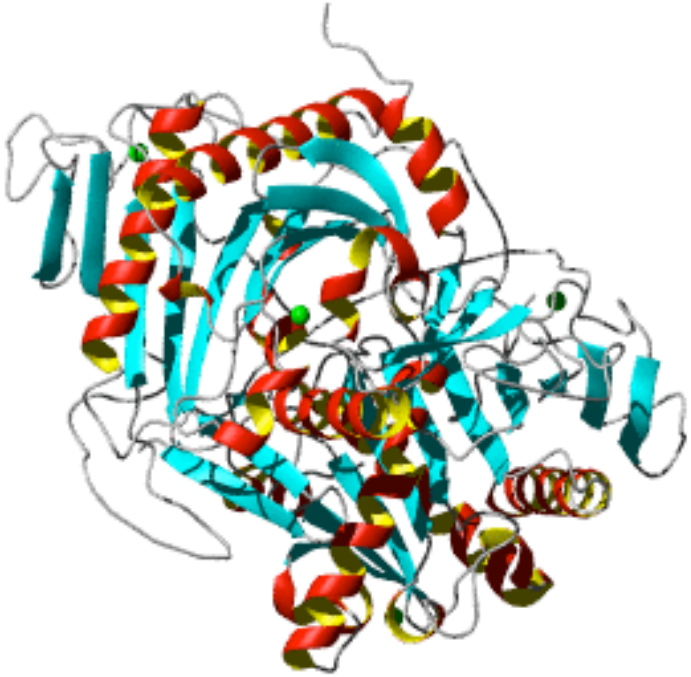


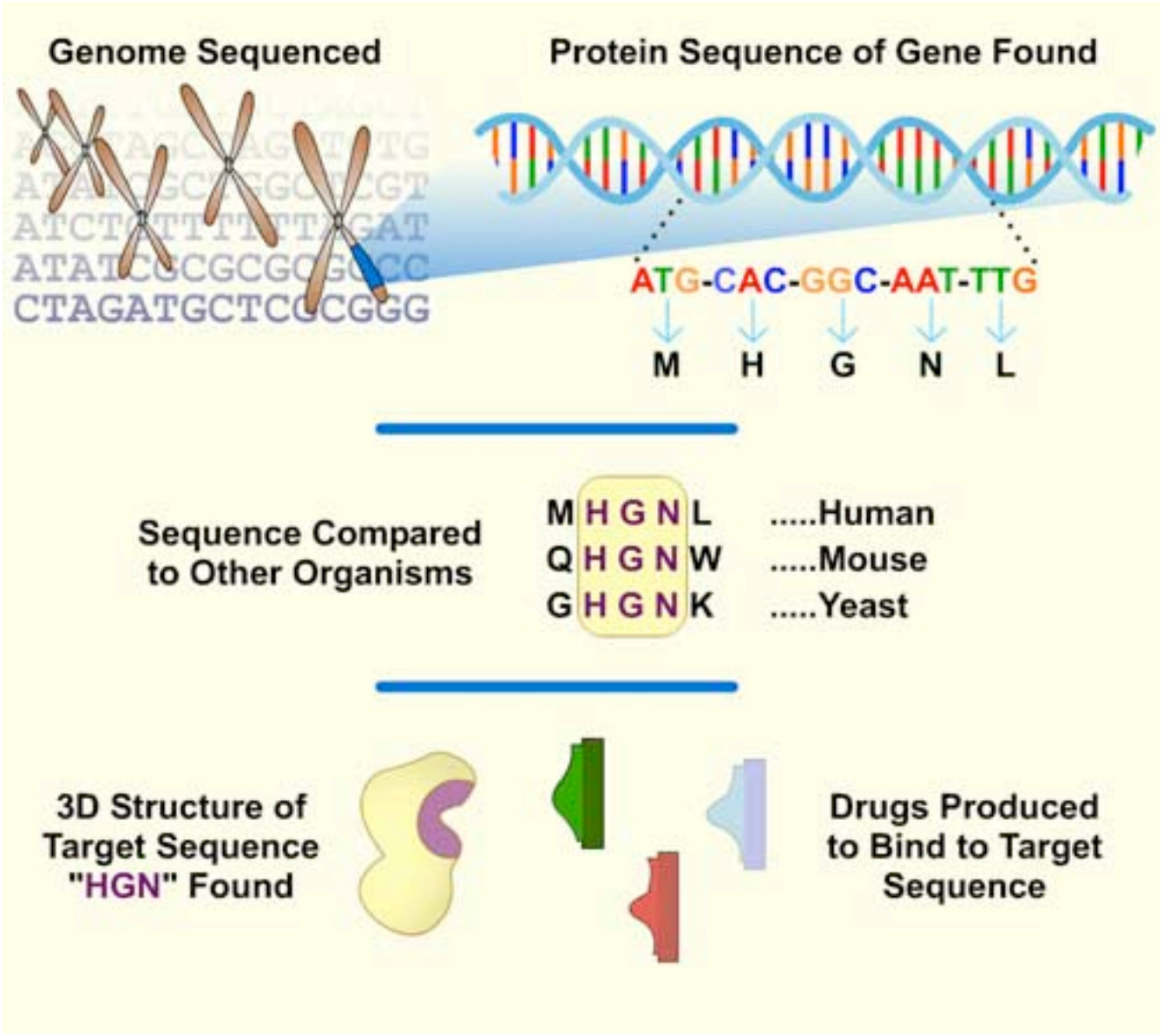
Hands-off Condor for Biology Applications

