Predicting Secondary Structure of All-Helical Proteins Using Hidden Markov Support Vector Machines

Blaise Gassend, Charles W. O'Donnell,
William Thies, Andrew Lee,
Marten van Dijk, and Srinivas Devadas

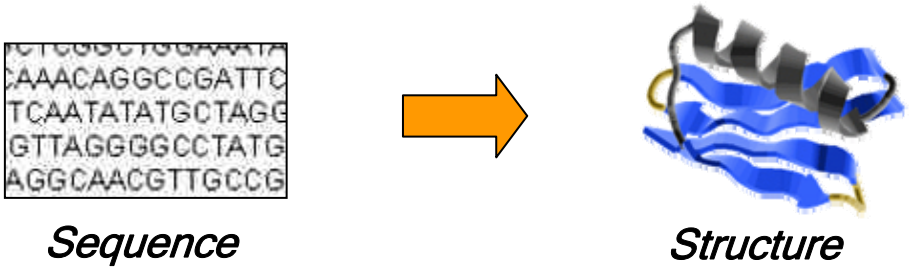
Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

Workshop on Pattern Recognition in Bioinformatics – August 20, 2006

Protein Structure Prediction

- **Classical problem: given sequence, predict structure**



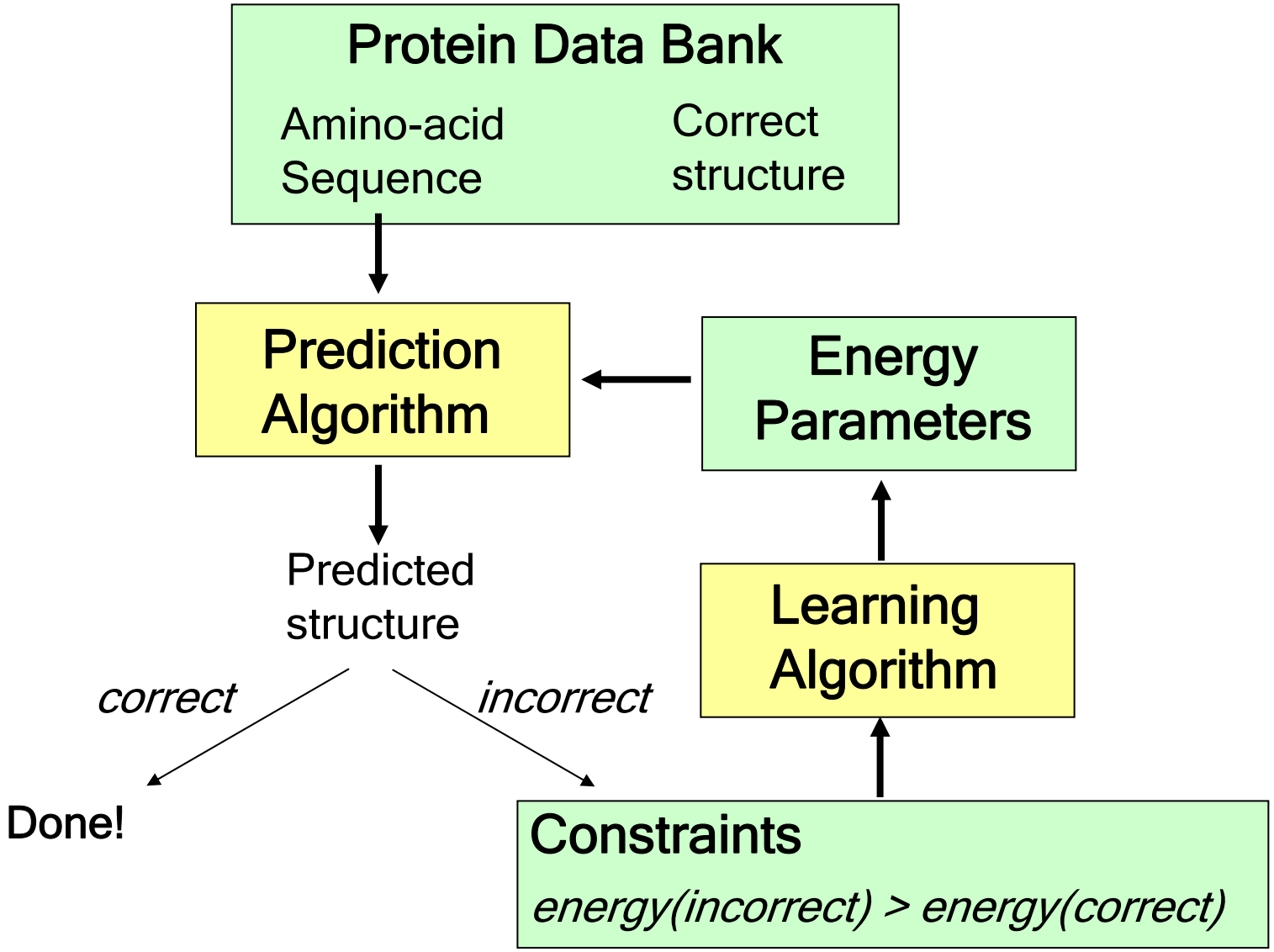
Sequence *Structure*

- **High-level approaches**

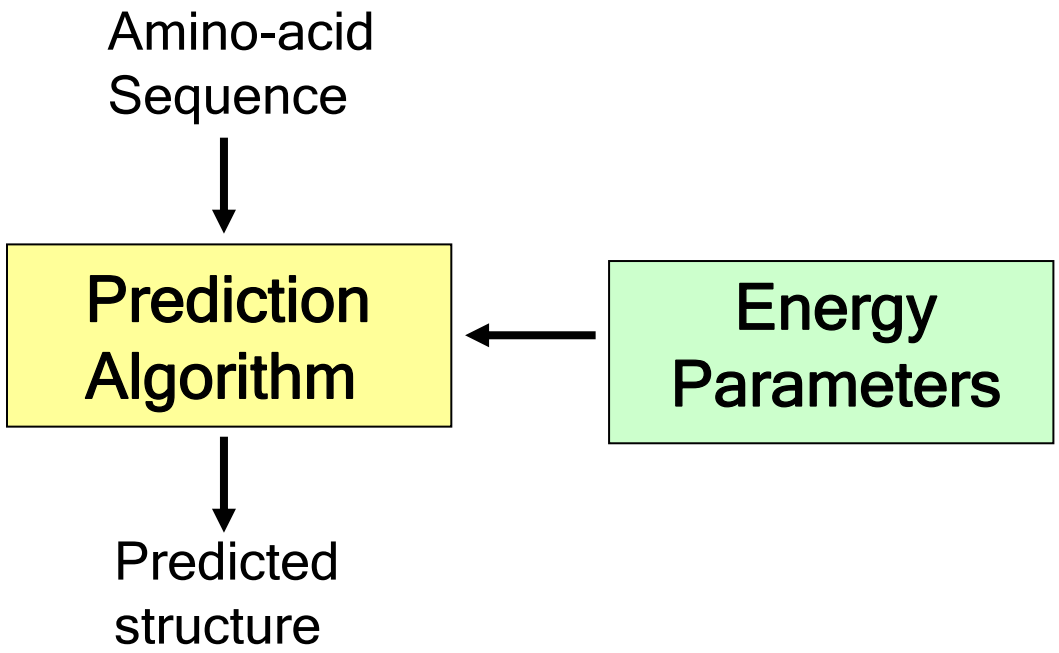
1. Energy-minimization (ab-initio) techniques
   - Elegant, but often lack correct parameters
2. Homology-based techniques
   - Useful, but hard to predict new proteins

Our approach:
Use energy minimization, but
learn parameters from existing proteins

Our Framework (Training)



