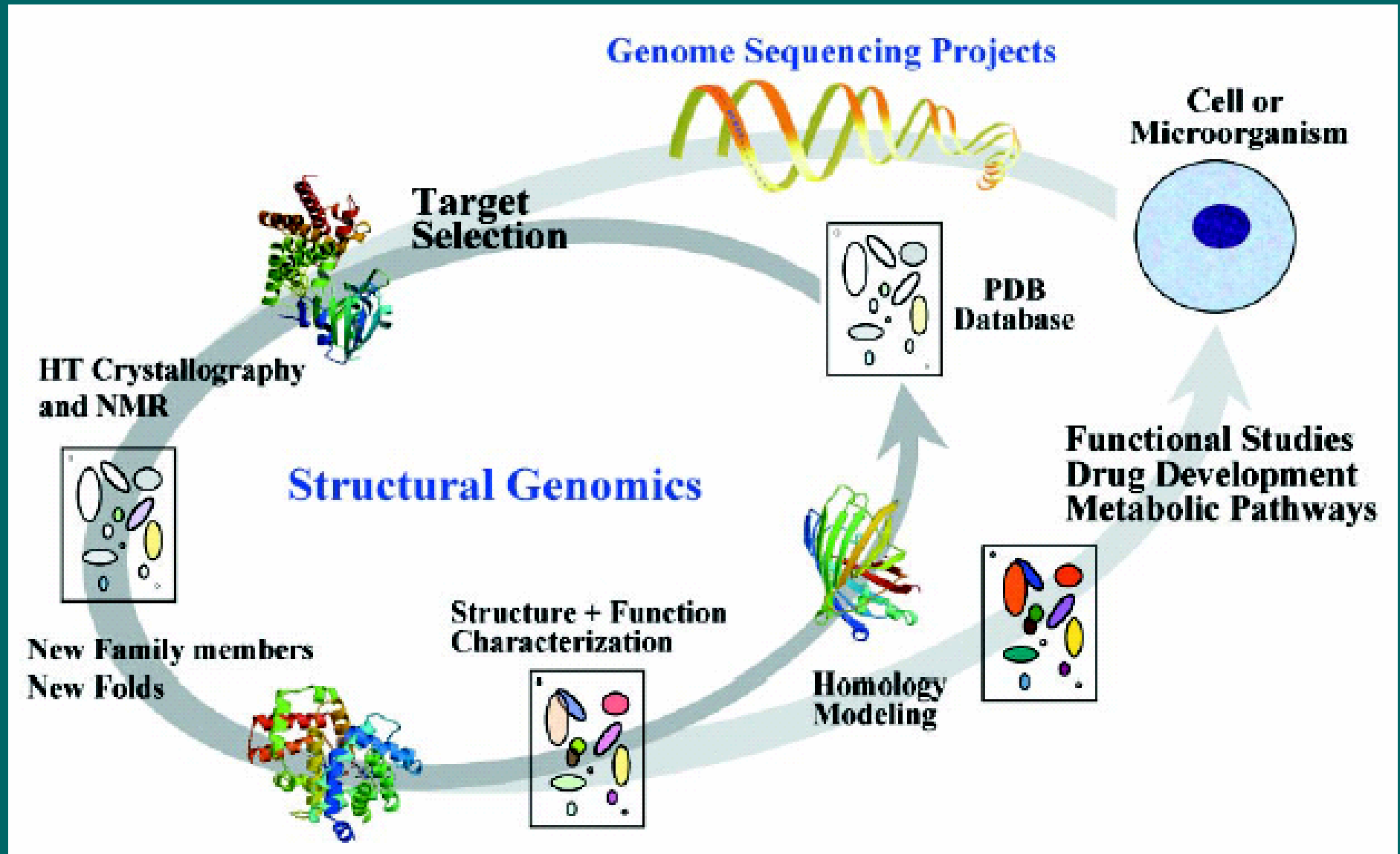