Hamid R. Eghbalnia

University of Wisconsin-Madison*

*University of Cincinnati starting July 2008

Genomics

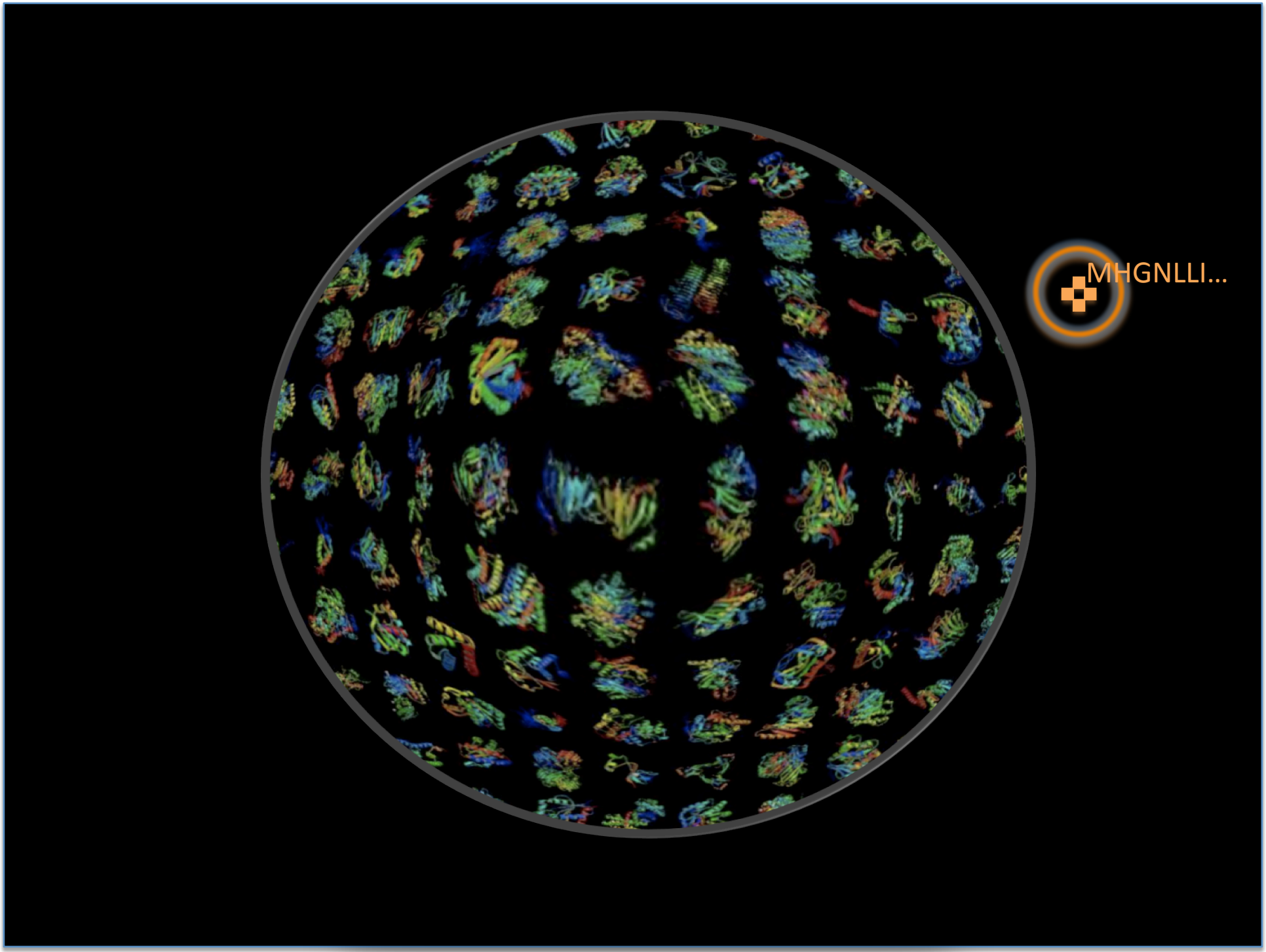

Genome Sequenced
Protein Sequence of Gene Found
ATG-CAC-GGC-AAT-TTG
M H G N L
Sequence Compared to Other Organisms
M H G N L .....Human
Q H G N W .....Mouse
G H G N K .....Yeast
3D Structure of Target Sequence "HGN" Found
Drugs Produced to Bind to Target Sequence

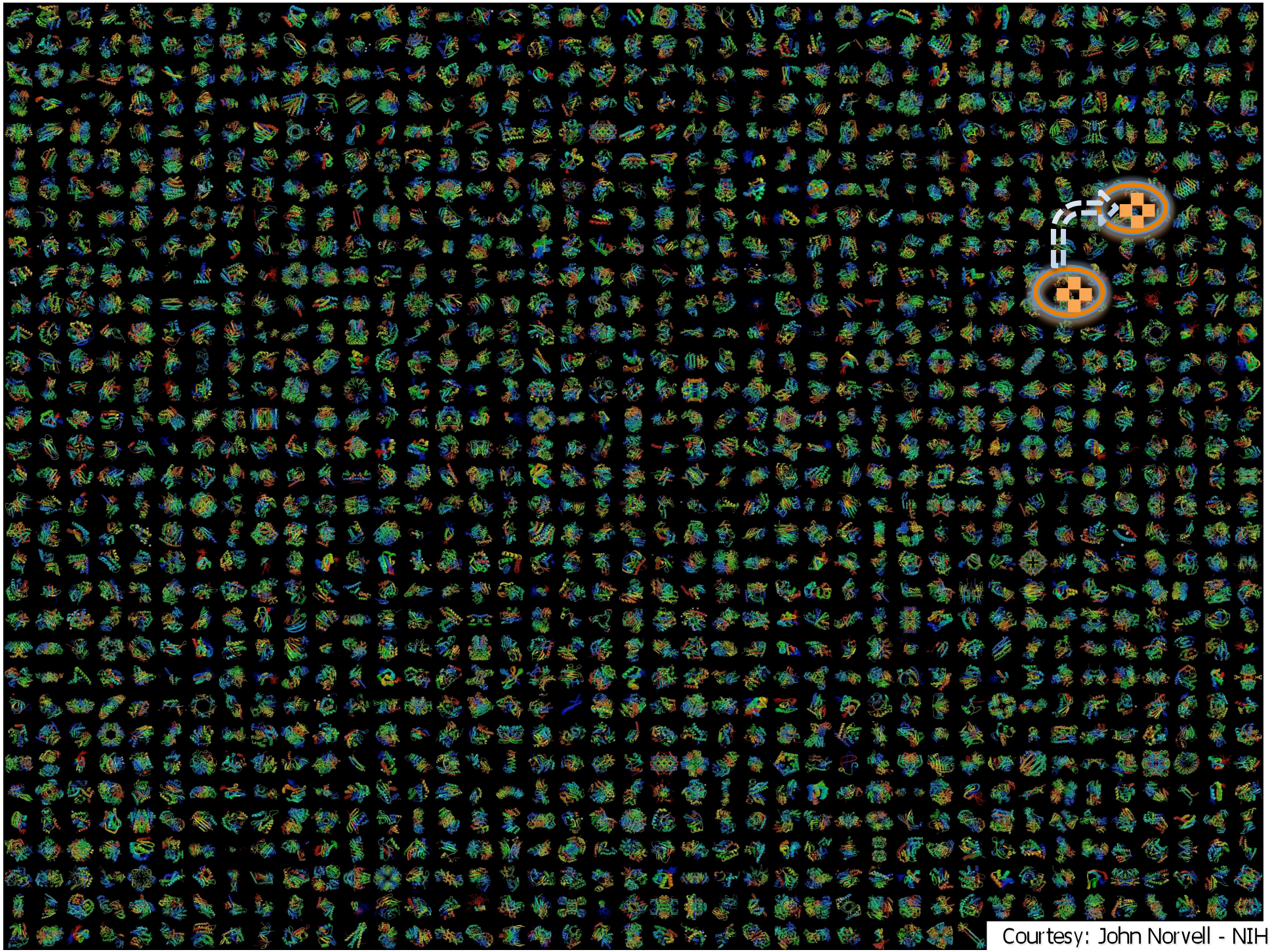
Protein Structure Initiative (PSI) Mission Statement