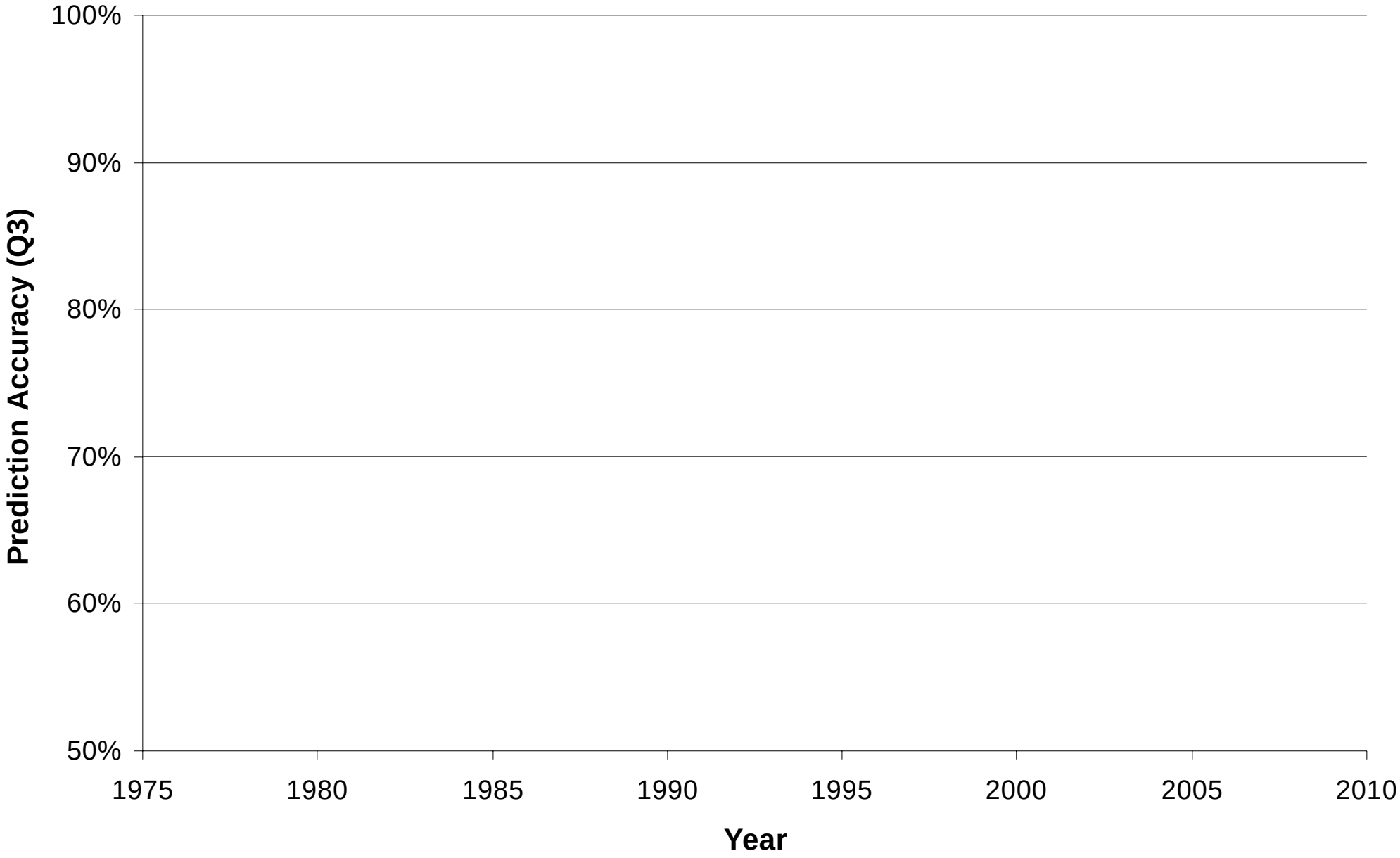
Our Framework (Testing)



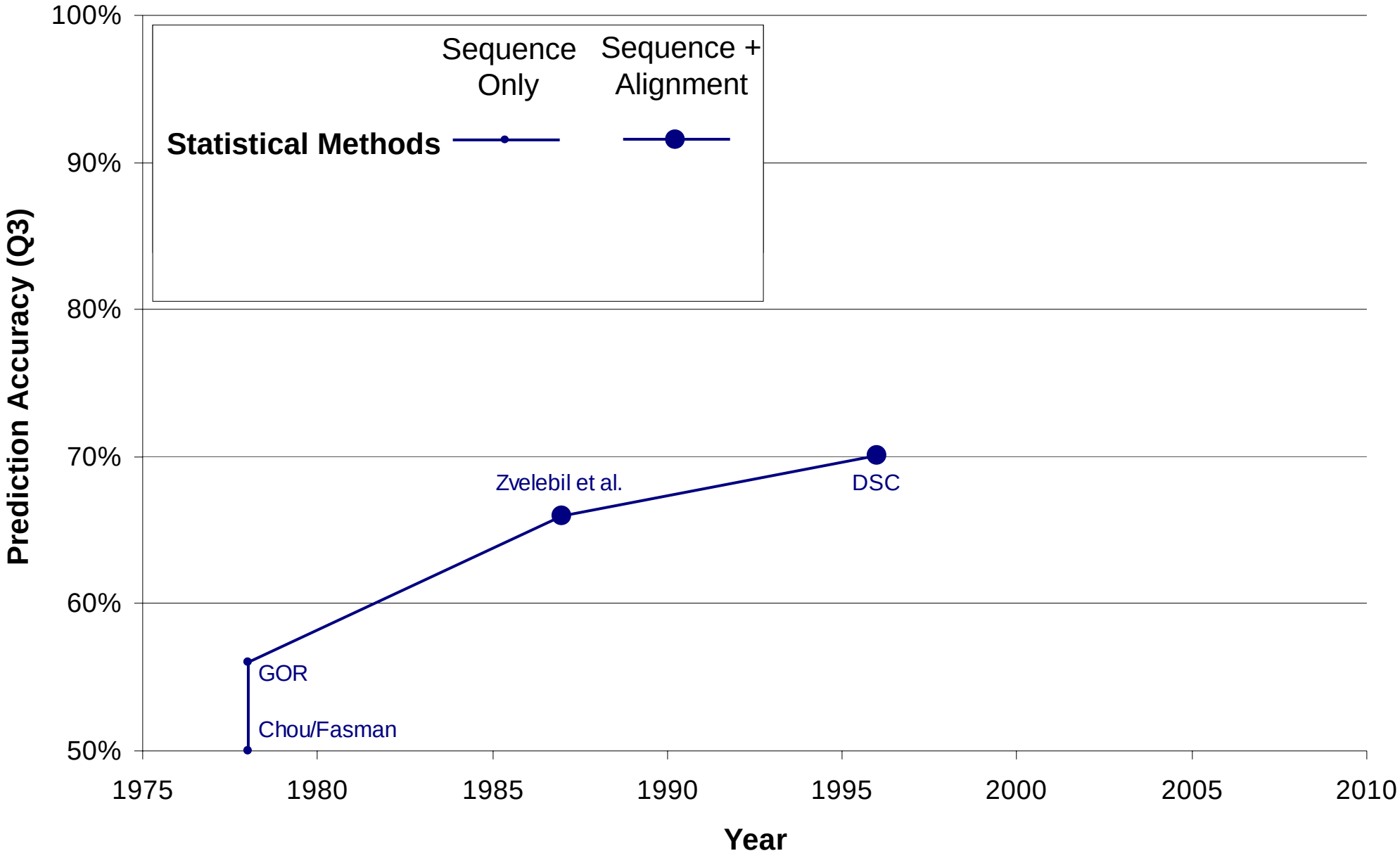
Initial Focus: Secondary Structure

- **Classify each residue as alpha helix, beta strand, coil**
  - In this paper, restrict to all-alpha proteins
- **Applications:**
  - Informing tertiary structure predictors
  - Identification of homologous proteins
  - Identification of active sites (coils)