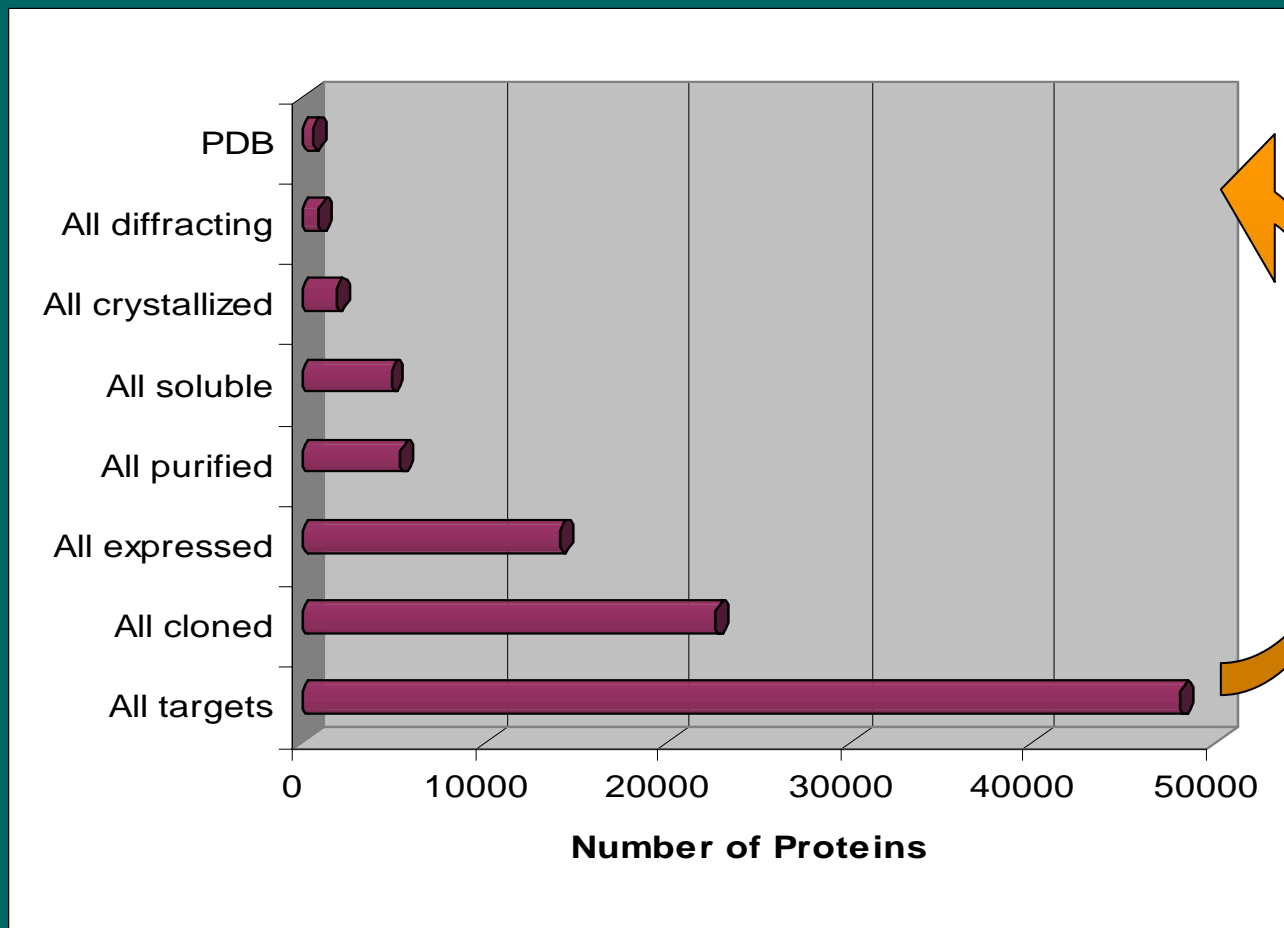
Will my protein crystallize?