"To make the three-dimensional atomic level structures of most proteins easily available from knowledge of their corresponding DNA sequences."

Structural genomics

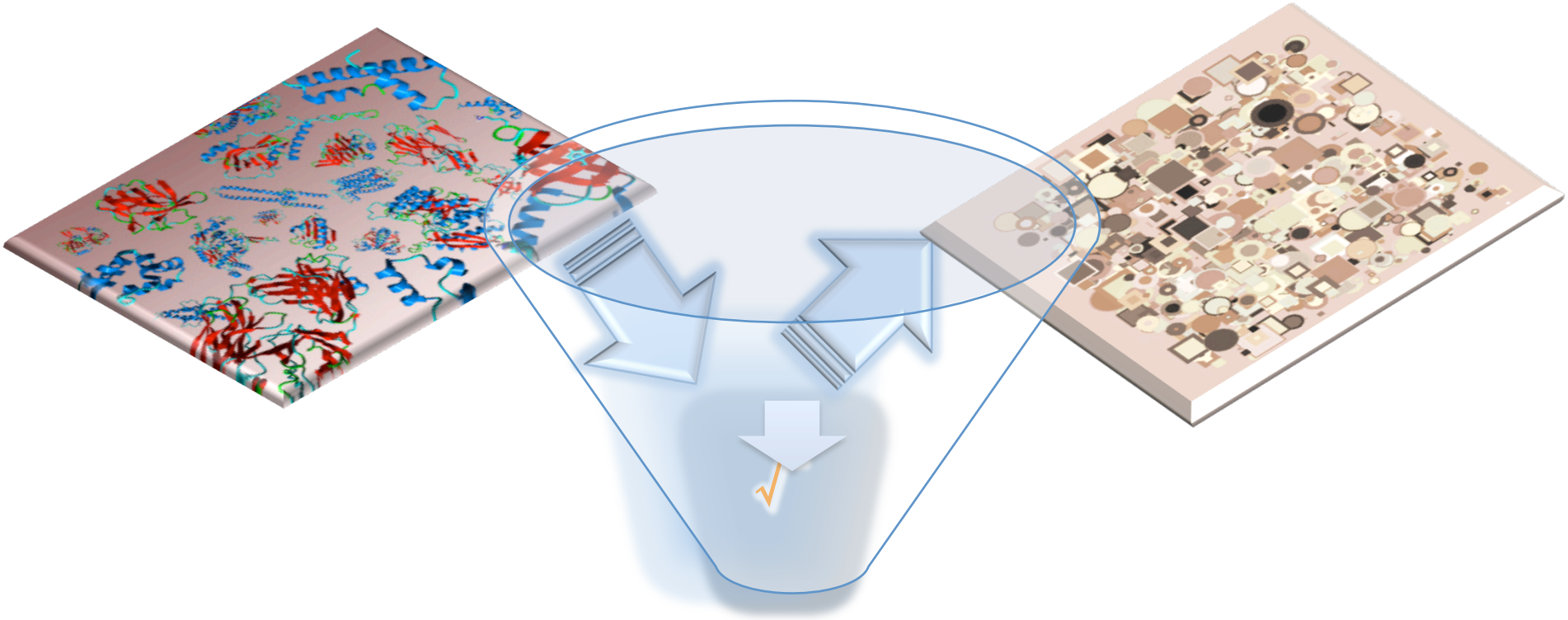
- Experimental determination of **key** protein structures
  - target selection
- **Modeling** members of the larger family
  - Model selection
- **Inferring** protein function
  - Inference
- Other use of the new structures


MHGNLLI...



Courtesy: John Norvell - NIH

Infer protein function from data



How good is structural modeling?