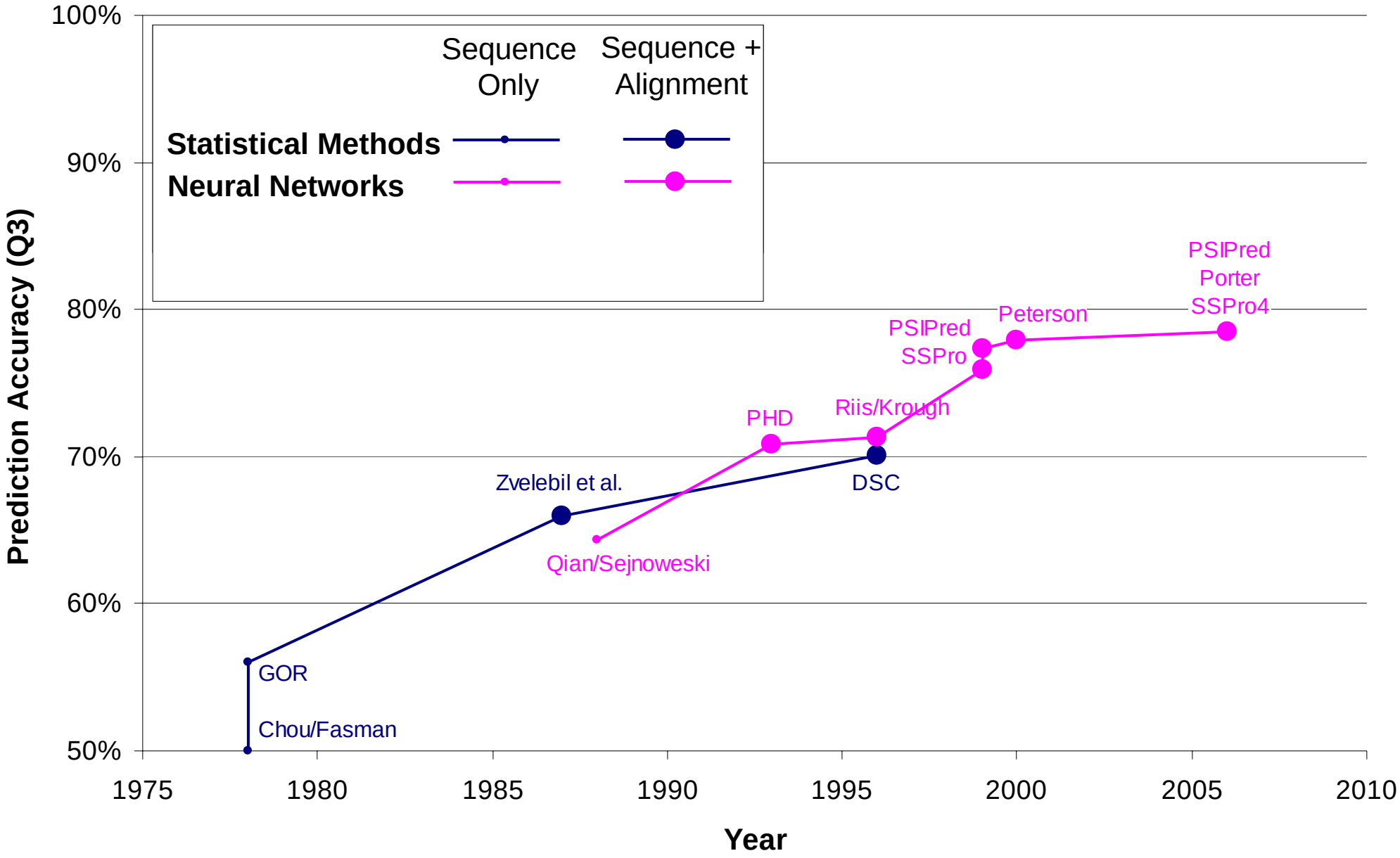
Secondary Structure Predictors



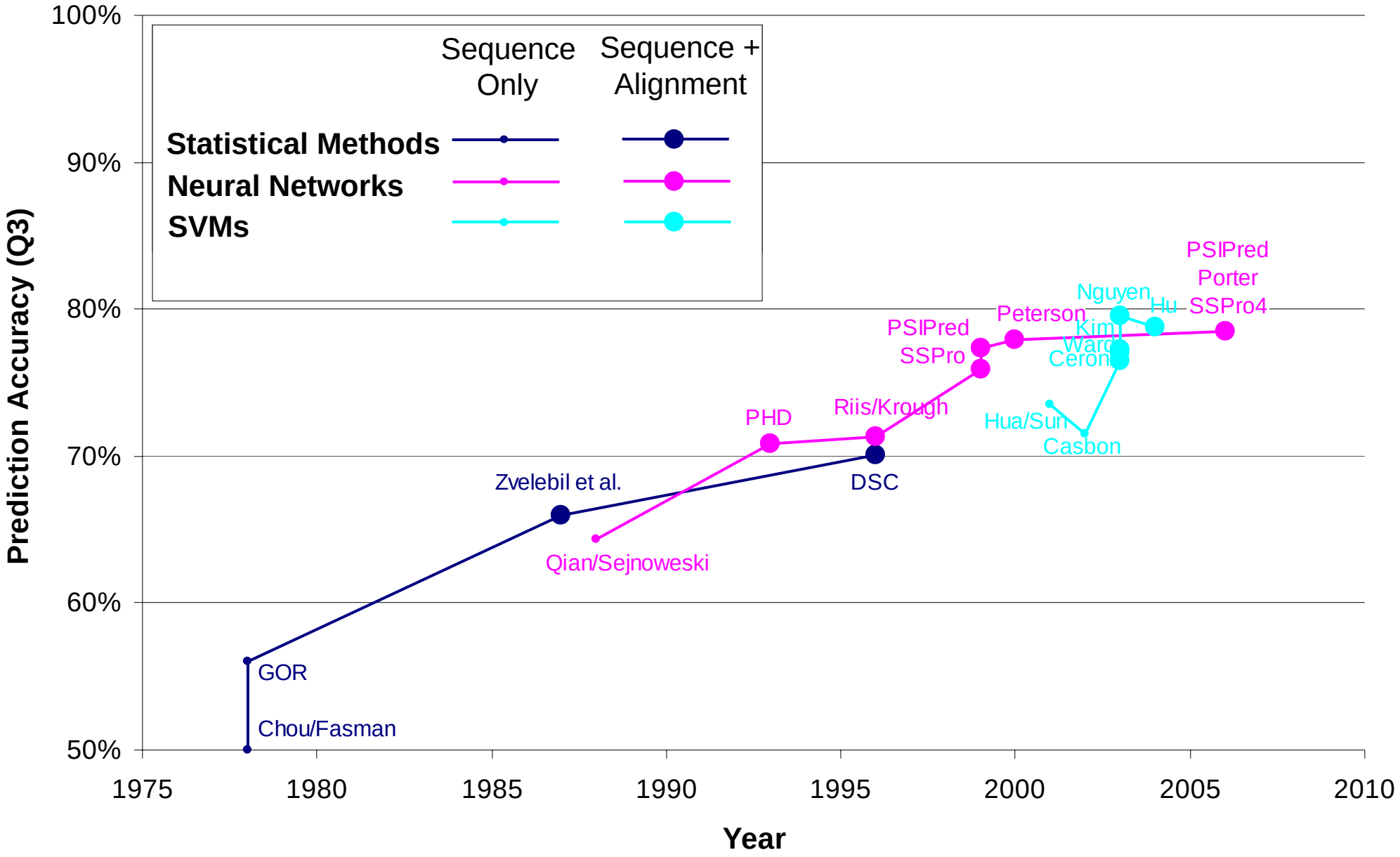
Secondary Structure Predictors



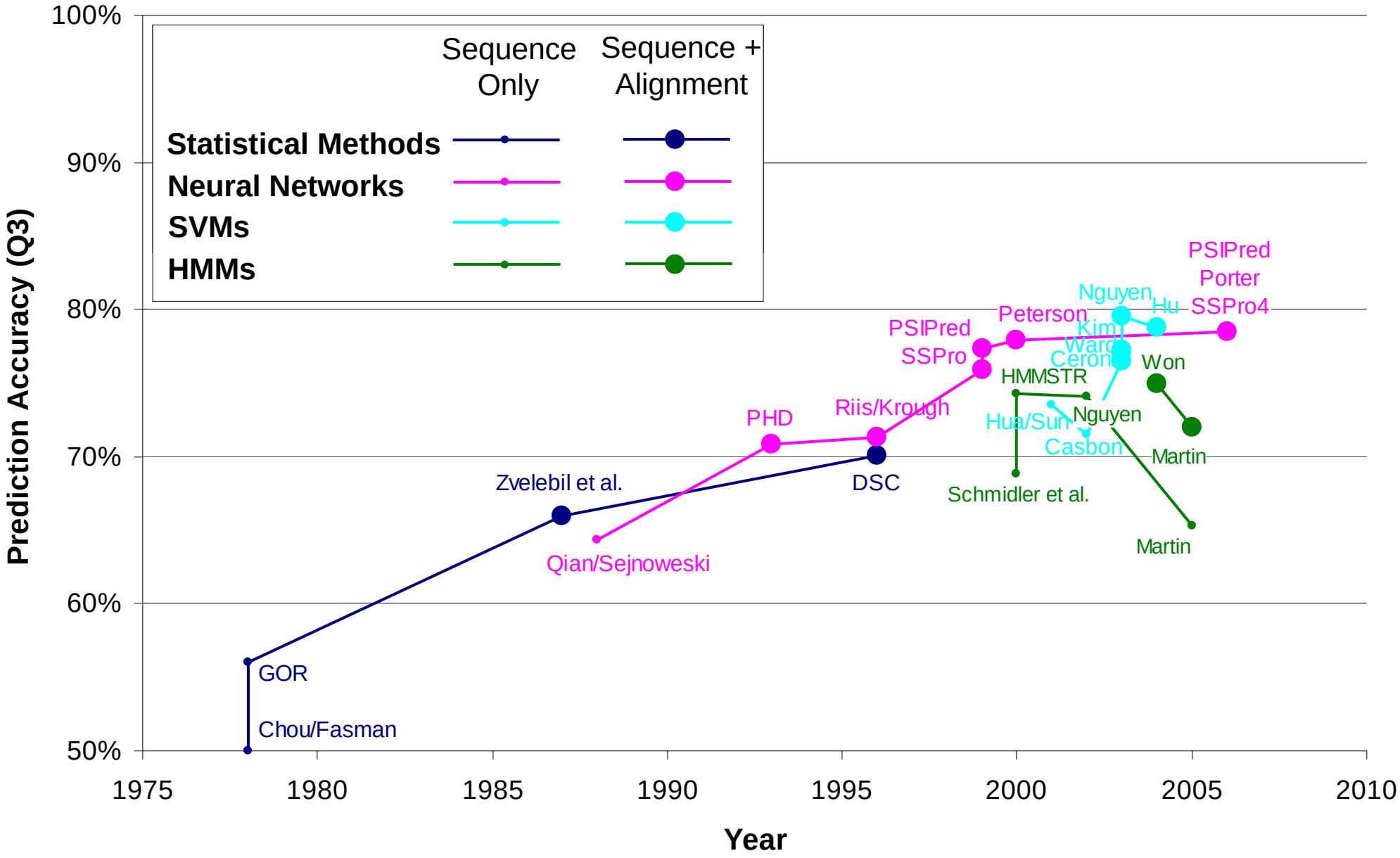
Secondary Structure Predictors



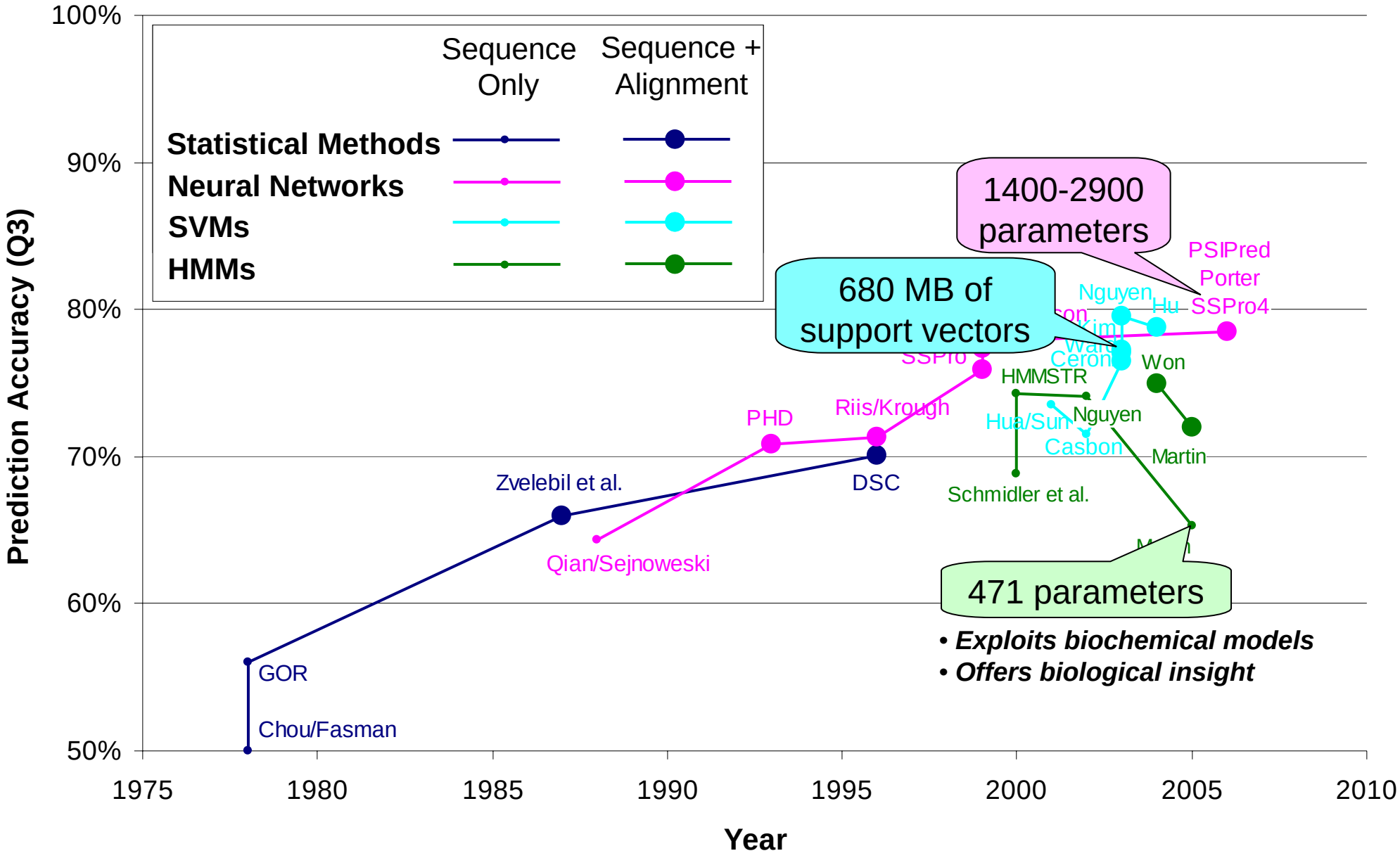
