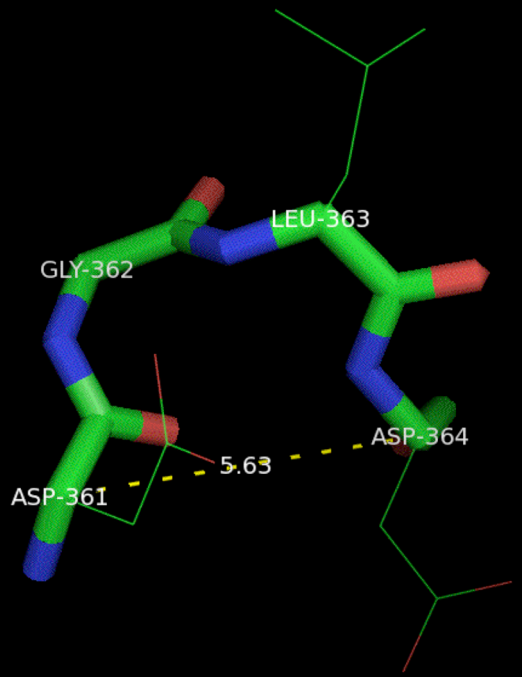
Gln

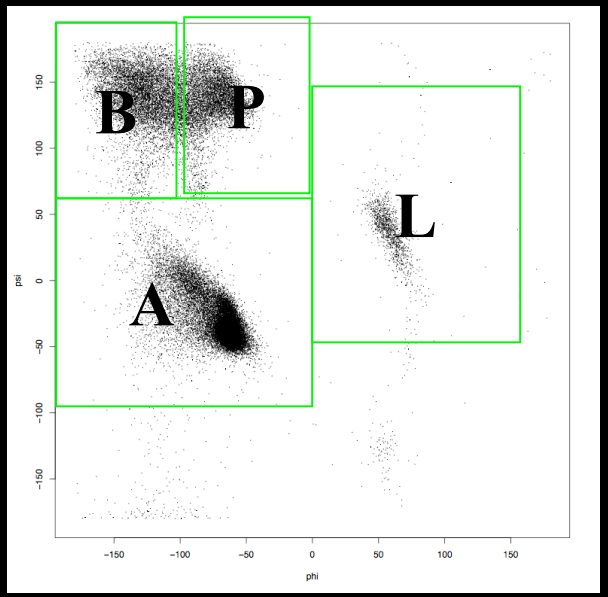
Ramachandran Data

10495 Loops len≥8

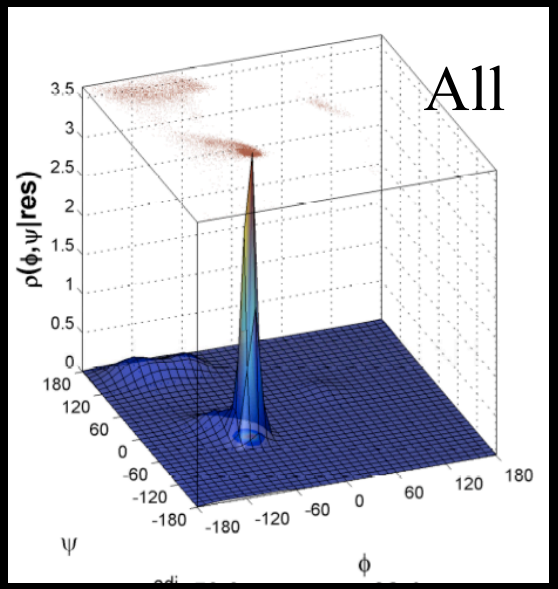
Turn	47%
Coil	36%
310H	13%
Bridge	4%

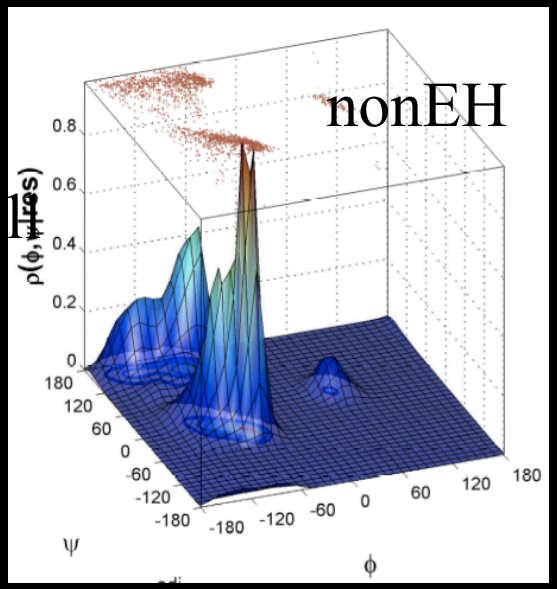
	Loop	Turn	Coil	310H
A	38%	42%	18%	96%
B	23%	19%	32%	0%
P	28%	25%	41%	0%
L	10%	13%	9%	3%