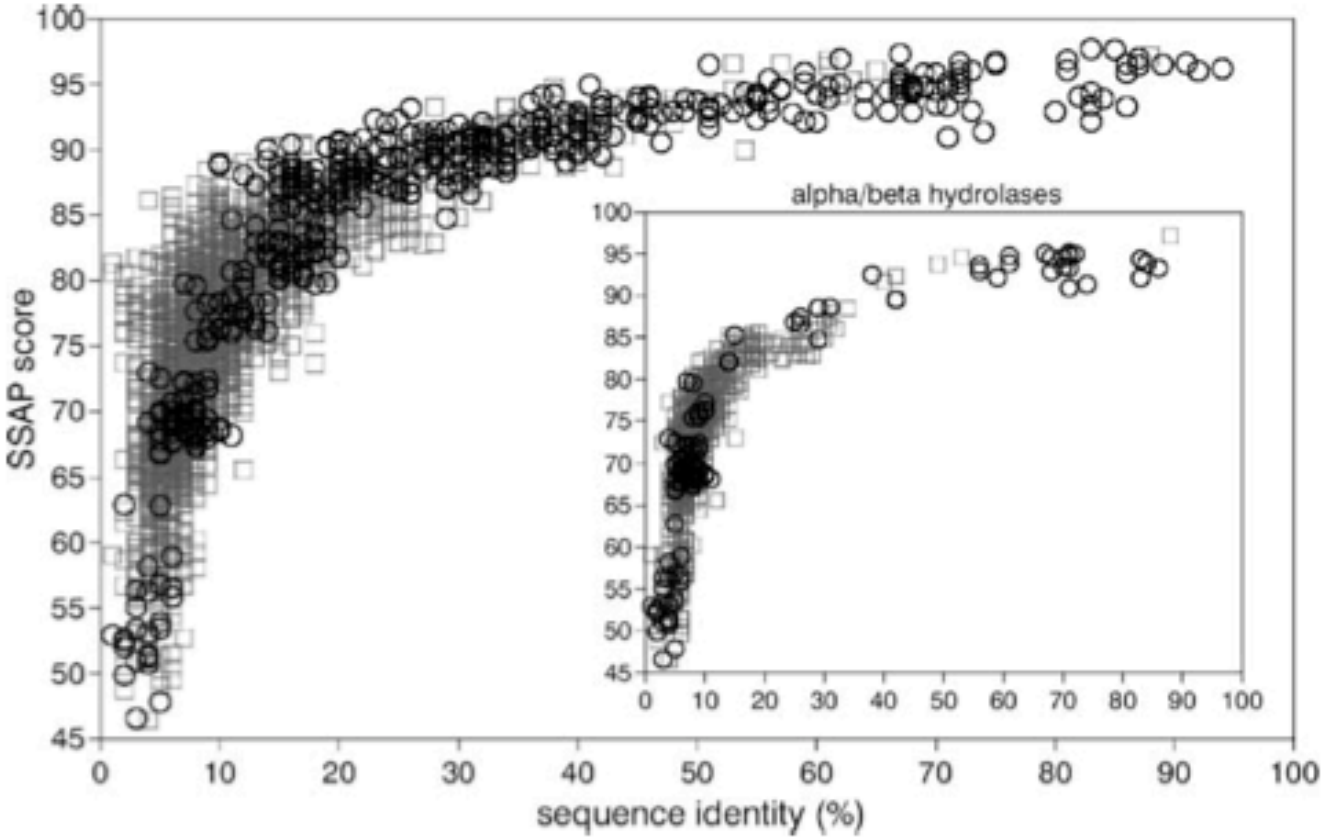
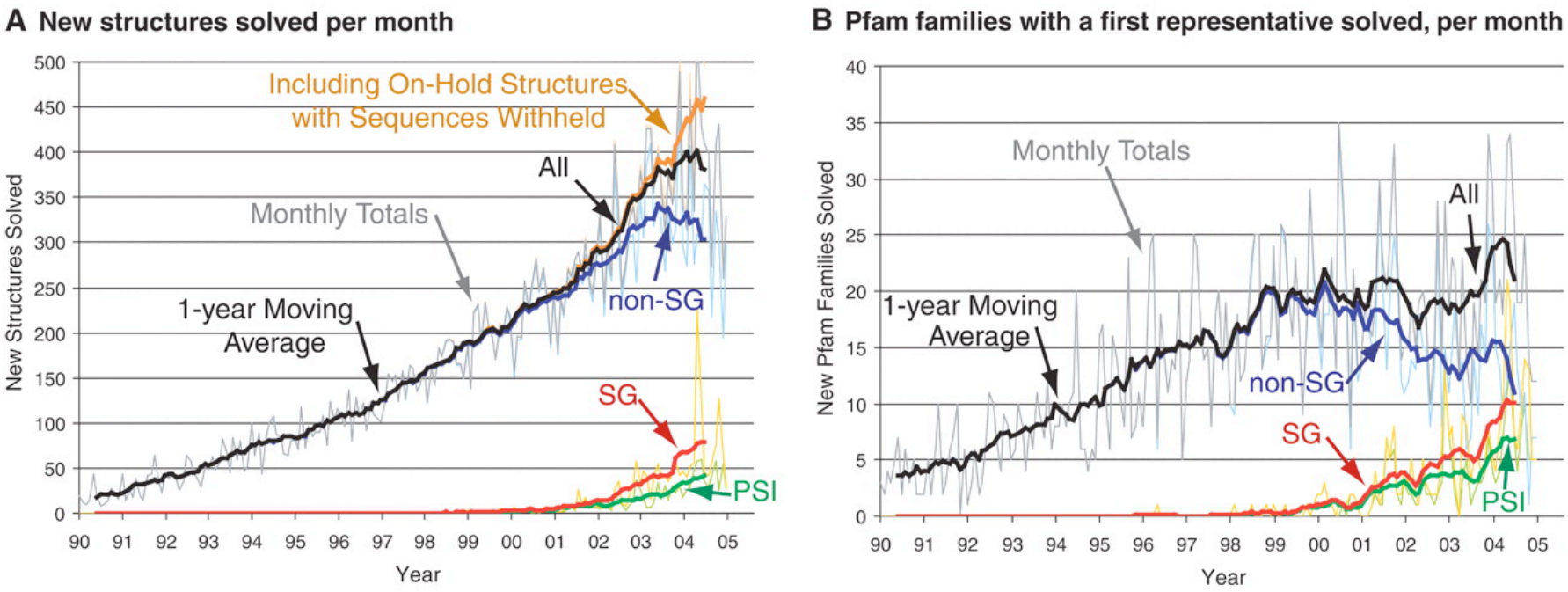


Fig. 1. Correlation between structure similarity (measured by the SSAP structure comparison algorithm, 0–100) and sequence similarity (measured by sequence identity) for all pairs of homologous domain structures in the CATH domain database.