Dmitrij Frishman
Technische Universität München

Target selection and structural genomics



Structural genomics pipeline



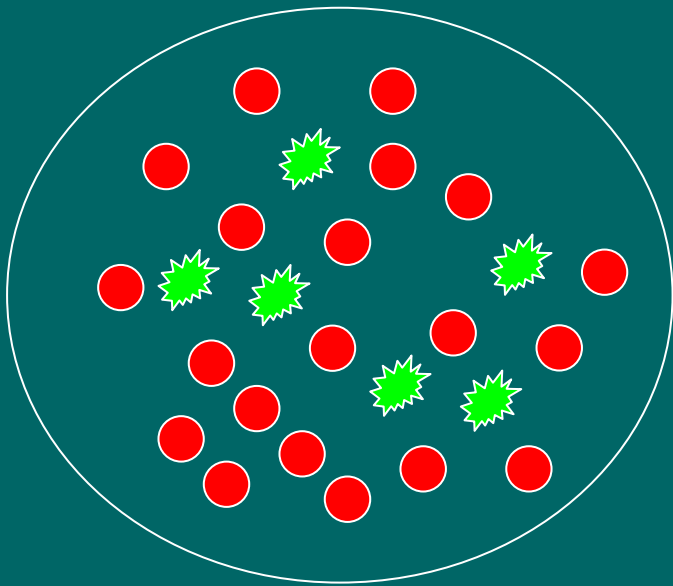
2.6%

Quest for the negative examples

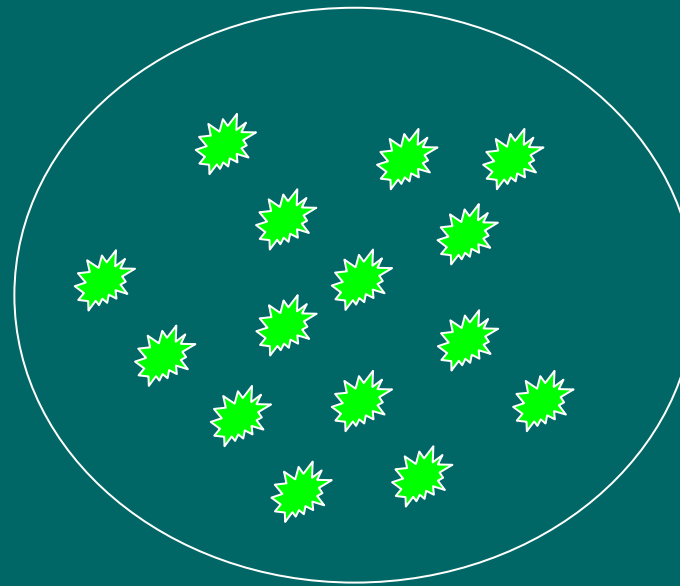
Ideally we would need a sufficiently large set of proteins experimentally shown to be recalcitrant for crystallization