Secondary Structure Predictors



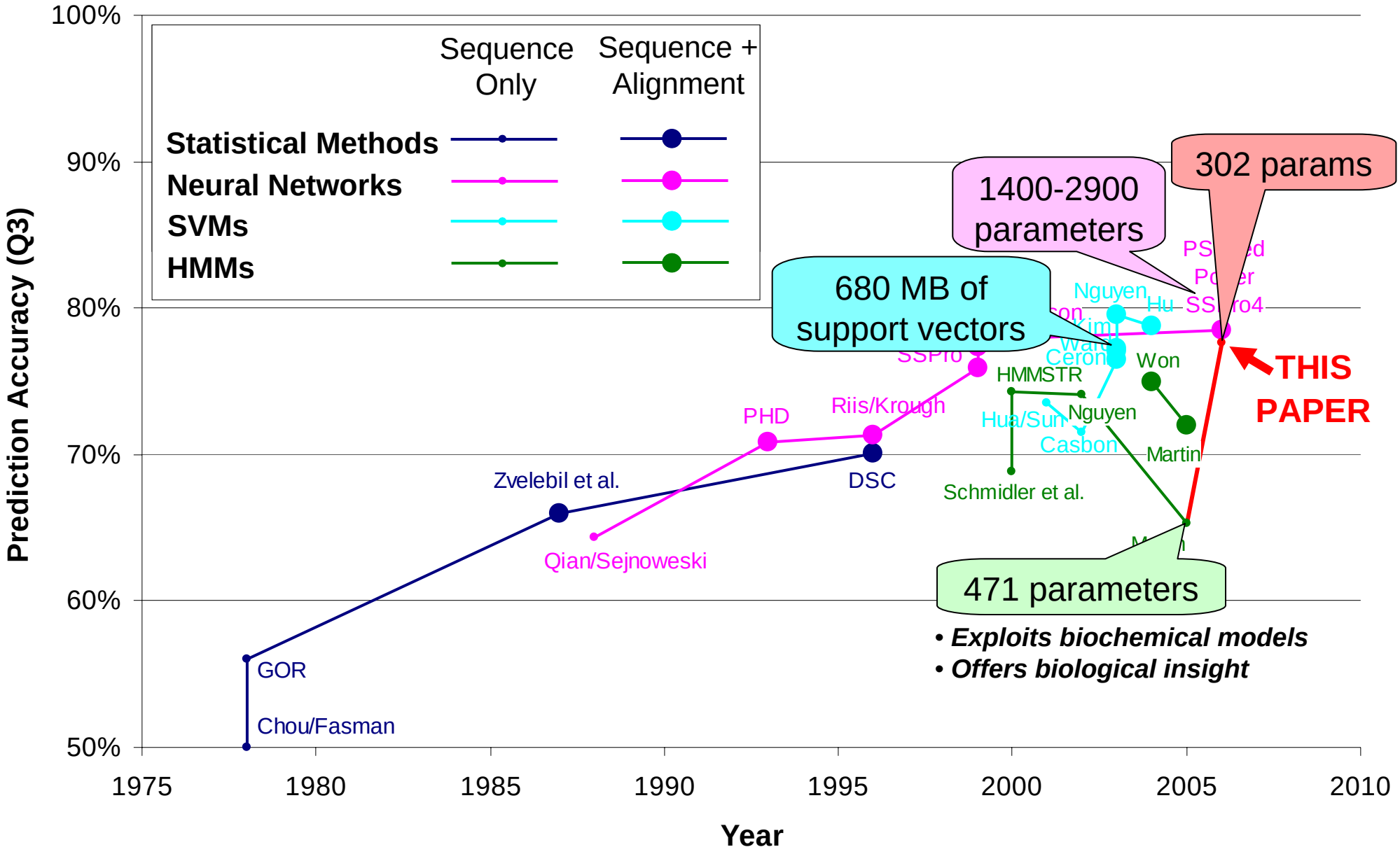
Secondary Structure Predictors



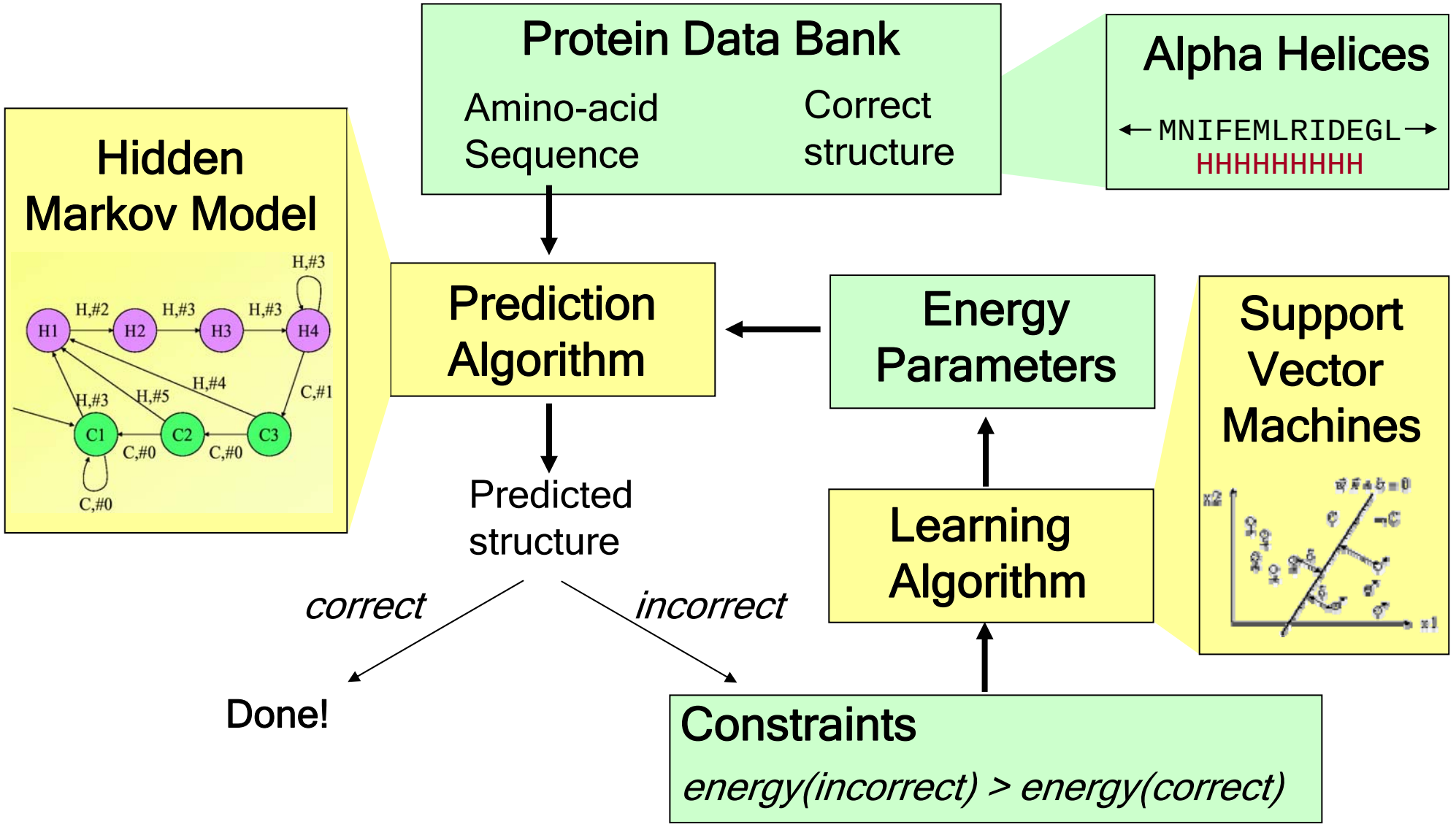
Secondary Structure Predictors



Secondary Structure Predictors



Our Framework Applied to Helix Prediction



Energy Parameters

Description of Energy Parameters	Number of Parameters	Name
Energy of residue R in a helix	20	H_R
Energy of residue R at offset i (-3…3) from N-cap	140	$N_{R,i}$
Energy of residue R at offset i (-3…3) from C-cap	140	$C_{R,i}$
Penalty for coils of length 1 or 2	2	
	302 Total	