Redfern, Orengo et al., J. Chromatography B 815:97 (2005)

Genseration of new structures



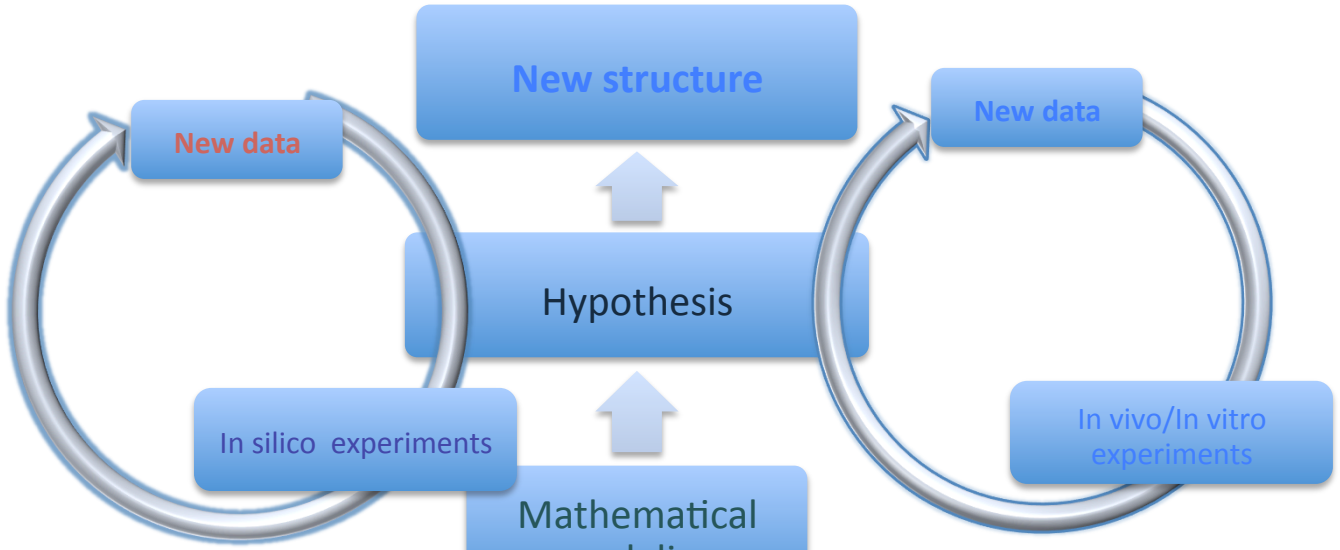
Chandonia and Brenner, Science 311:347 2006.

Sequences and Folds

- ~100,000 families of proteins that cannot be reliably modeled at present

- The structure universe of membrane proteins, and larger more dynamics complexes, remain mostly “unknown”