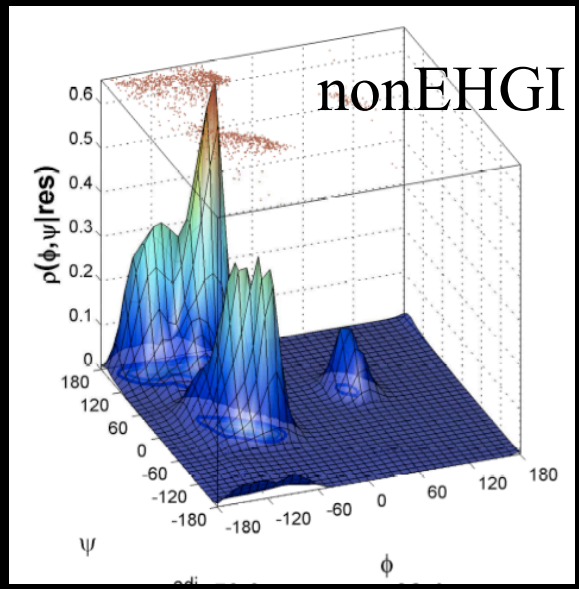


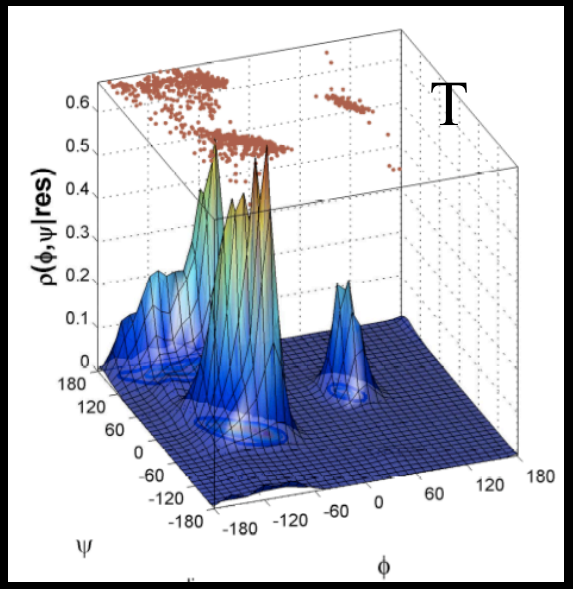


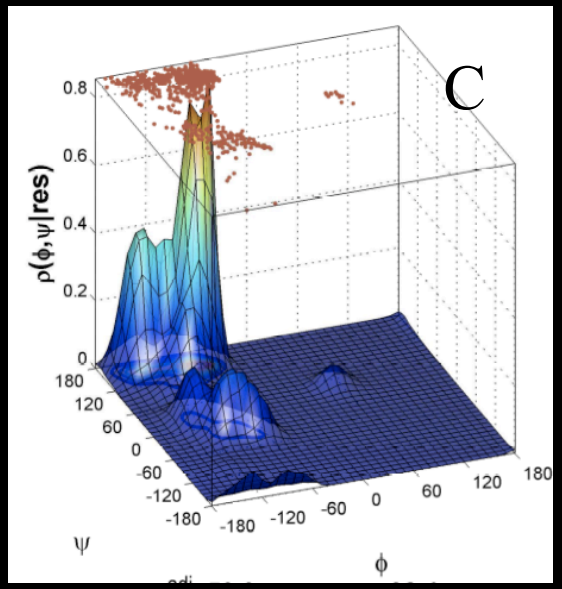
Rama potential: input sets

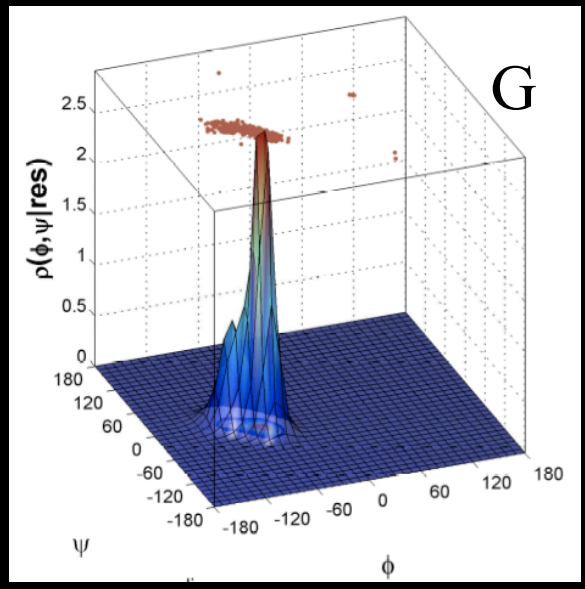




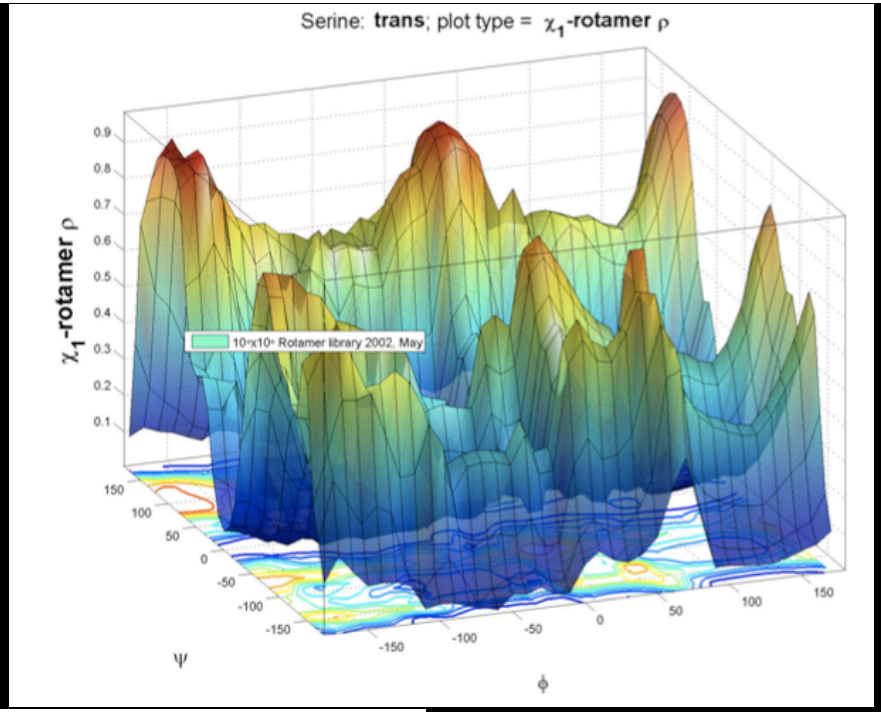








2002 Library



850 proteins

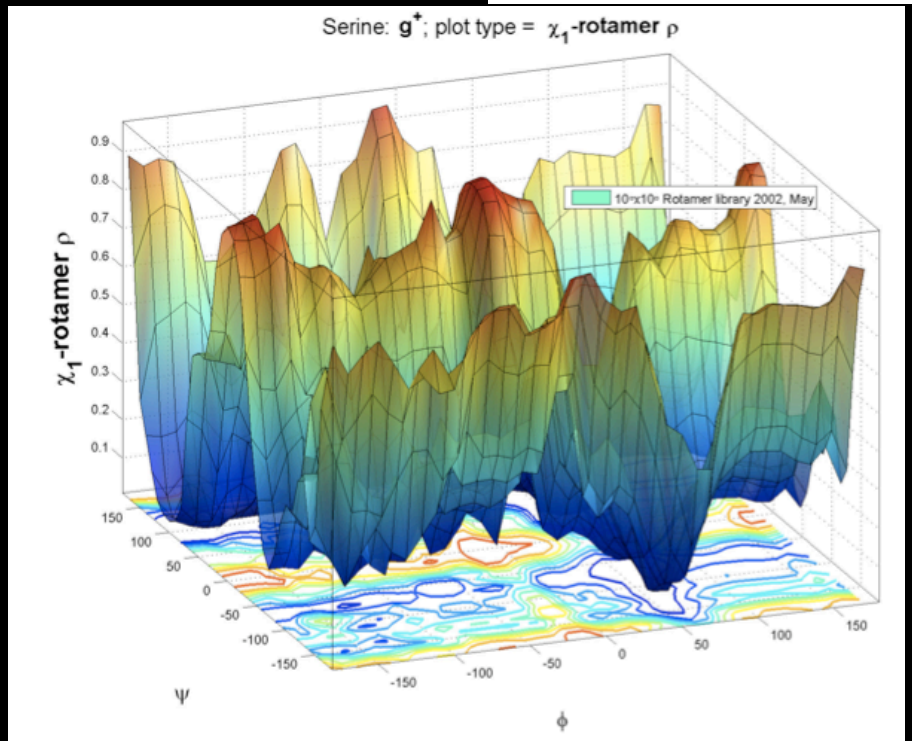
≤1.7Å

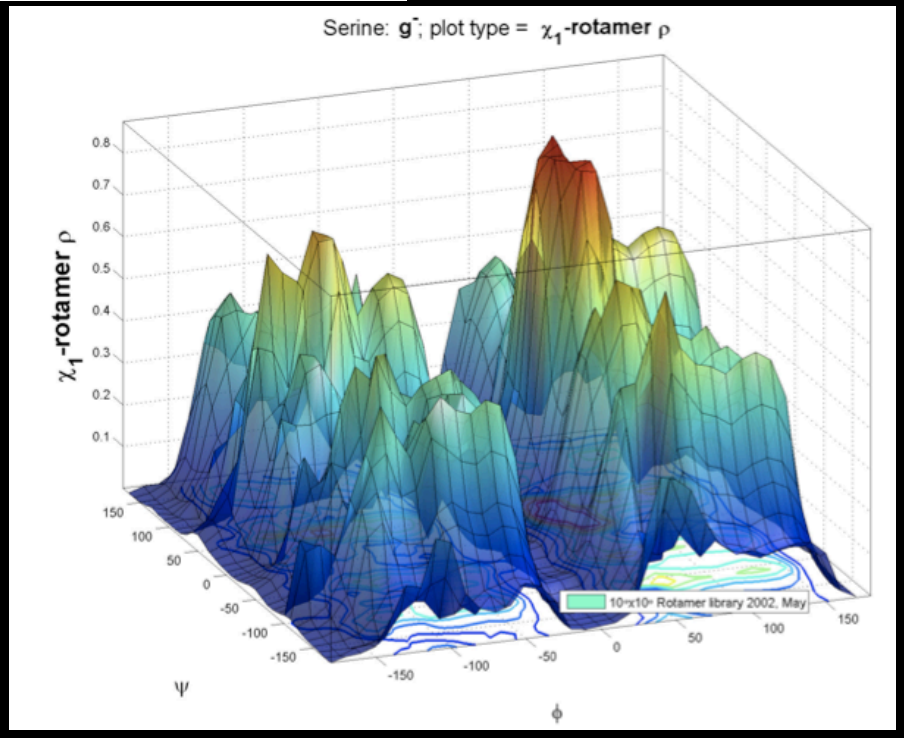
≤50% id

Asn/Gln/His flips

B-factor<40

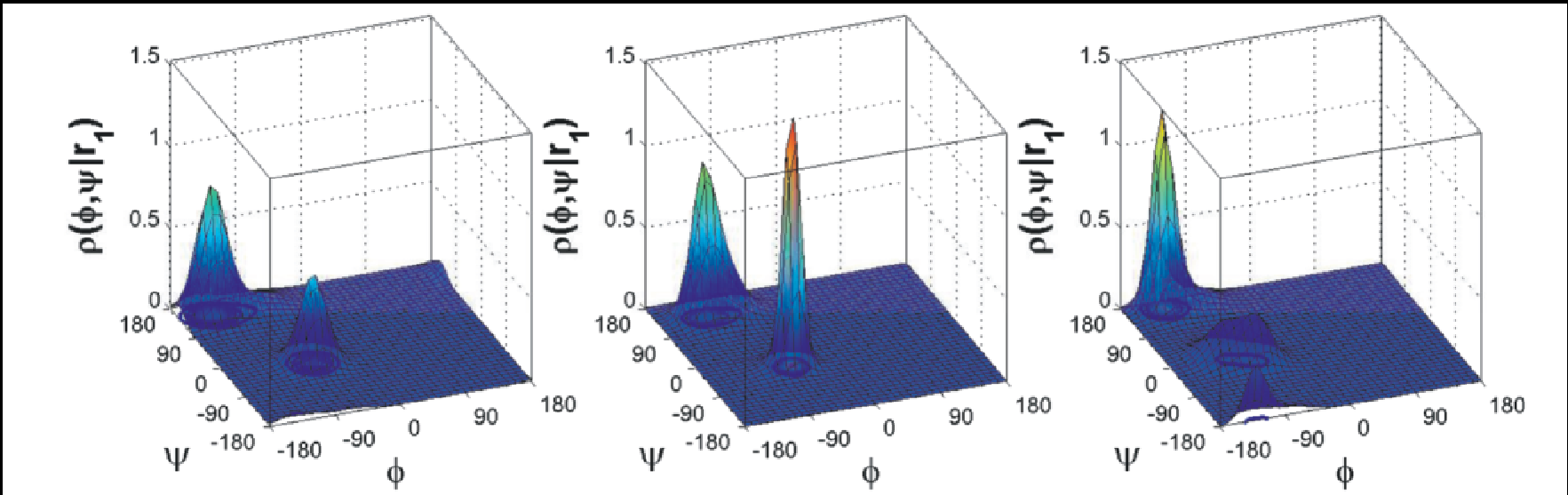
No contacts





Adaptive kernel density estimates
Ramachandran distribution for each rotamer of Valine

g+ trans g+



$$p(\phi,\psi|r) = \frac{1}{4\pi^2 N_r}\sum_{i=1}^{N_r}\frac{1}{\left(I_0(\kappa/\lambda_i)\right)^2}\exp\left(\frac{\kappa}{\lambda_i}\left(\cos(\phi_i-\phi)+\cos(\psi_i-\psi)\right)\right)$$

Bayes' rule

We can get $p(\phi,\psi|r)$. How do we get $p(r|\phi,\psi)$?