Energy Parameters

Description of Energy Parameters	Number of Parameters	Name
Energy of residue R in a helix	20	H_R
Energy of residue R at offset i (-3…3) from N-cap	140	$N_{R,i}$
Energy of residue R at offset i (-3…3) from C-cap	140	$C_{R,i}$
Penalty for coils of length 1 or 2	2	
	302 Total	

- **Example:**

  Sequence: **MNIFELRIDEGL**

  Structure: **HHHHHH**

  Energy =

Energy Parameters

Description of Energy Parameters	Number of Parameters	Name
Energy of residue R in a helix	20	H_R
Energy of residue R at offset i (-3…3) from N-cap	140	$N_{R,i}$
Energy of residue R at offset i (-3…3) from C-cap	140	$C_{R,i}$
Penalty for coils of length 1 or 2	2	
	302 Total	

- **Example:**

  Sequence: **MNIFELRIDEGL**

  Structure: **HHHHHH**

  Energy = $H_F + H_E + H_L + H_R + H_I + H_D$ (Helix)

Energy Parameters

Description of Energy Parameters	Number of Parameters	Name
Energy of residue R in a helix	20	H_R
Energy of residue R at offset i (-3…3) from N-cap	140	$N_{R,i}$
Energy of residue R at offset i (-3…3) from C-cap	140	$C_{R,i}$
Penalty for coils of length 1 or 2	2	
	302 Total	

- **Example:**

  Sequence: **MNIFELRIDEGL**

  Structure: **HHHHHH**

$$\text{Energy} = H_F + H_E + H_L + H_R + H_I + H_D \quad \text{(Helix)}$$

$$+ N_{M,-3} + N_{N,-2} + N_{I,-1} + N_{F,0} + N_{E,1} + N_{L,2} + N_{R,3} \quad \text{(N-cap)}$$

Energy Parameters

Description of Energy Parameters	Number of Parameters	Name
Energy of residue R in a helix	20	H_R
Energy of residue R at offset i (-3…3) from N-cap	140	$N_{R,i}$
Energy of residue R at offset i (-3…3) from C-cap	140	$C_{R,i}$
Penalty for coils of length 1 or 2	2	
	302 Total	

- **Example:**

  Sequence: **MNIFELRIDEGL**

  Structure: **HHHHHH**

$$\text{Energy} = H_F + H_E + H_L + H_R + H_I + H_D \quad \text{(Helix)}$$
$$+ N_{M,-3} + N_{N,-2} + N_{I,-1} + N_{F,0} + N_{E,1} + N_{L,2} + N_{R,3} \quad \text{(N-cap)}$$
$$+ C_{L,-3} + C_{R,-2} + C_{I,-1} + C_{D,0} + C_{E,1} + C_{G,2} + C_{L,3} \quad \text{(C-cap)}$$

Energy Parameters

Description of Energy Parameters	Number of Parameters	Name
Energy of residue R in a helix	20	H_R
Energy of residue R at offset i (-3…3) from N-cap	140	$N_{R,i}$
Energy of residue R at offset i (-3…3) from C-cap	140	$C_{R,i}$
Penalty for coils of length 1 or 2	2	
	302 Total	

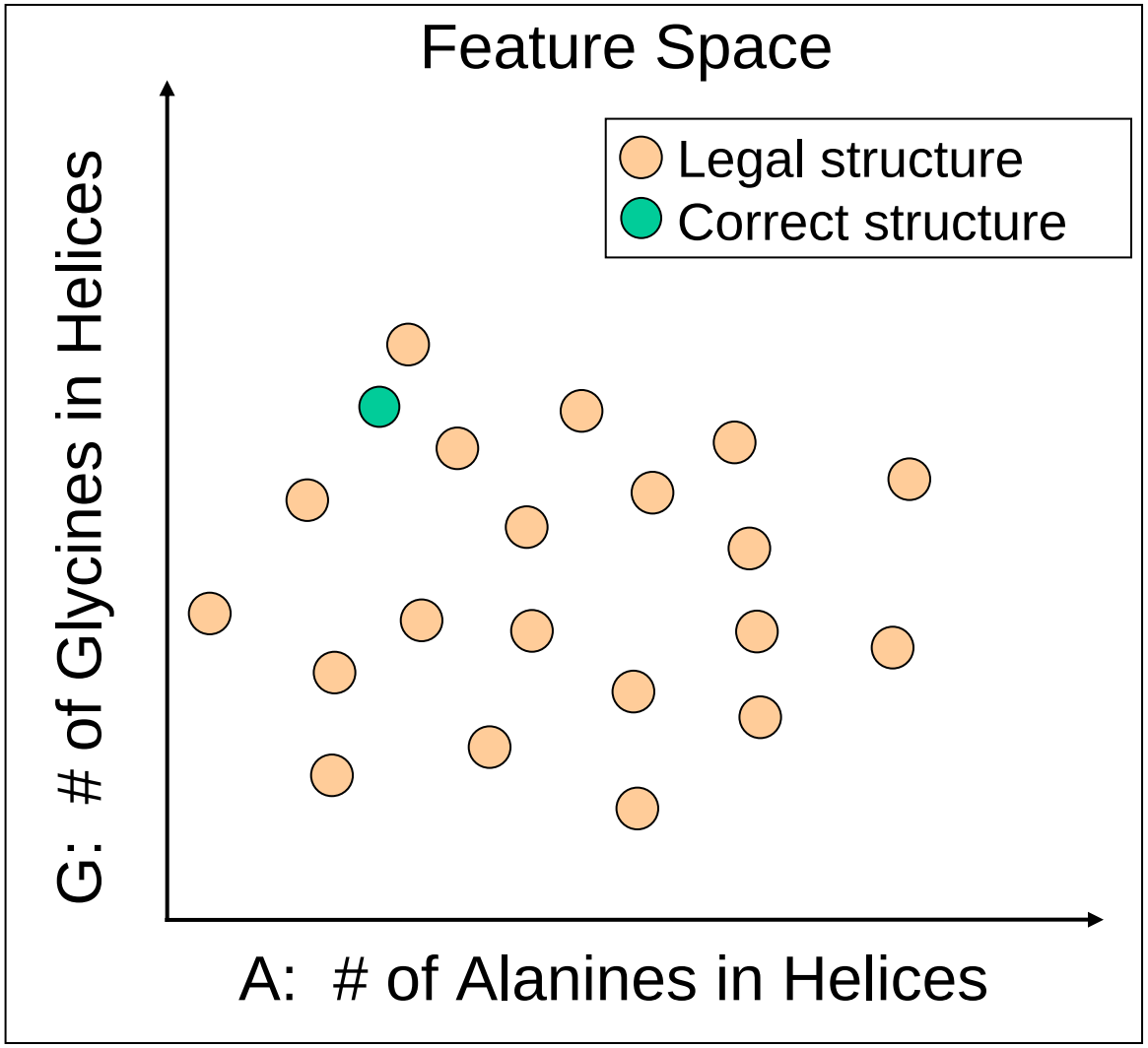
- **Example:**

  Sequence: **MNIFELRIDEGL**

  Structure: **HHHHHH**

$$\text{Energy} = H_F + H_E + H_L + H_R + H_I + H_D \quad \text{(Helix)}$$
$$+ N_{M,-3} + N_{N,-2} + N_{I,-1} + N_{F,0} + N_{E,1} + N_{L,2} + N_{R,3} \quad \text{(N-cap)}$$
$$+ C_{L,-3} + C_{R,-2} + C_{I,-1} + C_{D,0} + C_{E,1} + C_{G,2} + C_{L,3} \quad \text{(C-cap)}$$