- Unknown structure
- ★ Known structure

Swiss-Prot



PDB

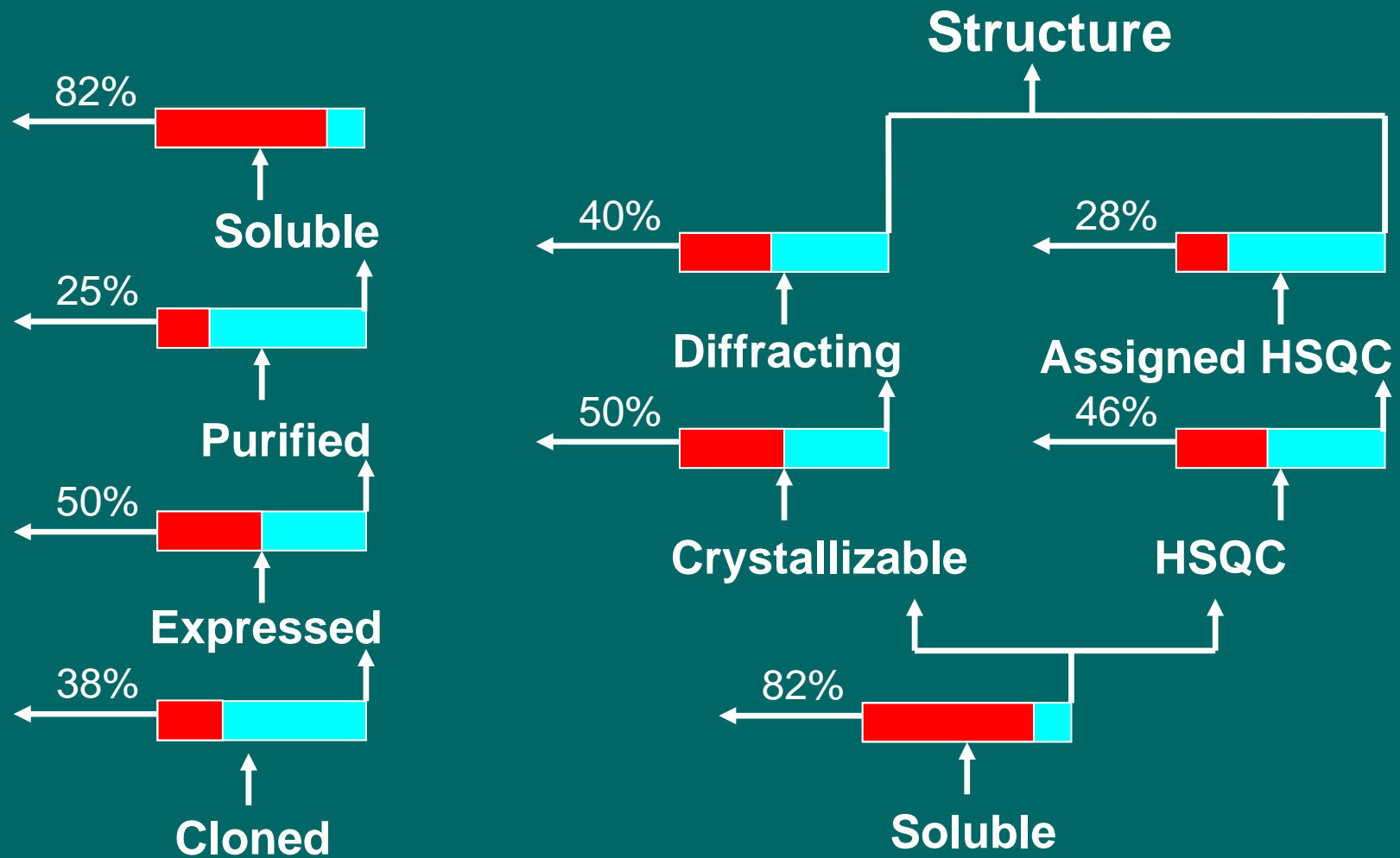


Crystallizability

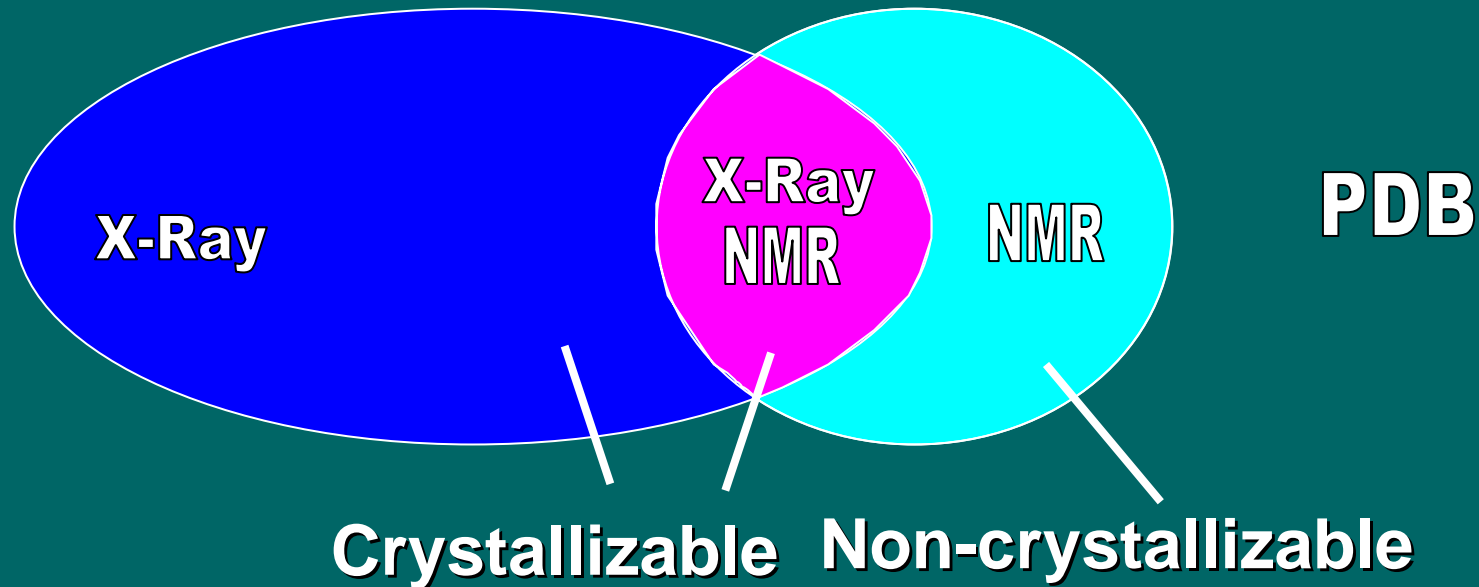
- The propensity of a given protein to yield a well-diffracting crystal under a given range of experimental conditions
 - individual protein trait
 - correlates with its amino acid sequence?
- The ability to predict experimentally tractable proteins from sequence alone would be beneficial for designing rational target selection strategies in structural genomics