Structure from sparse data



Putative structural alterations

PDB , BMRB , …

Prior knowledge built using experimental data

Structural changes

Biological structure

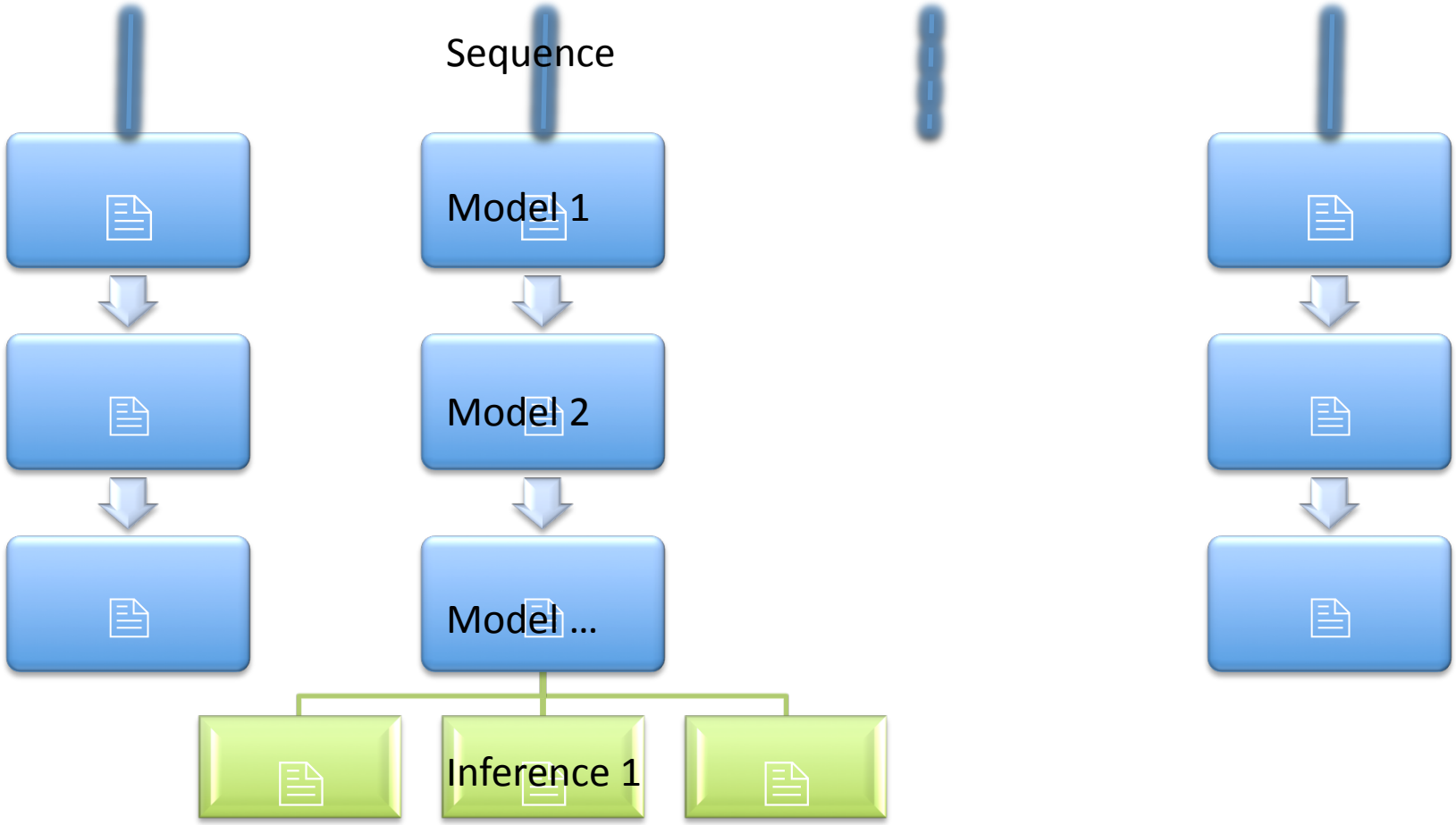
Questions of reliability

- Accuracy
  - How accurate is software package X in modeling a portion of my problem?
- Precision
  - How precise is software package X in modeling a portion of my problem?
- Extension
  - I use X to model one part, and Y to model another part, what is my accuracy and precision for X+Y?

Questions of predictability

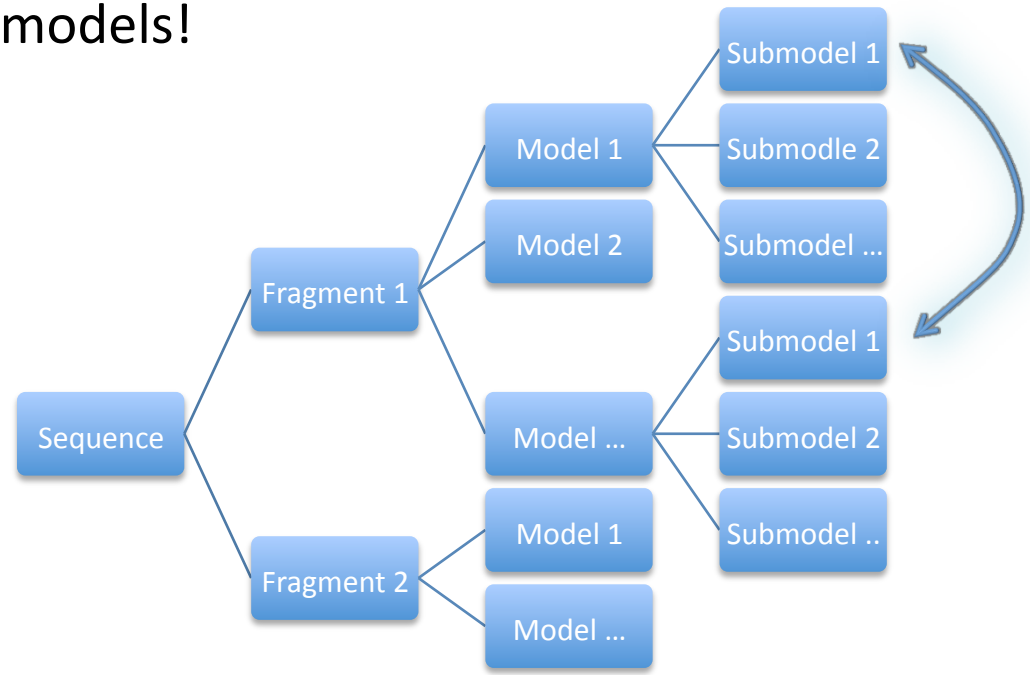
- How does one combine tools to achieve "reliable" results?
  - Directly using existing tools
  - In combination with "home grown" tools
- How much inference can we afford?