Invert using Bayes' rule:

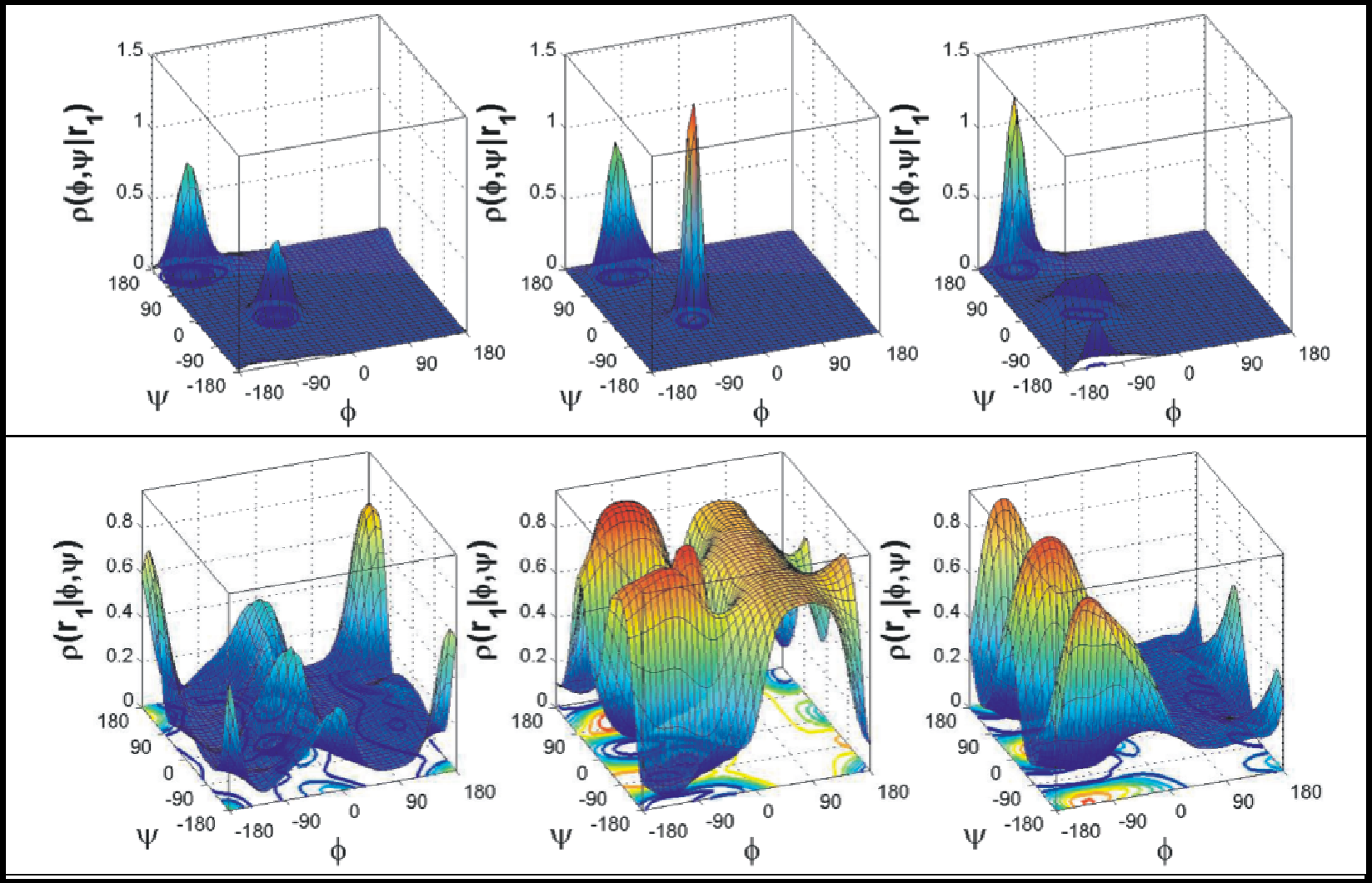
$$p(A,B) = p(A|B)p(B) = p(B|A)p(A)$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

$$p(r|\phi,\psi) = \frac{p(\phi,\psi|r)p(r)}{\sum_{r'} p(\phi,\psi|r')p(r')}$$

Valine

g+ trans g-



Kernel widths
Cross validation of log likelihood

In ϕ,ψ space, rotamer-dep

$$L(r;\kappa) = \prod_i p(\phi_i, \psi_i | r)$$

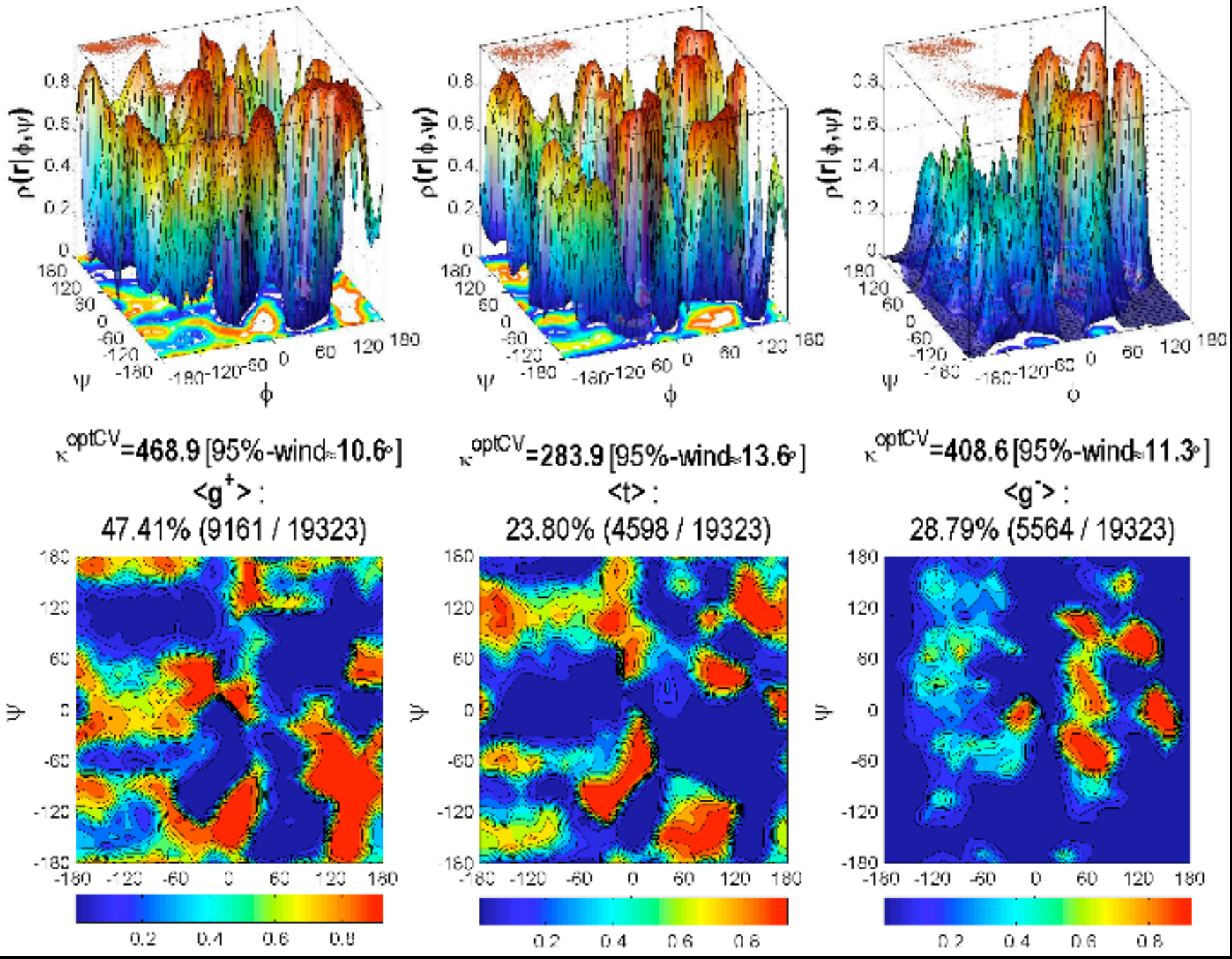
$$\log L(r;\kappa) = \sum_i \ln p(\phi_i, \psi_i | r)$$

Maximize log L for each rotamer as a function of κ by calculating p on 90% of data and evaluating L on 10%

In ϕ,ψ space, rotamer-indep

$$L(\kappa) = \prod_i p(\phi_i, \psi_i)$$

$$\log L(\kappa) = \sum_i \ln p(\phi_i, \psi_i)$$


ρ(r|ϕ,ψ)
ϕ
ψ
κ^optCV=468.9 [95%-wind≈10.6°]
<g+> :
47.41% (9161 / 19323)
κ^optCV=283.9 [95%-wind≈13.6°]
<t> :
23.80% (4598 / 19323)
κ^optCV=408.6 [95%-wind≈11.3°]
<g-> :
28.79% (5564 / 19323)