Structural genomics pipeline

Recalcitrant proteins; branching picture



Idea: use NMR structures as negative set

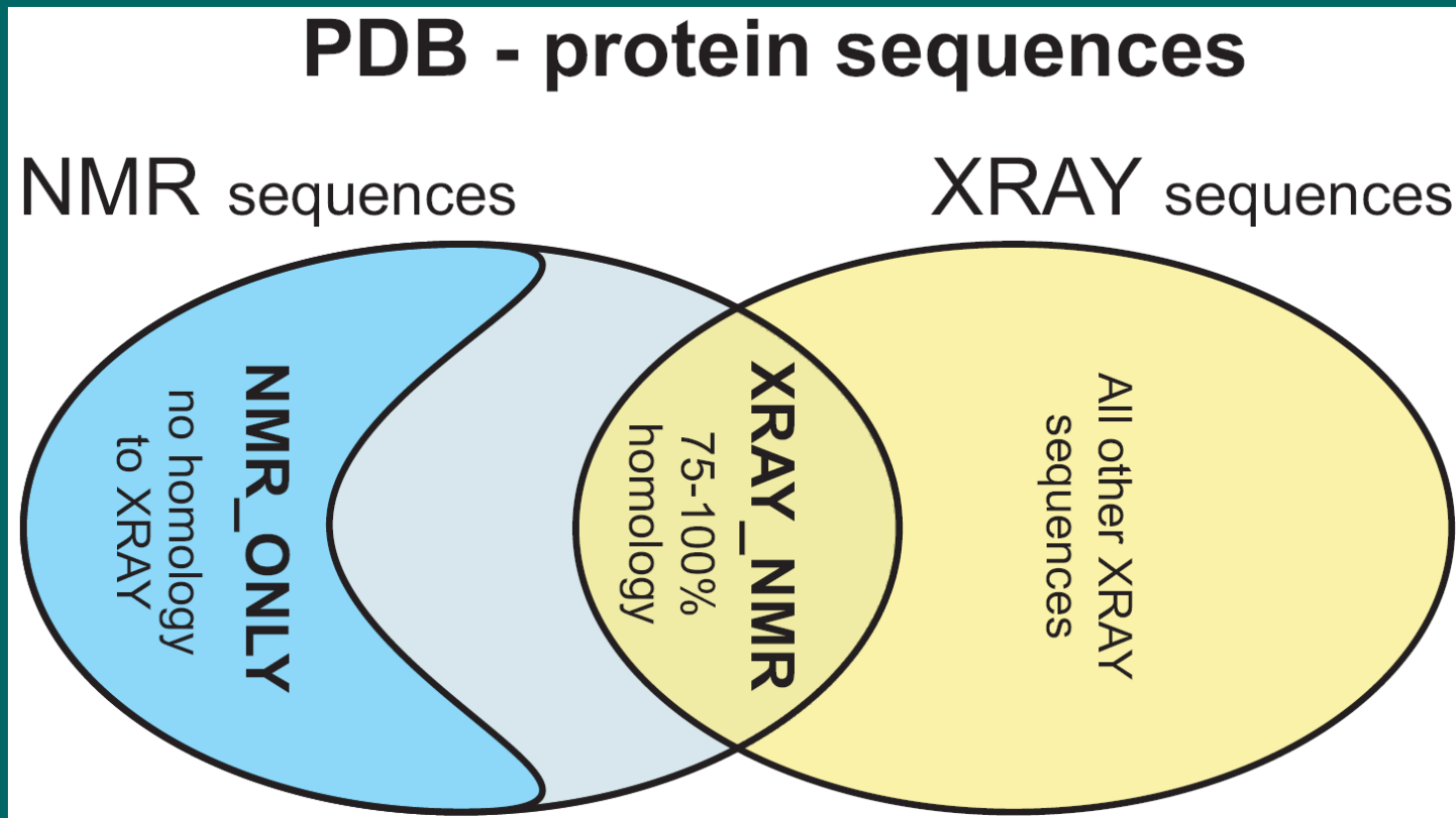


„Solution state NMR will also have a complementary role in post-genomic analysis, particularly considering that many protein targets do not provide crystals suitable for crystallographic analysis ...“