Computing at multiple granularities

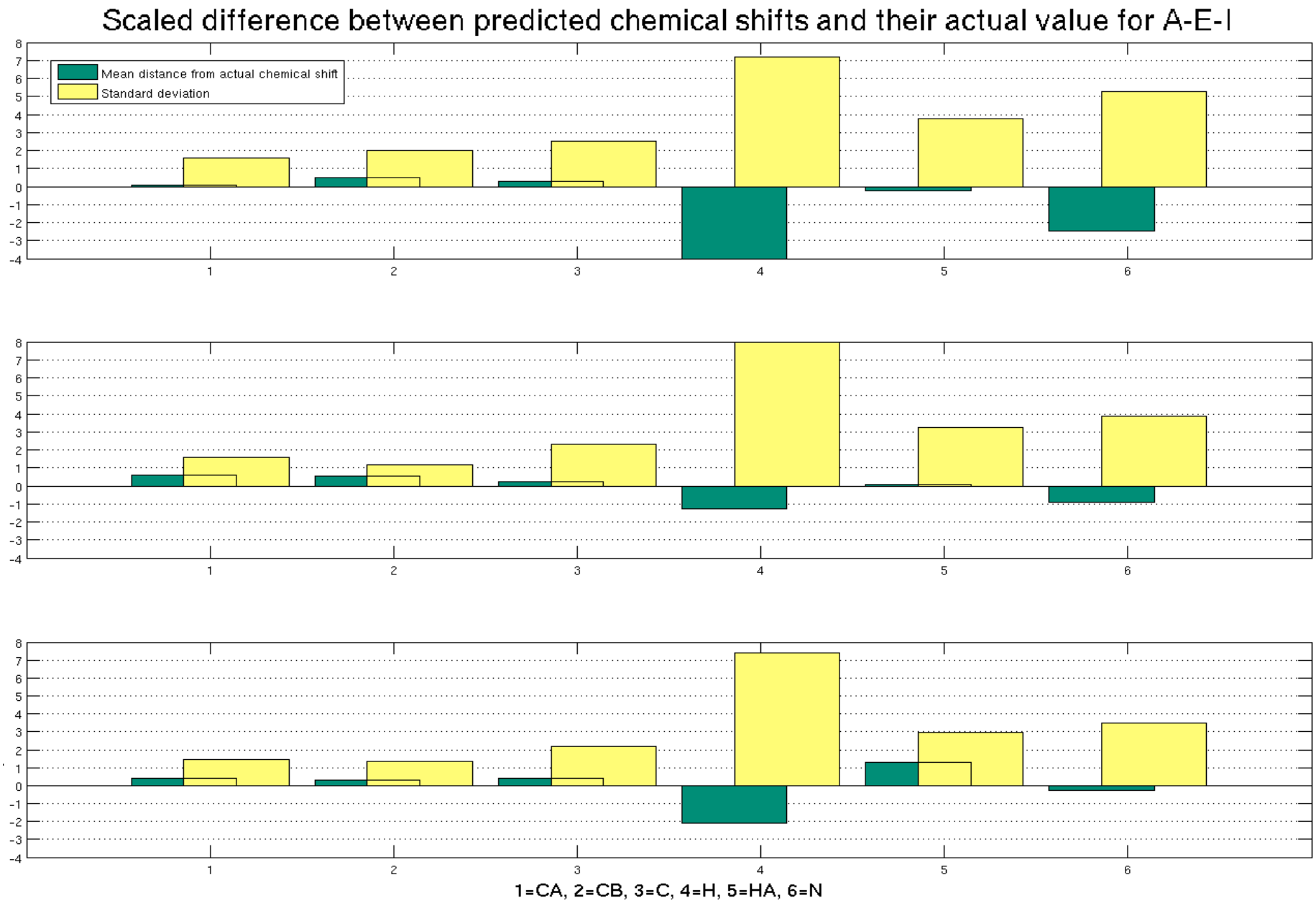


Example

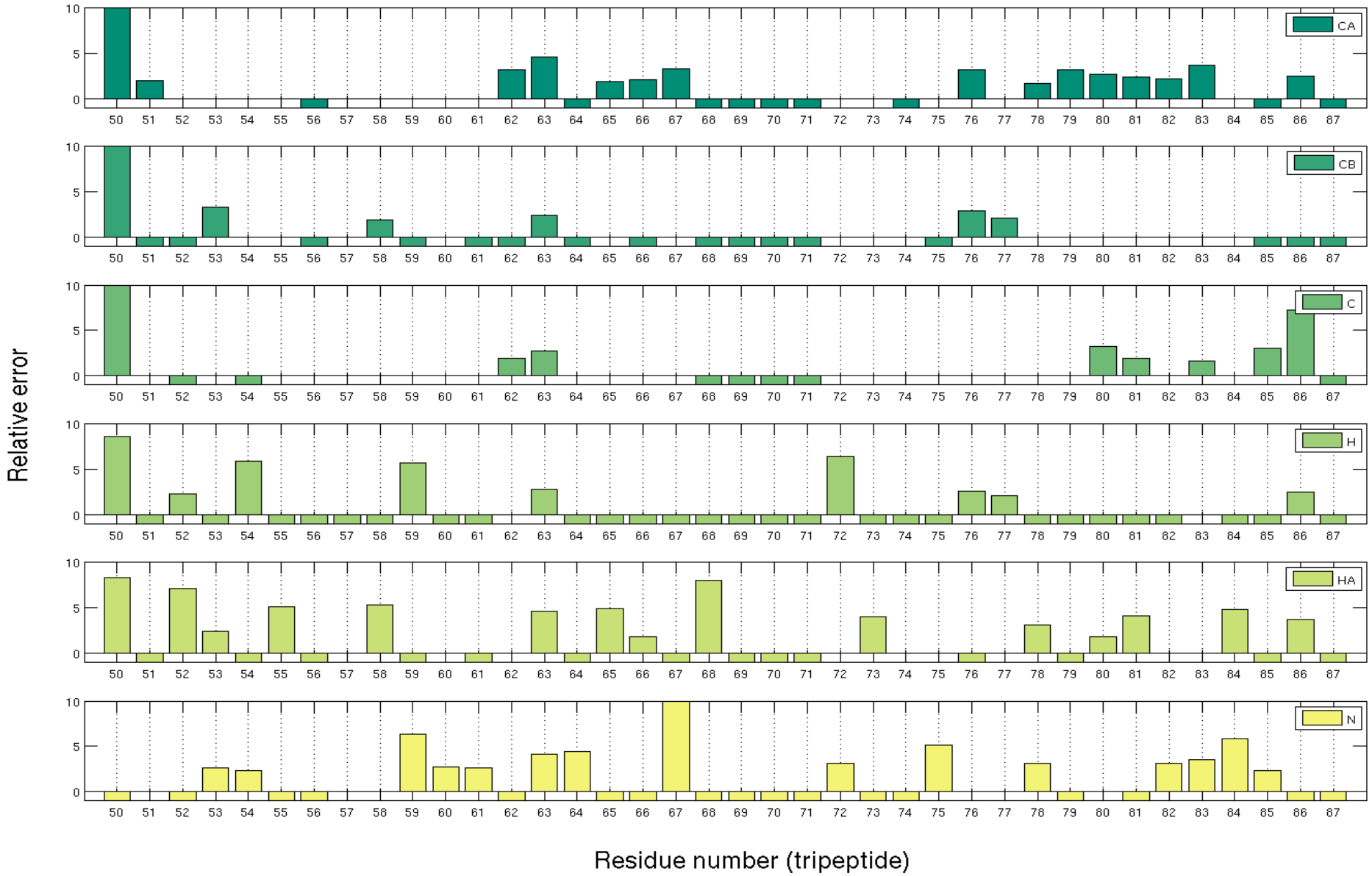
- Starting with a protein sequence MLLIGGP..
- Generate a fragment library L_1 L_2 … L_n
  - Generate multiple fragment libraries
  - For each library, generate N structure models $S_1 \ldots S_m$
  - For each model compare back-calculated properties to known experimental data $(L_k,S_j) \rightarrow P_{kj}$ $|P_{kj}-E_{kj}|=?$
    - Do this for each model for each library
- Select models
  - Need to cross-check models!
- Iterate