Correct way

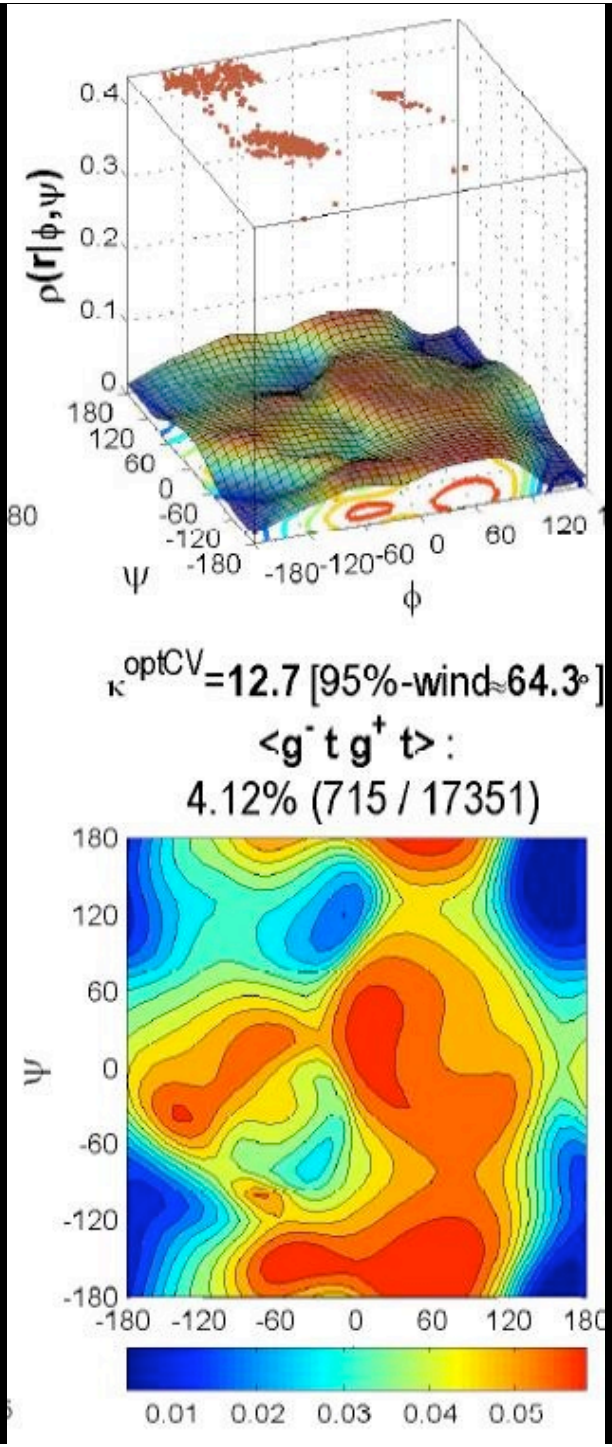
In *r* space

$$L(\kappa) = \prod_i p\left(r \middle| \phi_i, \psi_i\right)$$

$$\log L(\kappa) = \sum_i \log p\left(r \middle| \phi_i, \psi_i\right)$$

Same (ϕ,ψ) --> same κ

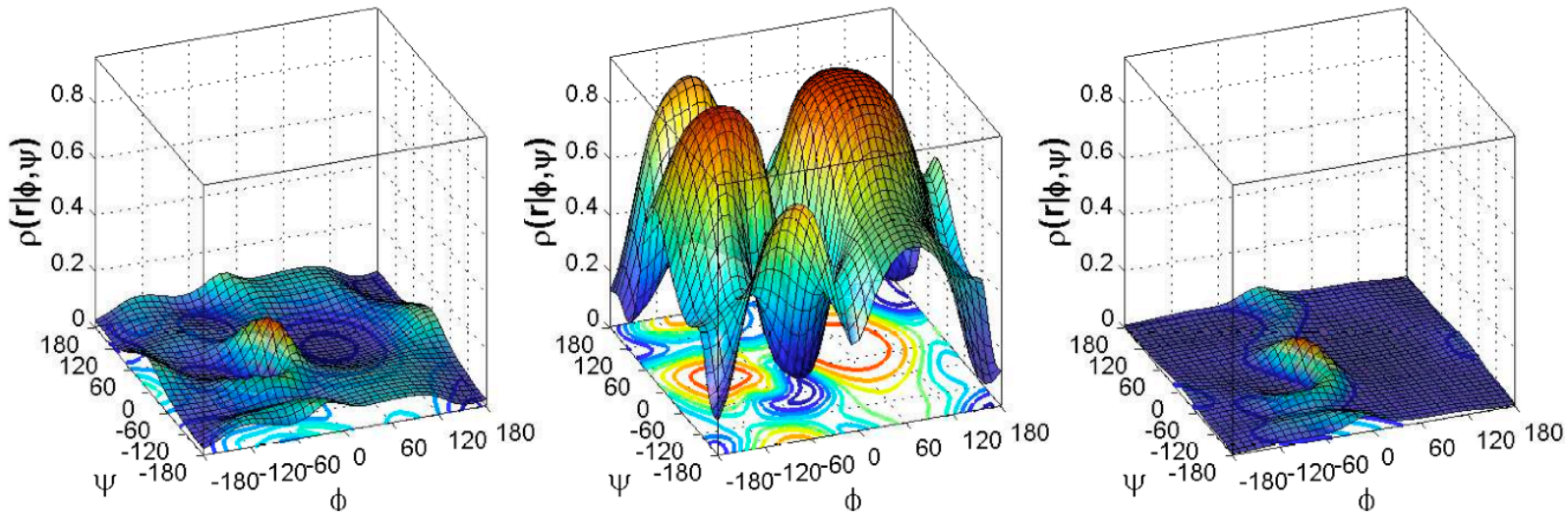
$$\lambda_i = \left(\frac{\left(\prod_{j=1,N} \hat{f}\left(\phi_j, \psi_j\right) \right)^{\frac{1}{N}}}{\hat{f}\left(\phi_i, \psi_i\right)} \right)^{\frac{1}{2}}$$

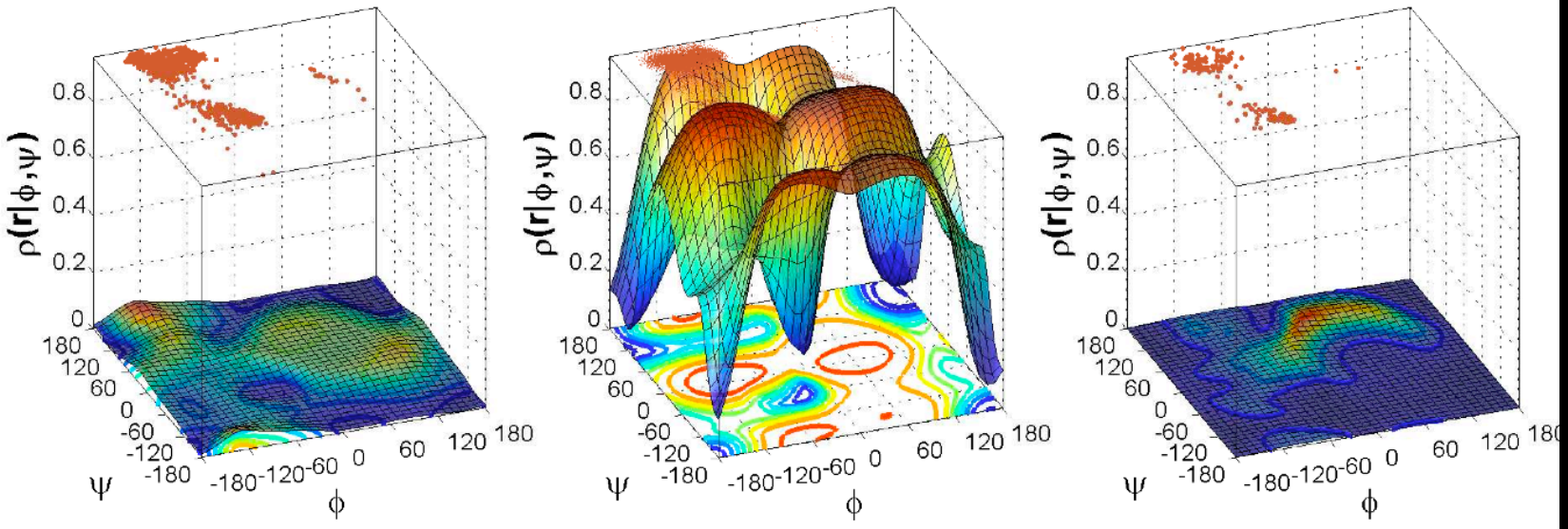


July 2008

Feb 2009

LEU κ_1^m = 25 [$2\sigma \approx$ 22.9°], α_0 = 0.5; dataPoints, adap





κ^{optCV}=25.3 [95%-wind≈45.5°]

$\langle g^- g^+ \rangle$:

3.21% (954 / 29742)

κ^{optCV}=25.3 [95%-wind≈45.5°]

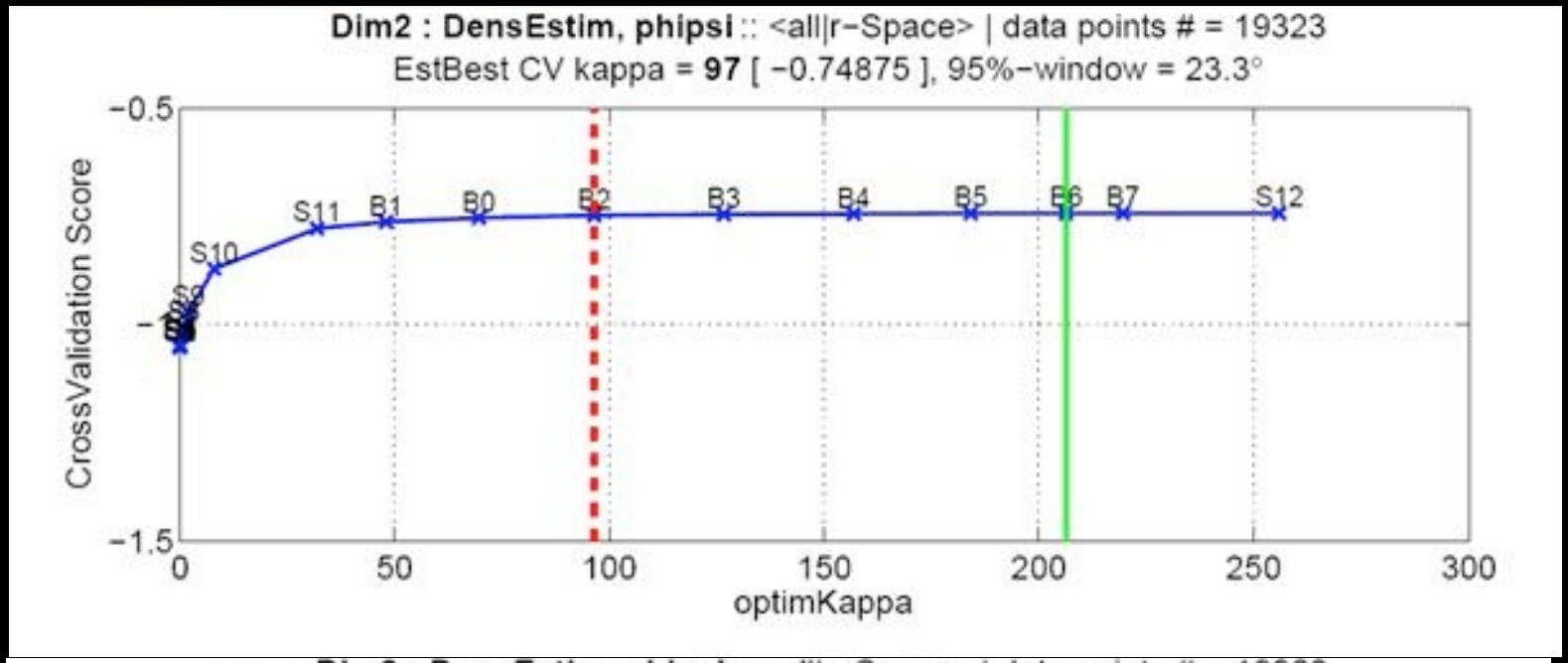
$\langle g^- t \rangle$:

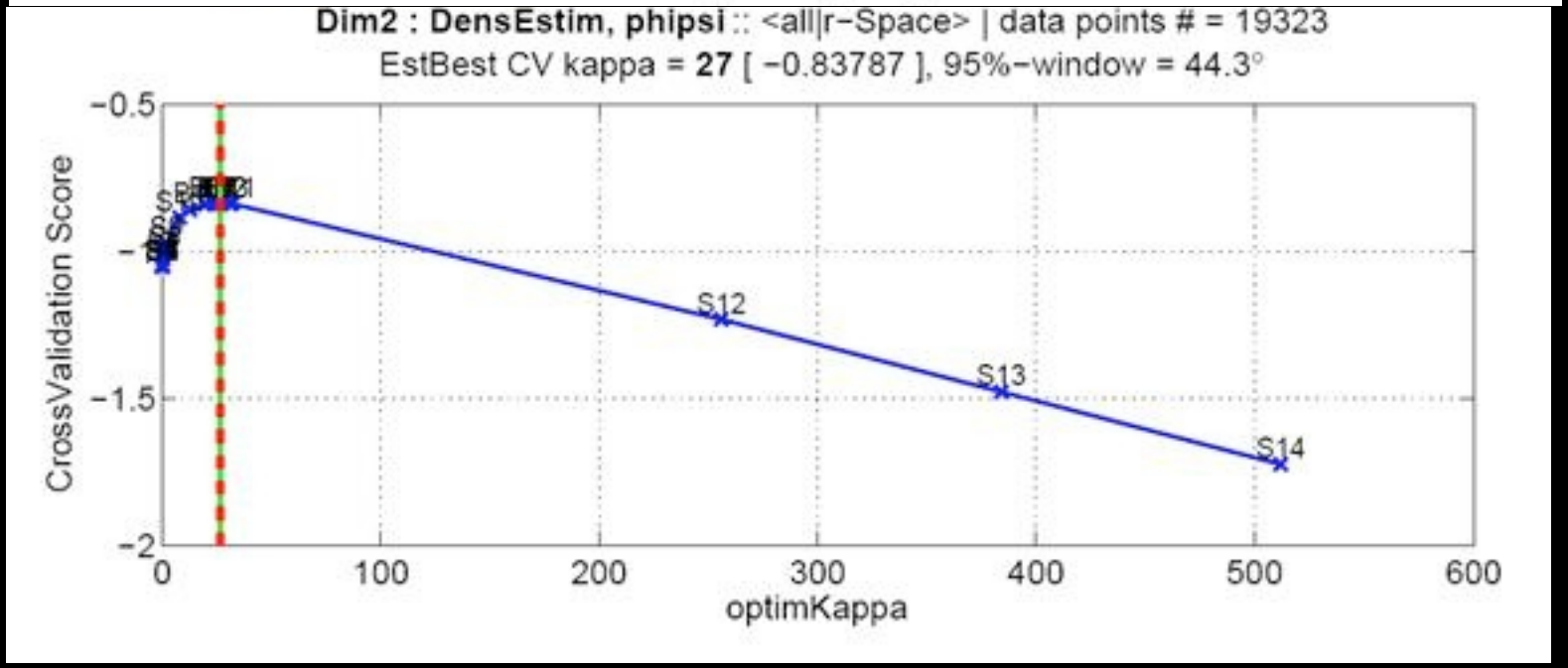
62.30% (18529 / 29742)

κ^{optCV}=25.3 [95%-wind≈45.5°]

$\langle g^- g^- \rangle$:

0.65% (192 / 29742)

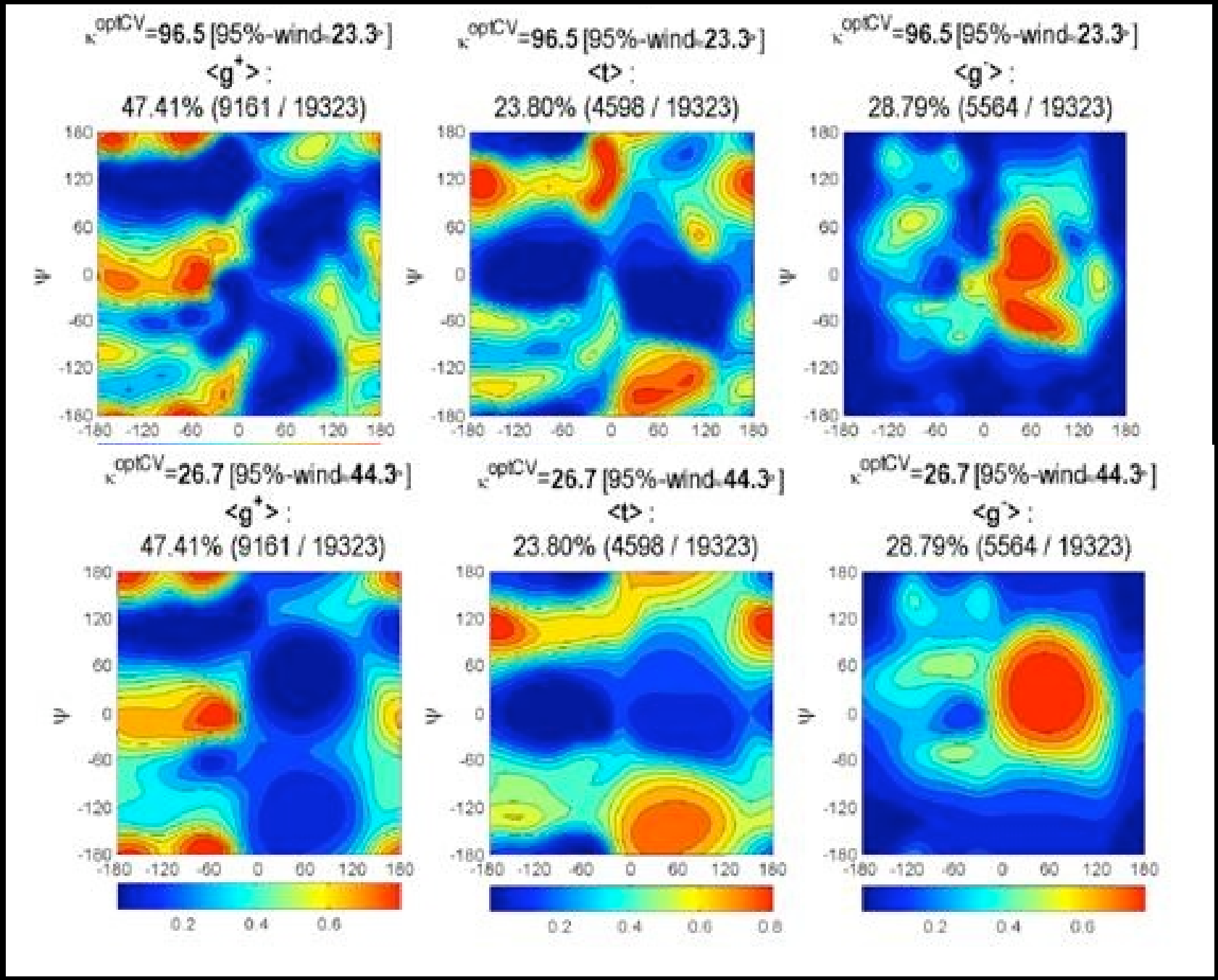

Dim2 : DensEstim, phipsi :: <all|r−Space> | data points # = 19323
EstBest CV kappa = 97 [−0.74875], 95%−window = 23.3°
CrossValidation Score
optimKappa
−0.5
−1.5
0
50
100
150
200
250
300
S10
S11
B1
B0
B2
B3
B4
B5
B6
B7
S12


Dim2 : DensEstim, phipsi :: <all|r−Space> | data points # = 19323
EstBest CV kappa = 27 [−0.83787], 95%−window = 44.3°
CrossValidation Score
optimKappa
−0.5
−1.5
−2
0
100
200
300
400
500
600
S12
S13
S14