„NMR is particularly valuable in structural genomics for analyzing protein structures that are outside the scope of crystallographic studies ...“

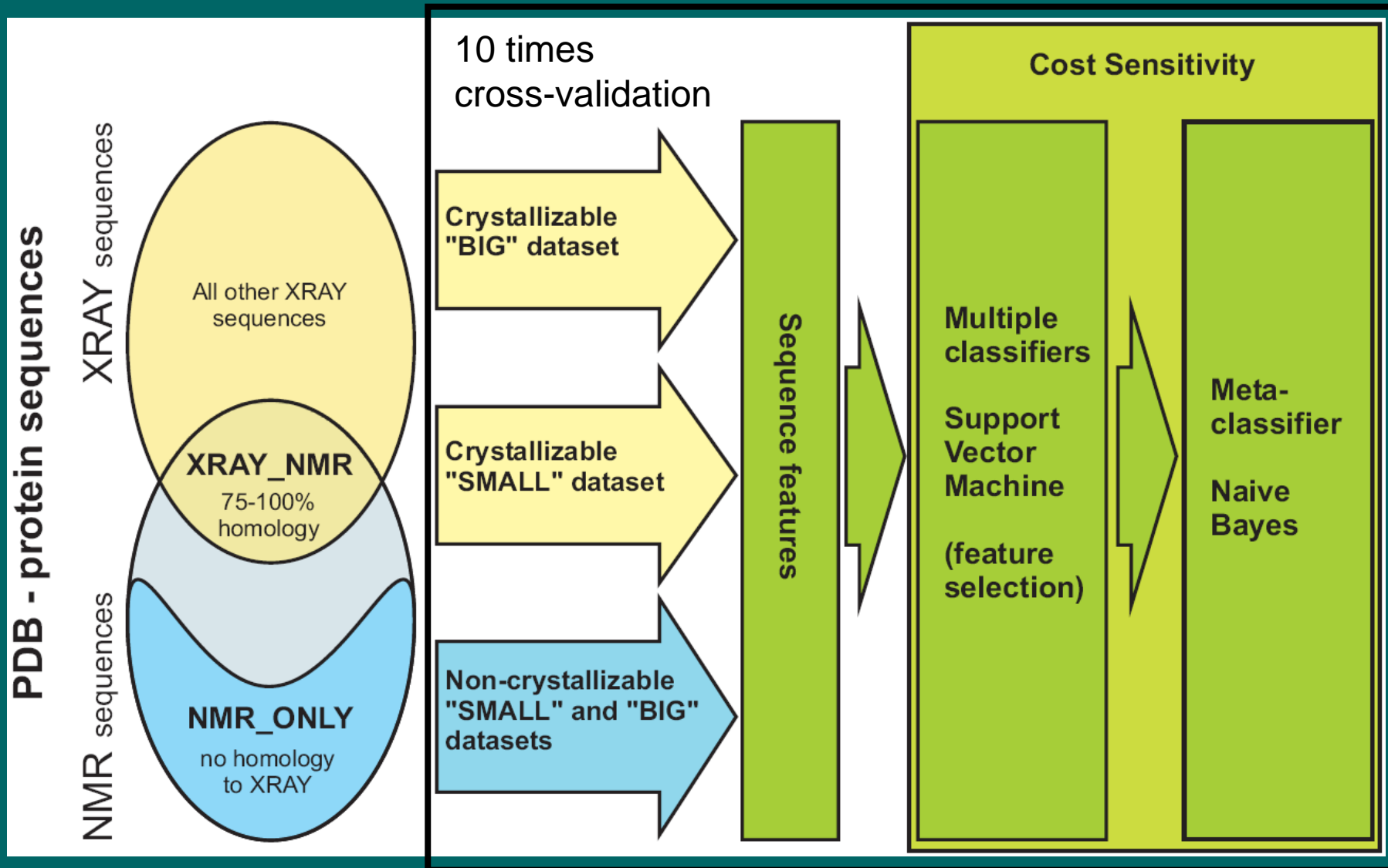
Montelione G, Zheng D, Huang YJ, Gunsalus KC and Szyperski T. Protein NMR spectroscopy in structural genomics. Nature Struct Biol., Structural Genomics Supplement, November 2000, 982-985