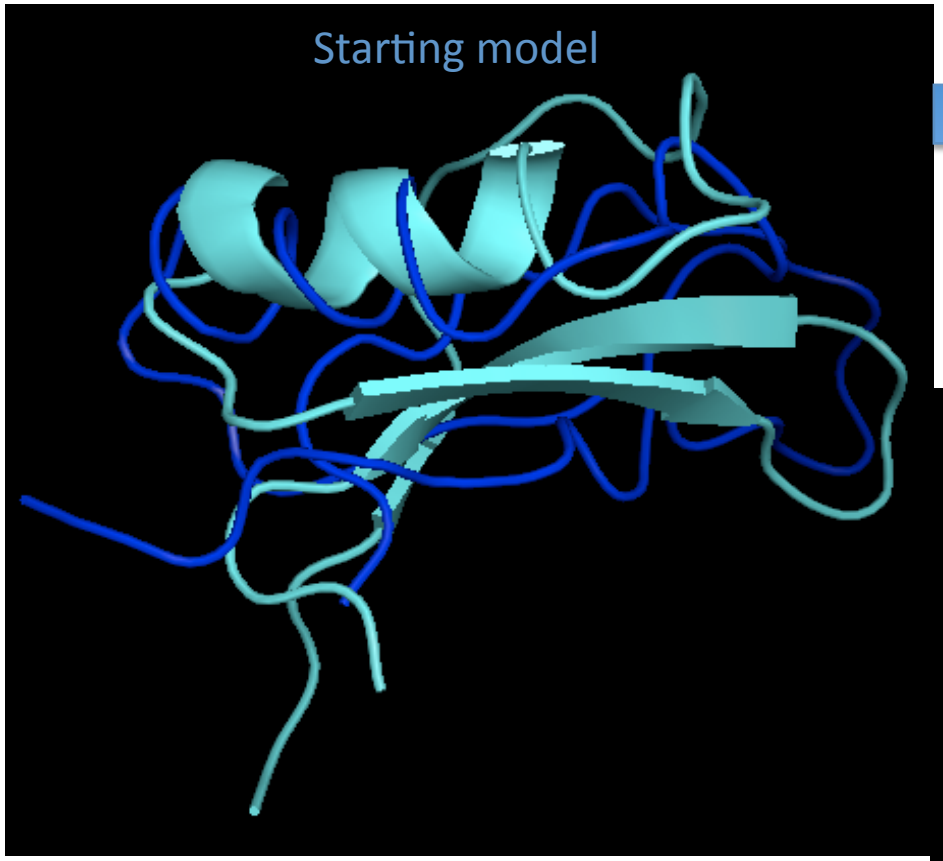
Predicting chemical shifts from structure



Example: P-gamma

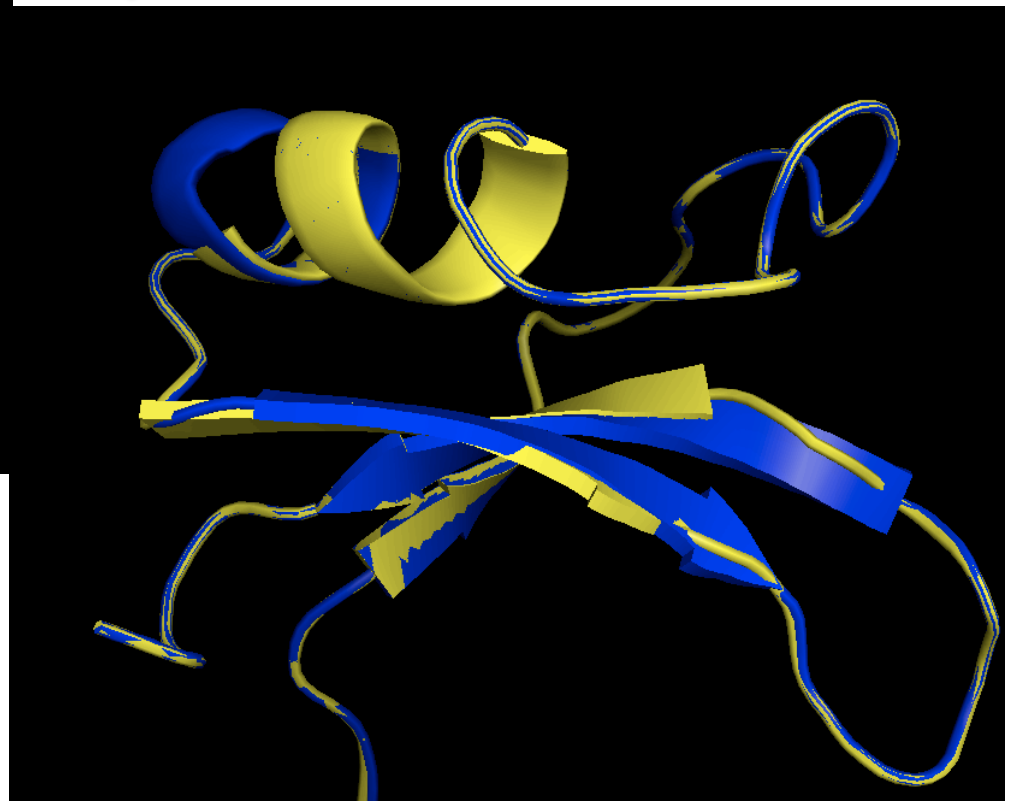


Example: Structure from sparse data



Model vs. pdb structure

Refined with force field



Acknowledgments

Arash Bahrami
Marco Tonelli
Fariba Assadi-Porter
Mohammad Sedighi
John Markley
Eldon Ulrich
Zach Miller
Miron Livny

NIH K22 LM8992

NIH P41 RR02301