No prior

With Prior

June 2009



Design term

Current

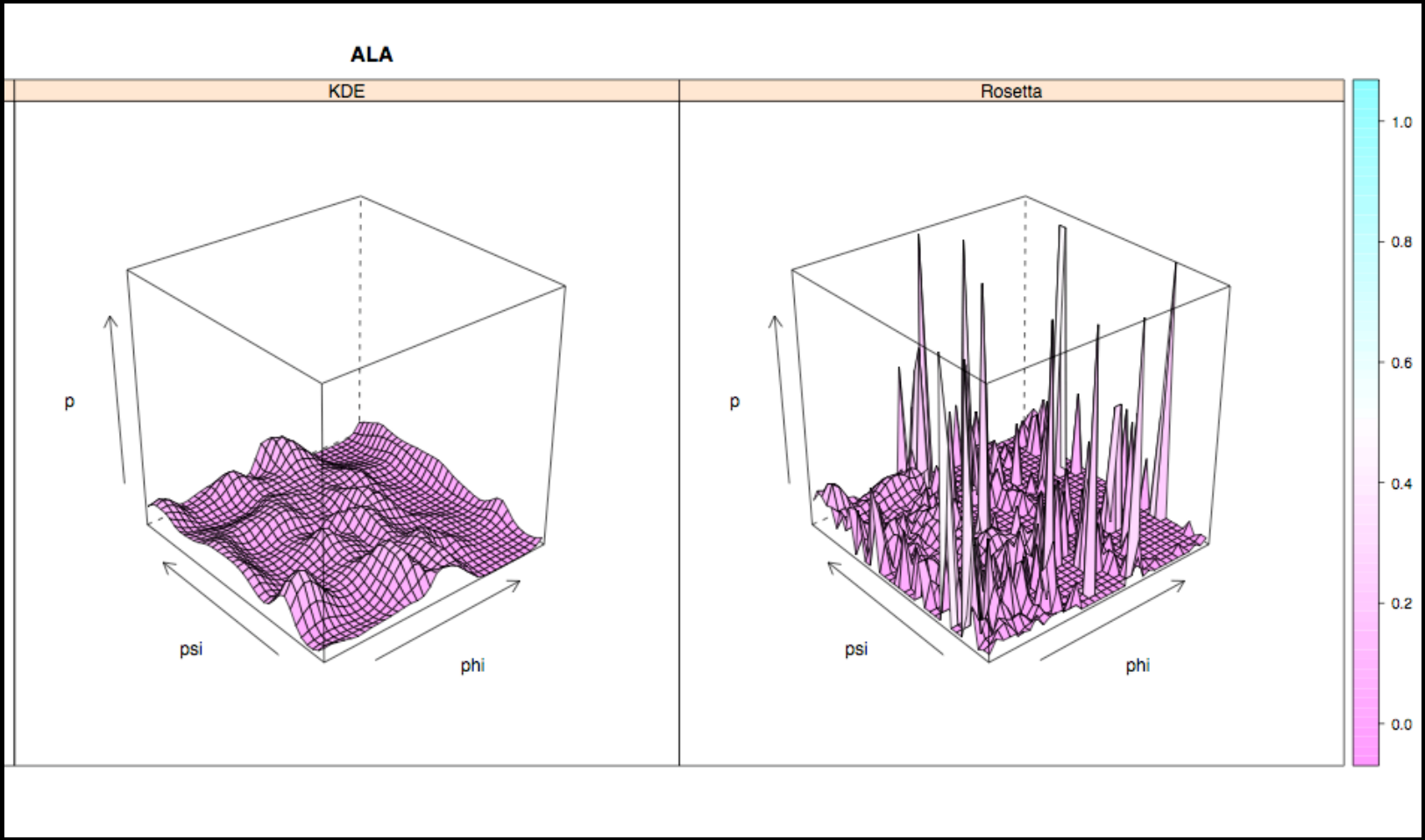
$$p\left(Res\middle|\phi,\psi\right)=\frac{N\left(Res\middle|\phi\pm\Delta,\psi\pm\Delta\right)}{N\left(\phi\pm\Delta,\psi\pm\Delta\right)}$$

Proposed

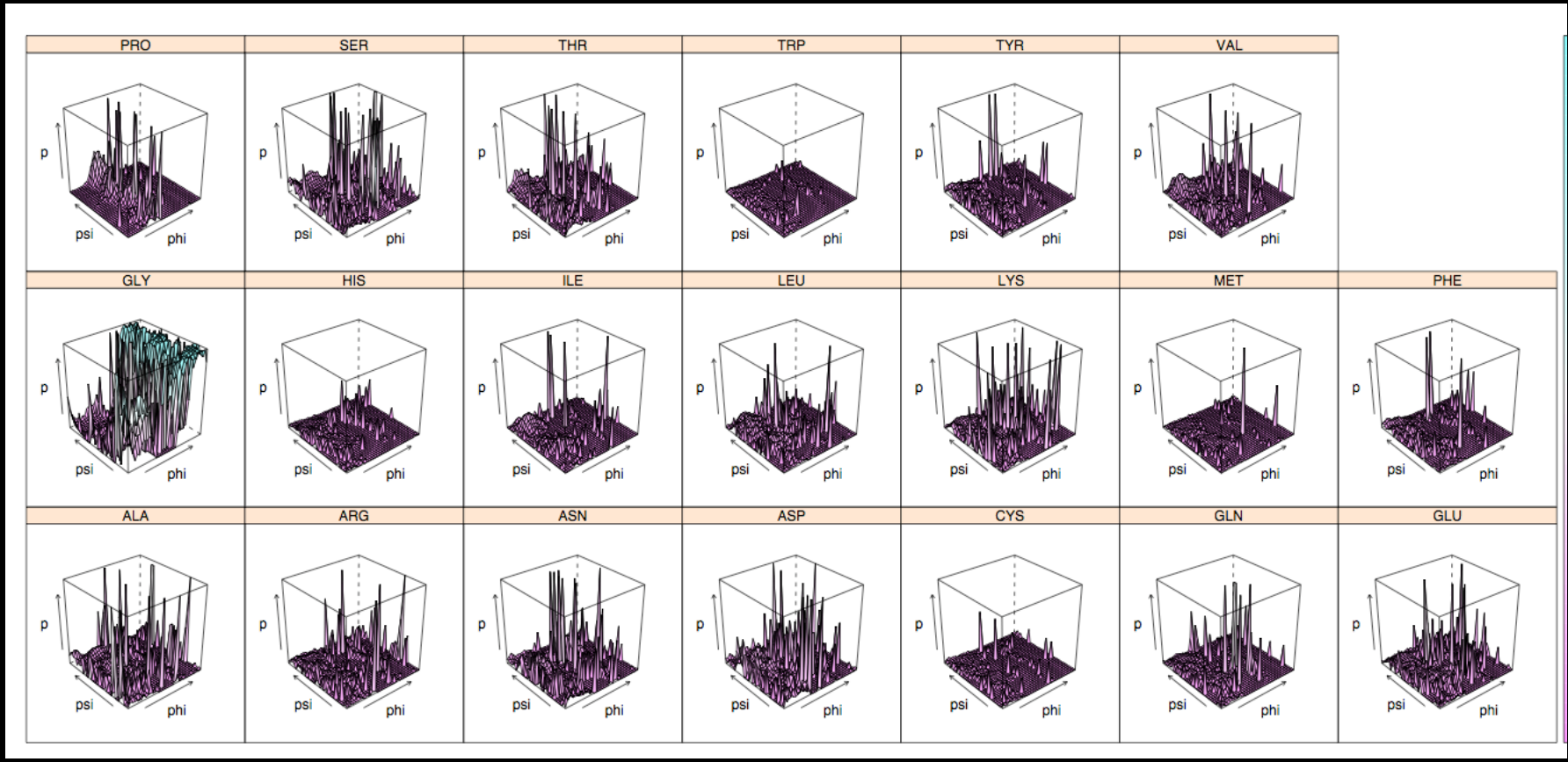
$$p\left(Res\middle|\phi,\psi\right)=\frac{p\left(\phi,\psi\middle|Res\right)p\left(Res\right)}{\sum_{i}p\left(\phi,\psi\middle|Res_{i}\right)p\left(Res_{i}\right)}$$