Building the training dataset

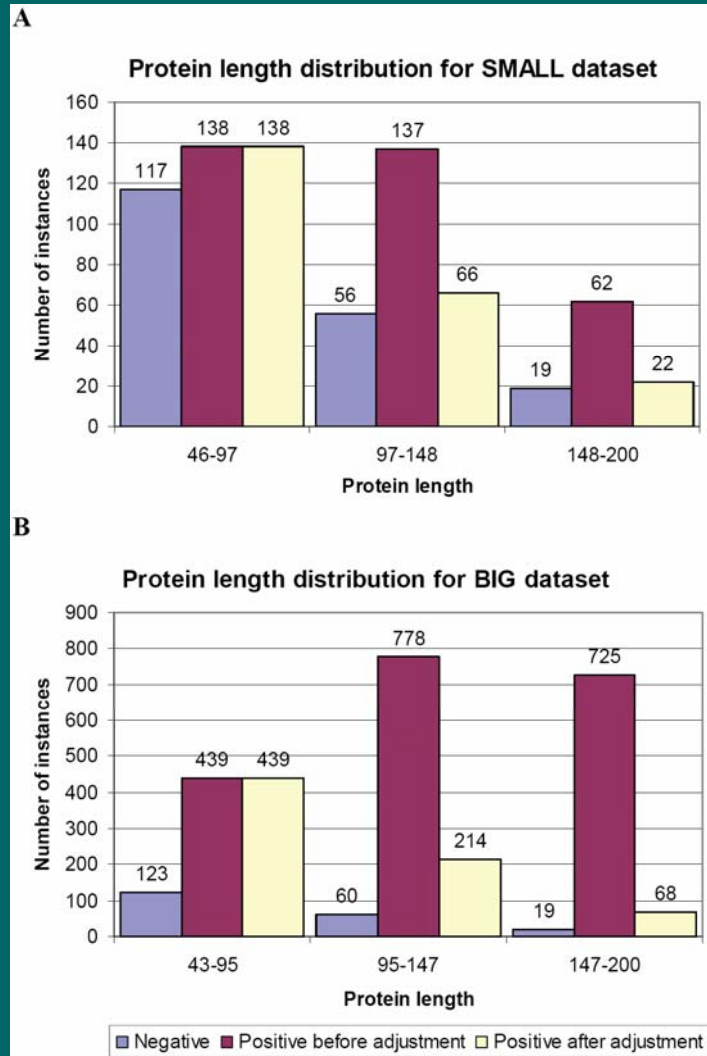


- 1) proteins with X-ray structures, which are hence crystallizable;
- 2) proteins with NMR (Nuclear Magnetic Resonance) structures that also have an X-ray structure, or bear a high sequence similarity (blast 75-100% identity, 10% length difference) to proteins with X-ray structures, and;
- 3) proteins with NMR structures only (based on BLAST bit score cutoff 30) .

Method



Excluding obvious differences



Protein length distributions of datasets were adjusted to avoid predictions prejudiced by protein size.

To obtain datasets not biased with respect to sequence length we constructed sub-samples from the original three datasets with comparable size distributions

Kolmogorov-Smirnov test was performed, with the null hypothesis being that their length distributions are identical.