Learning the Parameters



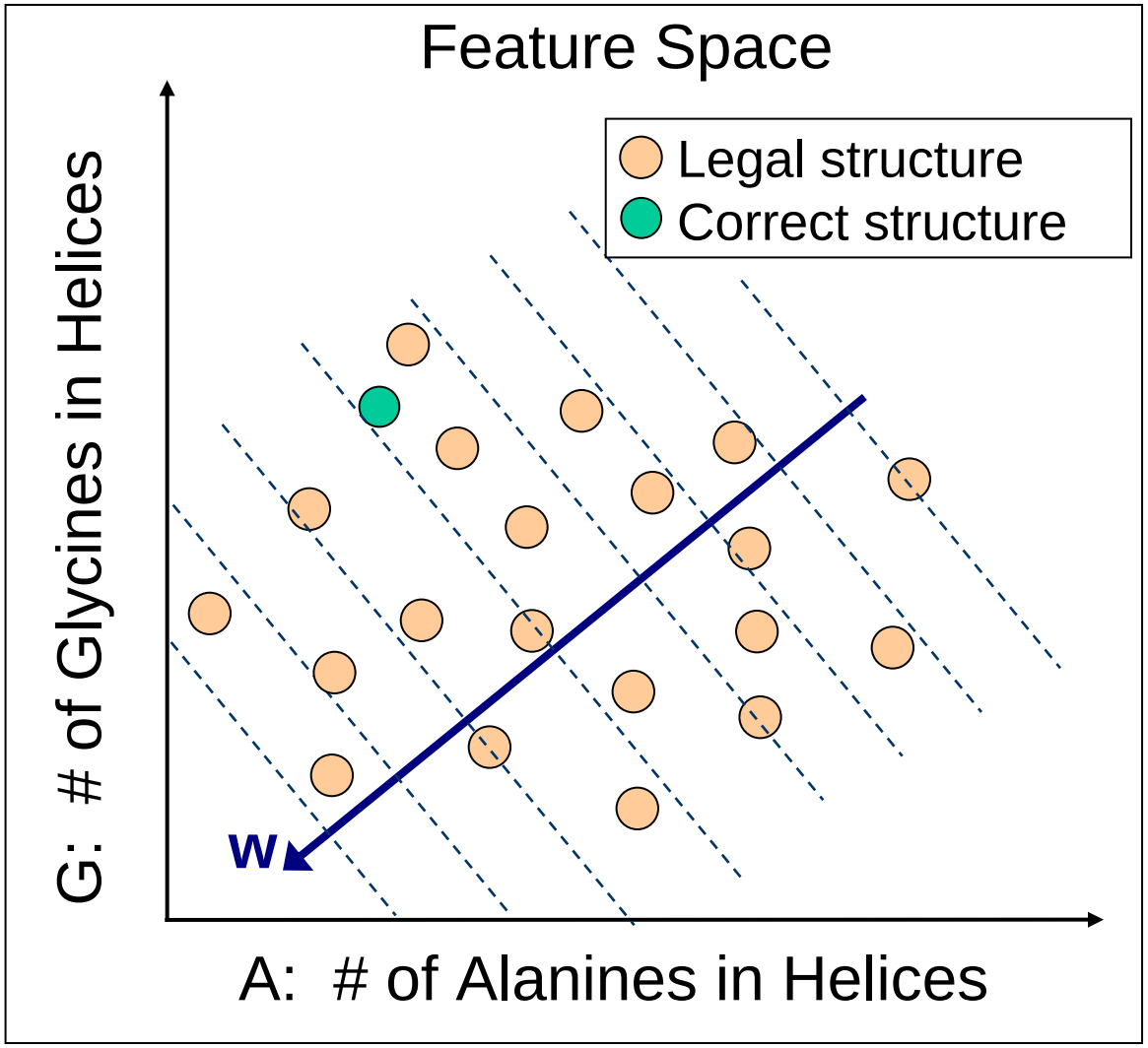
Energy (○) = H_A*A + H_G*G

= **w** · [A G]

where **w** represents the energy parameters [H_A H_G]

Highest energy in direction of energy parameters **w**

Learning the Parameters



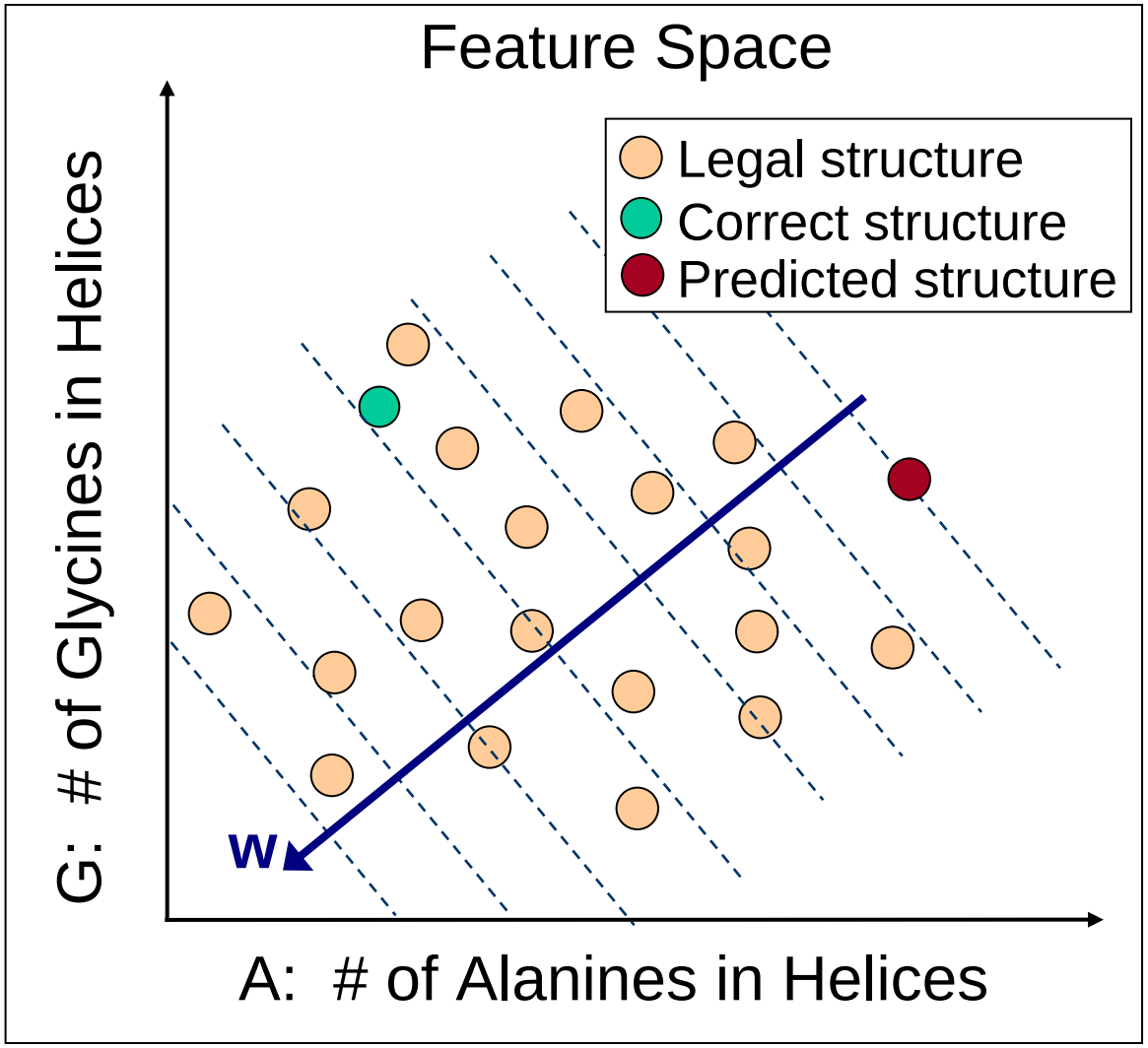
Energy (○) = H_A*A + H_G*G

= **w** · [A G]

where **w** represents the energy parameters [H_A H_G]