$p\left(\phi,\psi\middle|Res\right)$ comes from kernel density estimates

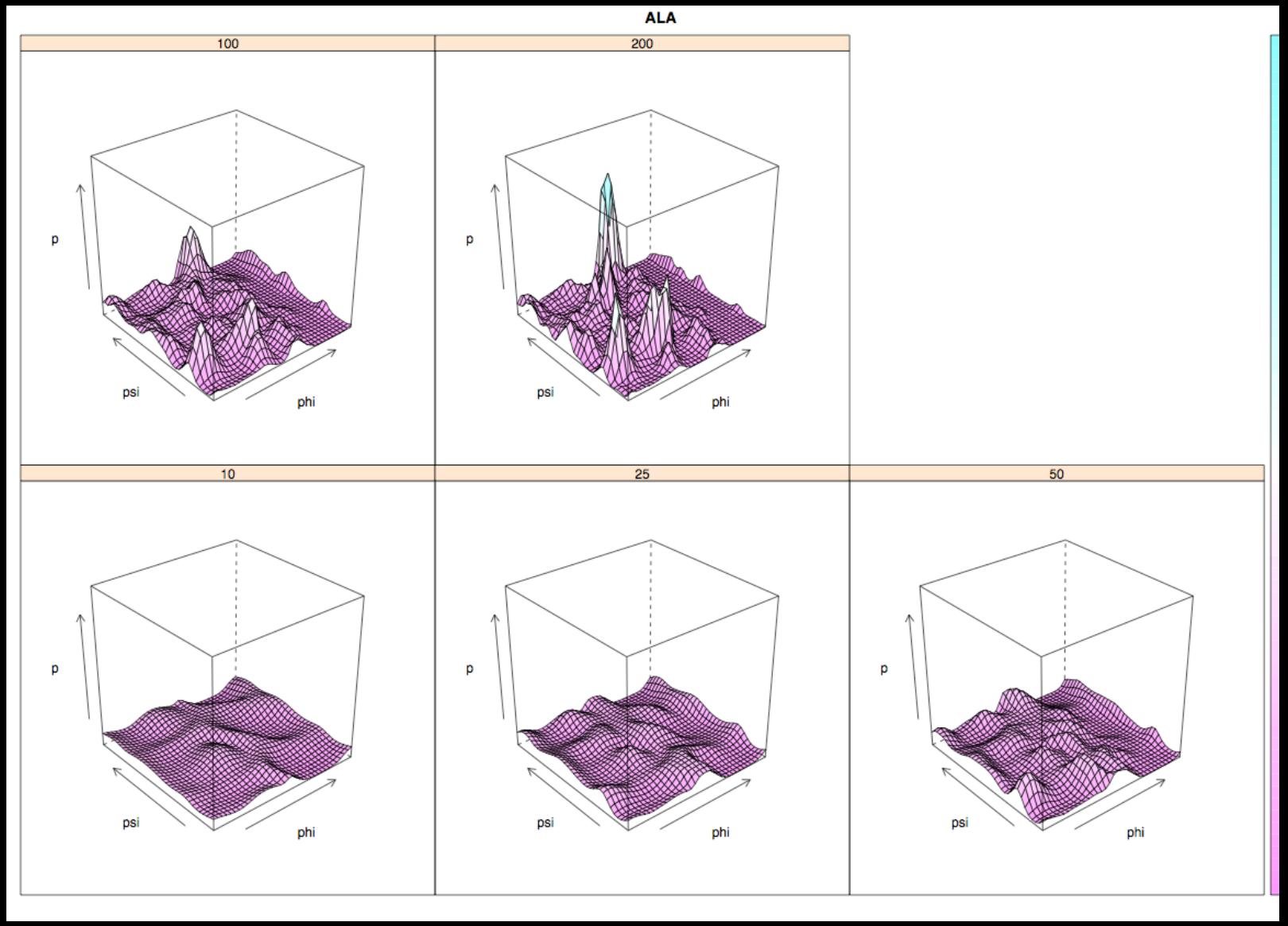
Propensities



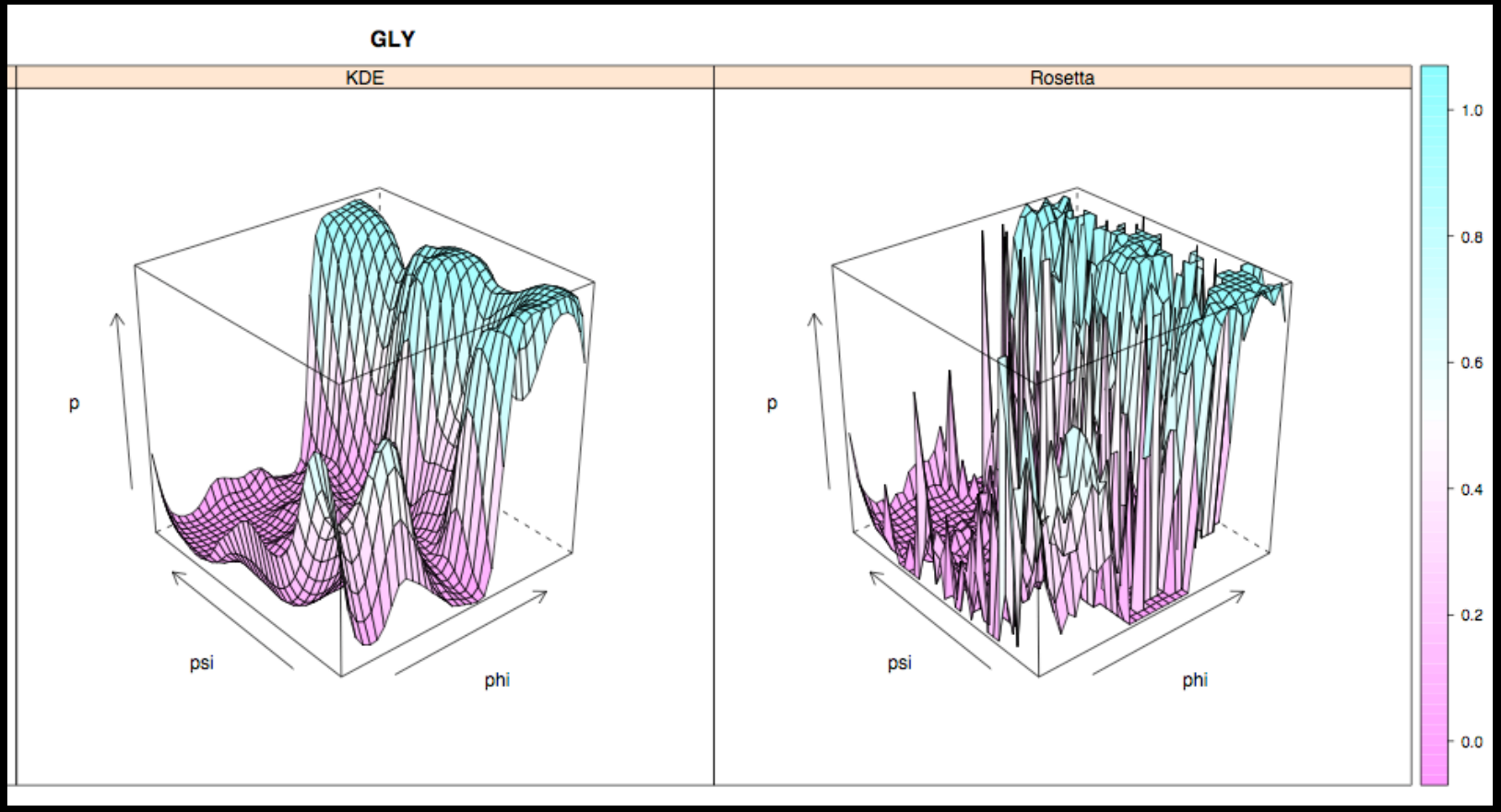
Rosetta's current design term



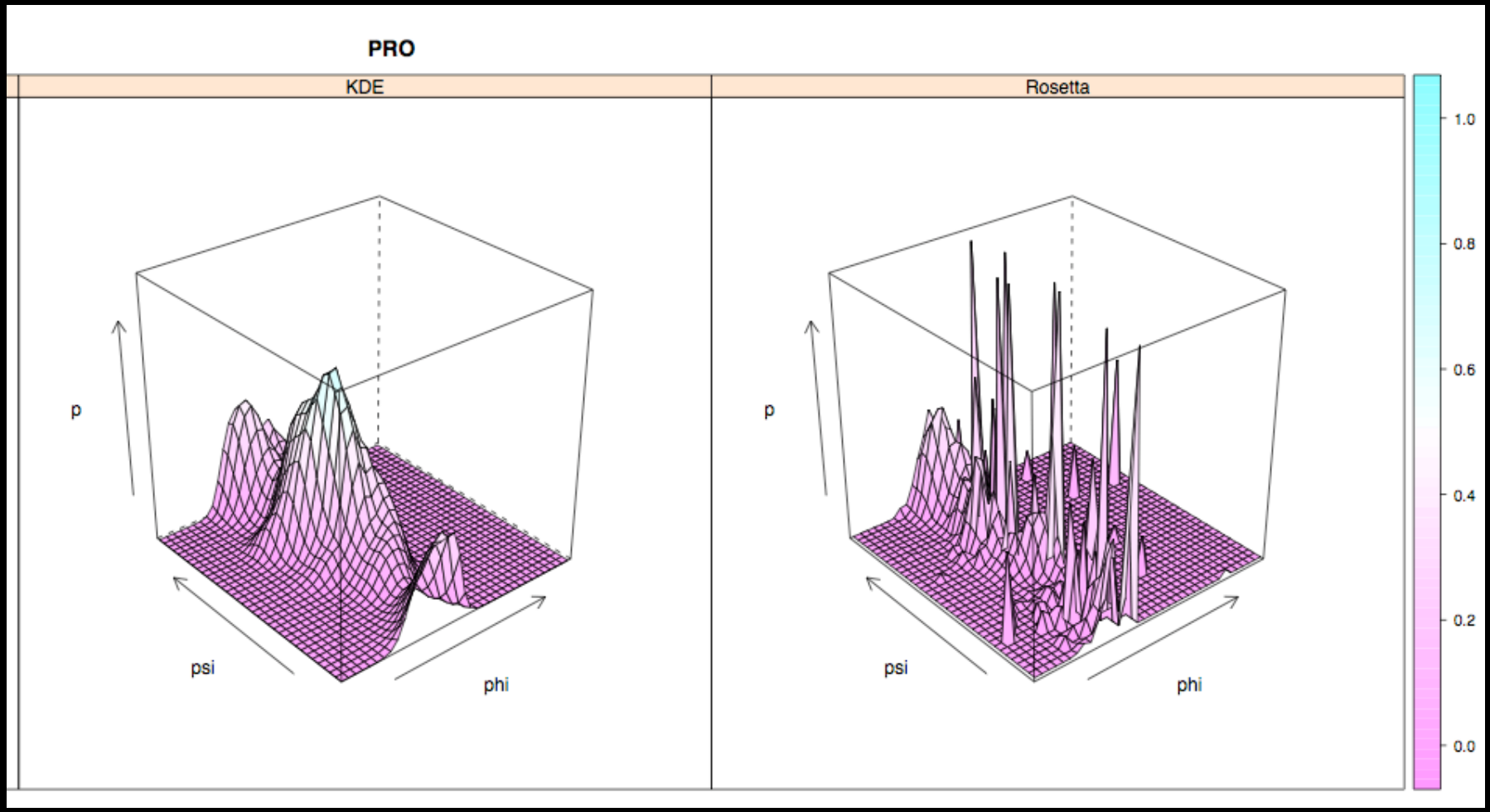
Kappa



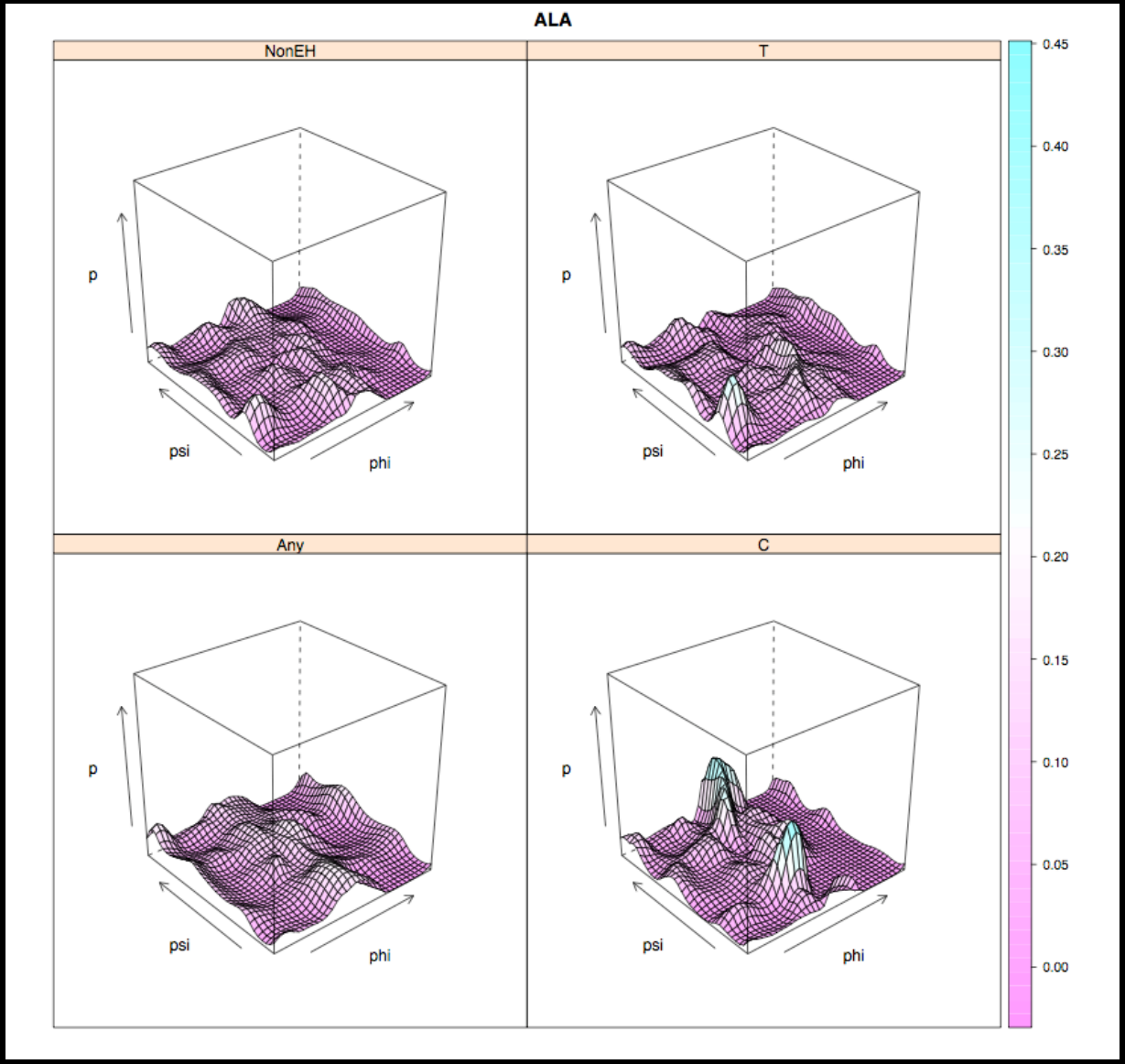
Glycine



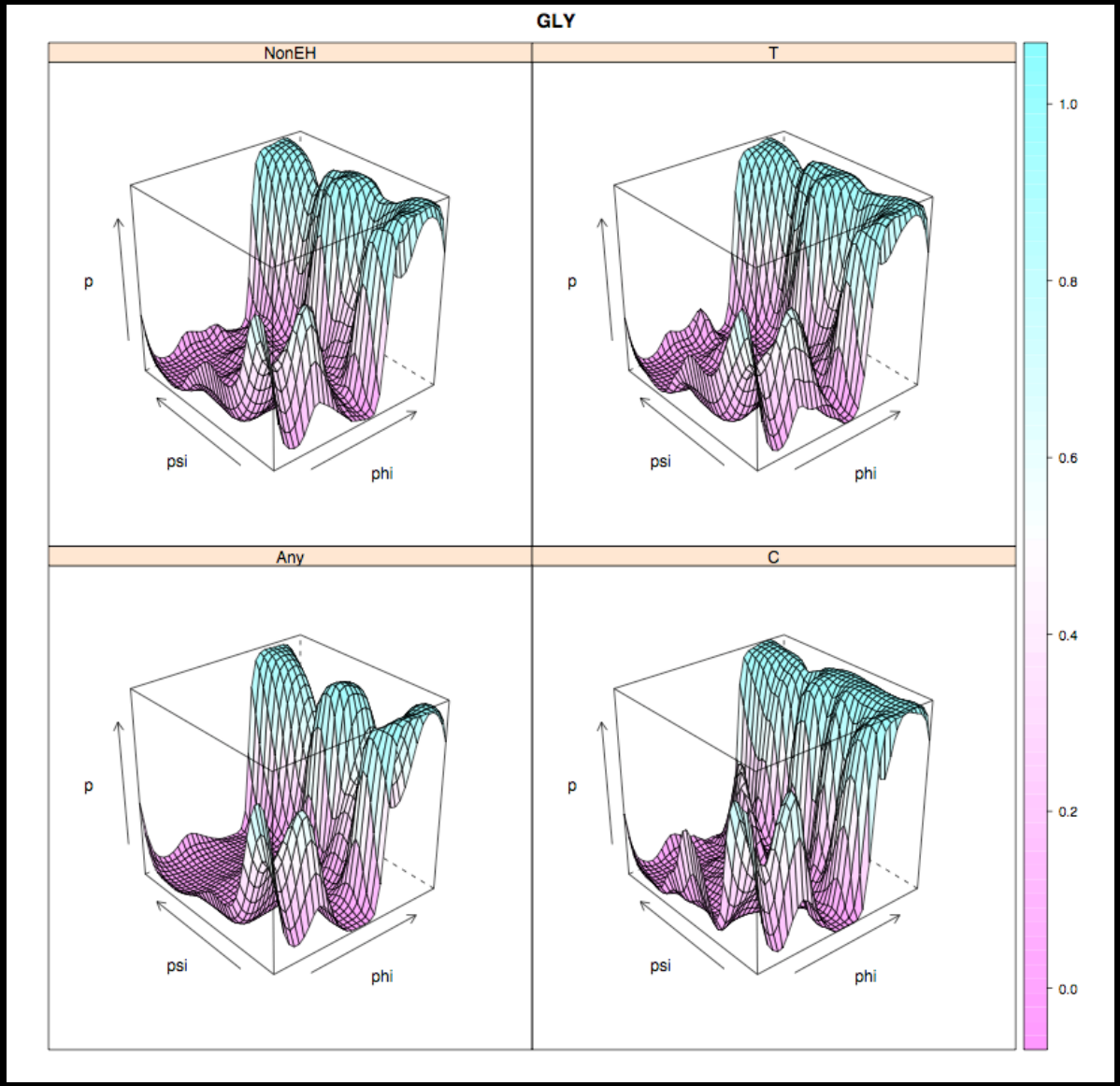