Amino acid groupings according to different hydrophobicity scales

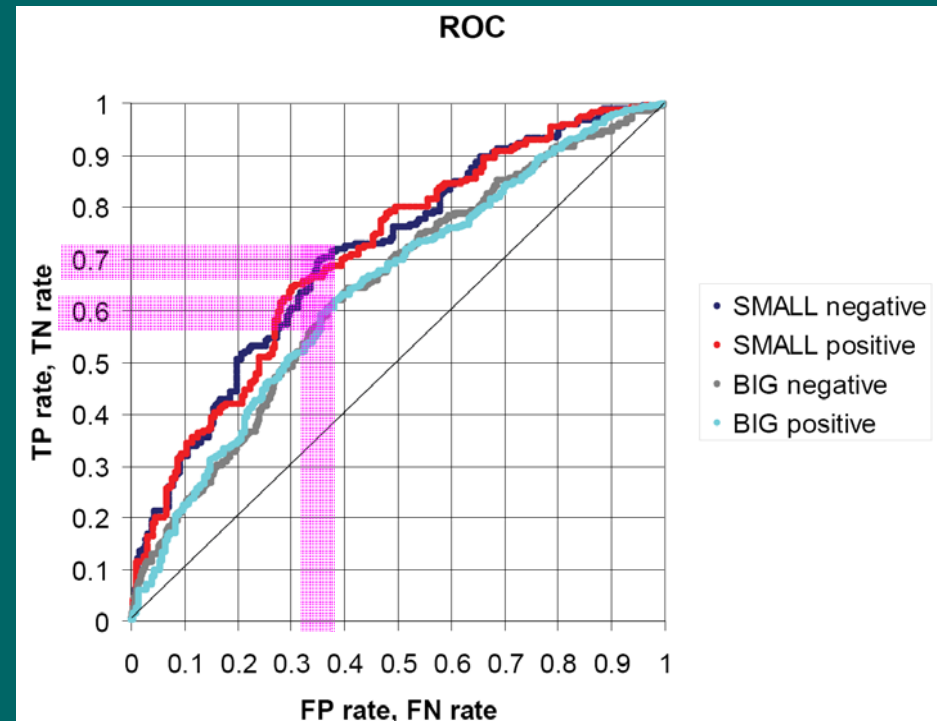
Group/Scale	GES	Kyte & Doolittle	Rose
-	F, M, I, L, V, C, W, A, T, G, S (hydrophobic)	R, N, D, Q, E, H, K (hydrophilic)	R, N, D, Q, E, K, P, S (hydrophilic)
0	P, Y, H, Q, N (neutral)	G, P, S, T, W, Y (neutral)	A, G, H, T, Y (neutral)
+	E, K, D, R (hydrophilic)	A, C, I, L, M, F, V (hydrophobic)	C, I, L, M, F, W, V (hydrophobic)

Amino acid groupings according to different similarity matrices

Grouping schema	Similarity matrix	Clustering method	Level of dendrogram	Amino acid groups
G1	All-matrix	UPGMA	Up to 4 nodes from root	(A); (W, Y); (M, L, V, I, F); (C); (H); (E, Q, R, K); (N, D); (T, S, G); (P)
G2	All-matrix	UPGMA	Up to 5 nodes from root	(A); (Y); (W); (L, V, I, F); (M); (C); (H); (R, K); (Q); (E); (D); (N); (G); (T, S); (P)
G3	All-matrix	UPGMA	Up to 2 nodes from leaves	(P); (T, S, G); (N, D); (E); (Q, R, K); (H); (C); (M); (L, V, I, F); (Y, W); (A)
G4	Genetic code matrix	Neighbor joining	Up to 1 node from leaves	(W); (C); (G); (R); (Q, E); (A, V); (P); (L); (K, M); (T); (I); (S); (F, Y); (N); (H, D)
G5	Genetic code matrix	Minimal evolution	Up to 2 nodes from leaves	(G, C, W); (R); (S); (N, D, H); (F, Y); (L, P); (A, V); (Q, E); (T); (I); (K, M)
G6	Mutation cost matrix	UPGMA	Up to 2 nodes from leaves	(C); (P); (L, M); (I, V); (F); (W, Y); (G); (A); (T, S, H); (N, D); (K, R); (E, Q)
G7	All-matrix	UPGMA	3 groups	(N, D, G, P, S, T); (A, C, I, L, M, F, W, Y, V); (R, Q, E, H, K)

Classifier evaluation

- The meta-classifier had an accuracy of 67% and 60.9% for the SMALL and BIG datasets, respectively
- Classification based solely on sequence length information achieved a maximum accuracy of 53.4% (SMALL) and 50.6% (BIG)



Acknowledgments



Pawel Smialowsky



Antonio Martin



Jürgen Cox



Thorsten Schmidt



Andreas Kirschner