Proline



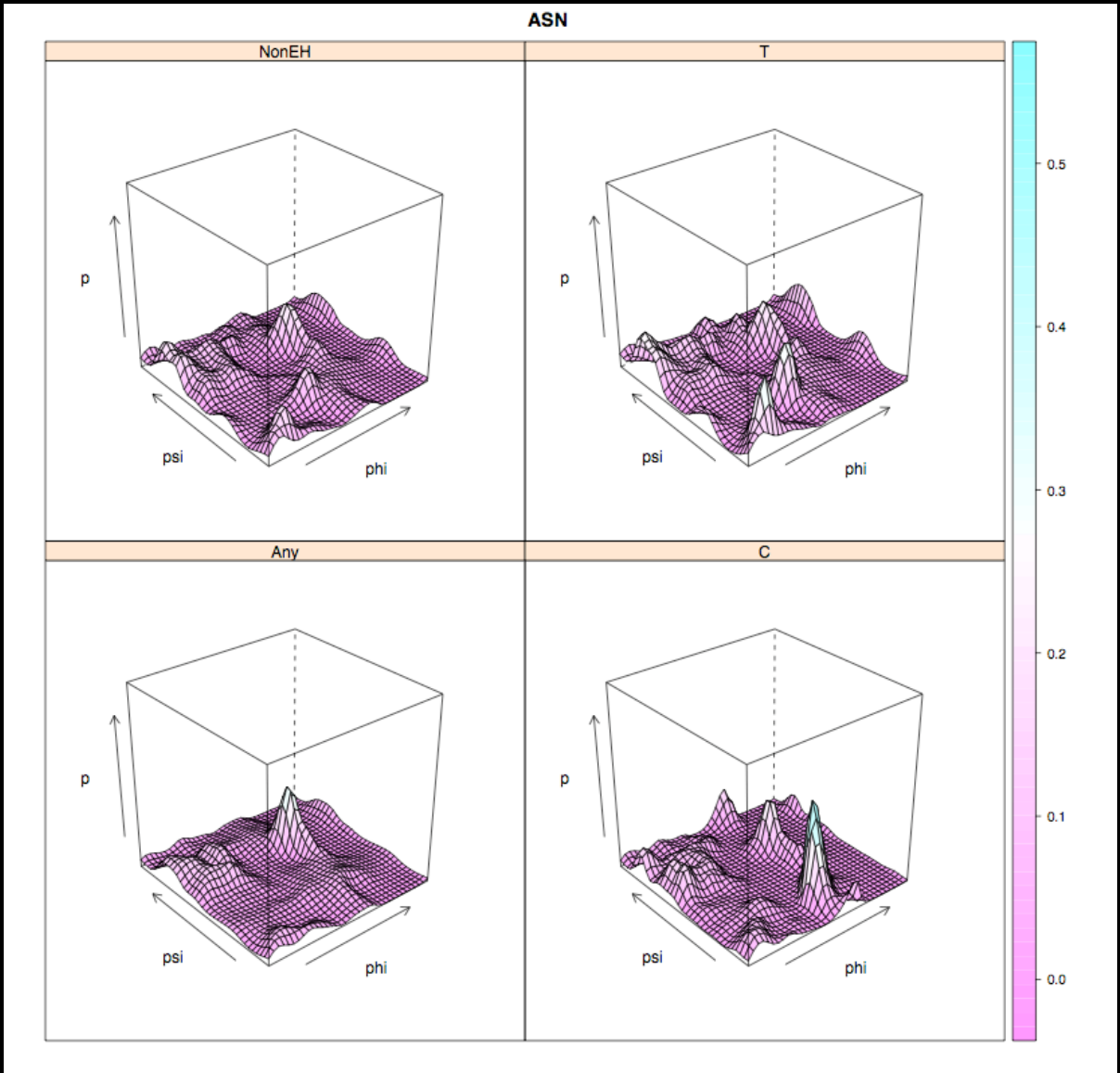
Propensity: Input set -- ALA



Propensity: Input set -- GLY



Propensity: Input set -- ASN



Acknowledgments

Maxim Shapovalov
Guoli Wang

NIH P20 GM76222-04
NIH R01 GM73784-04
NIH R01 GM84453-07

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

SIGN I
RECO
COMM
SIGN I
E-MAIL
PRINT
REPRI