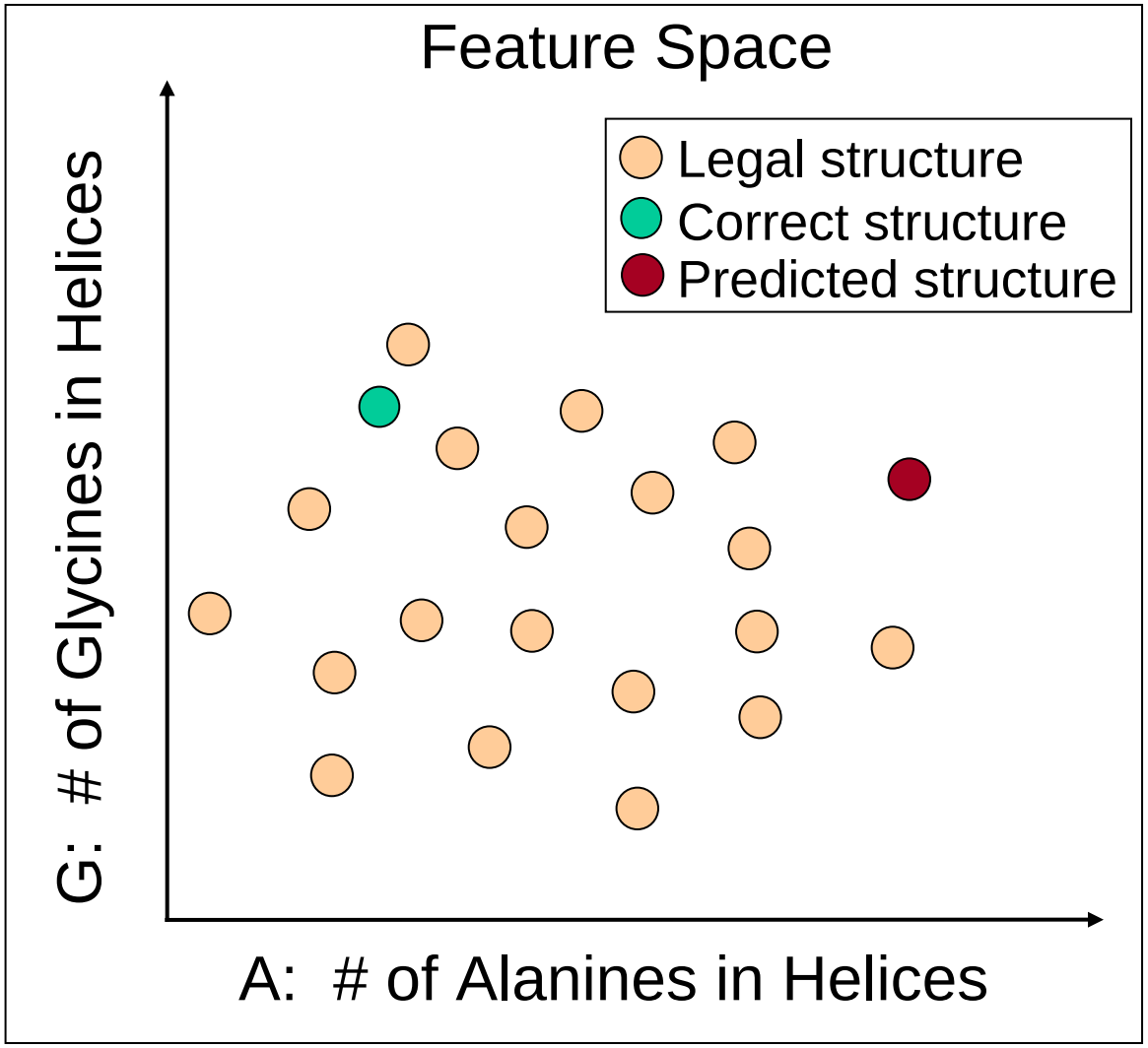
Highest energy in direction of energy parameters **w**

Learning the Parameters



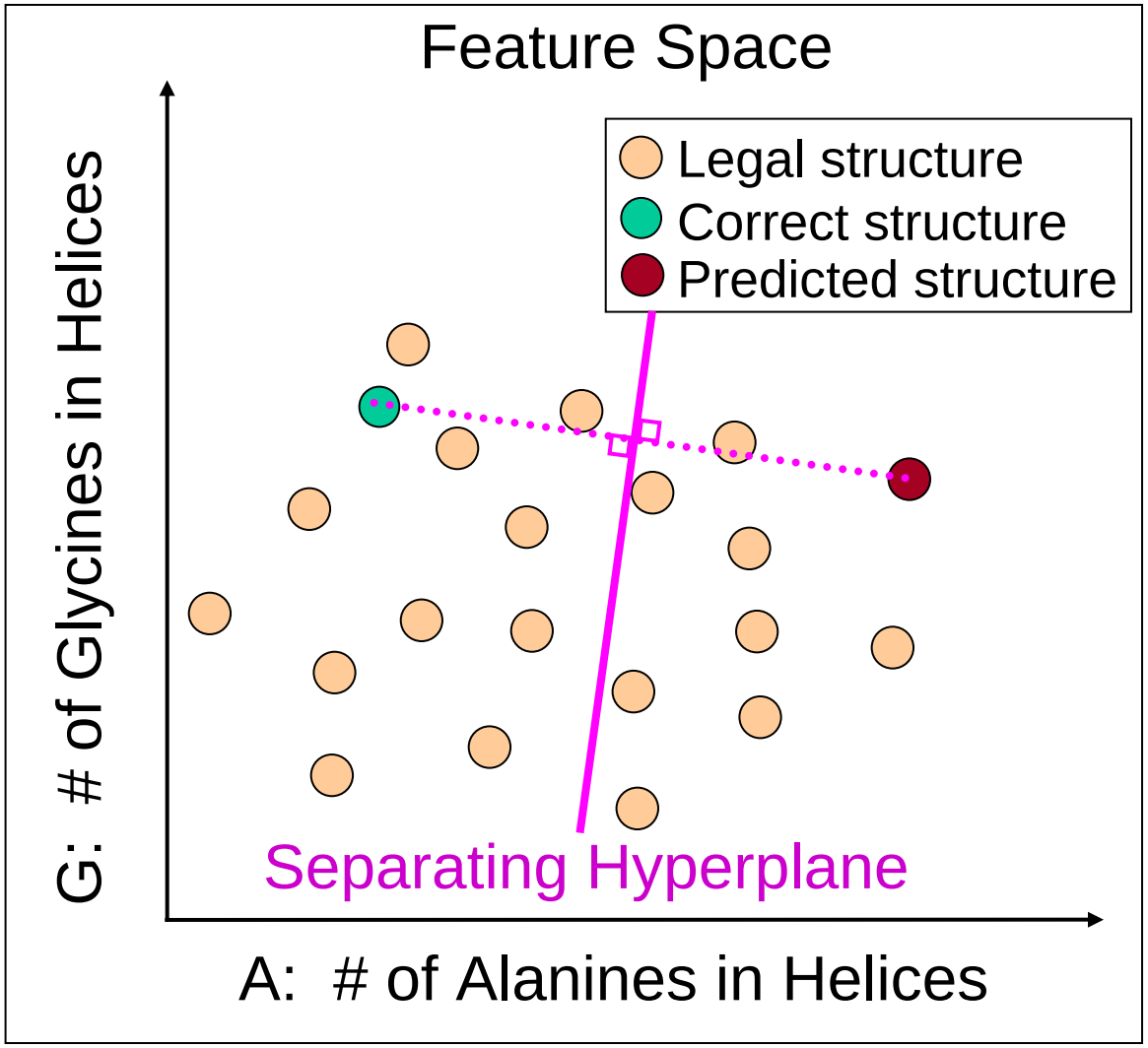
1. Predict stucture

Learning the Parameters



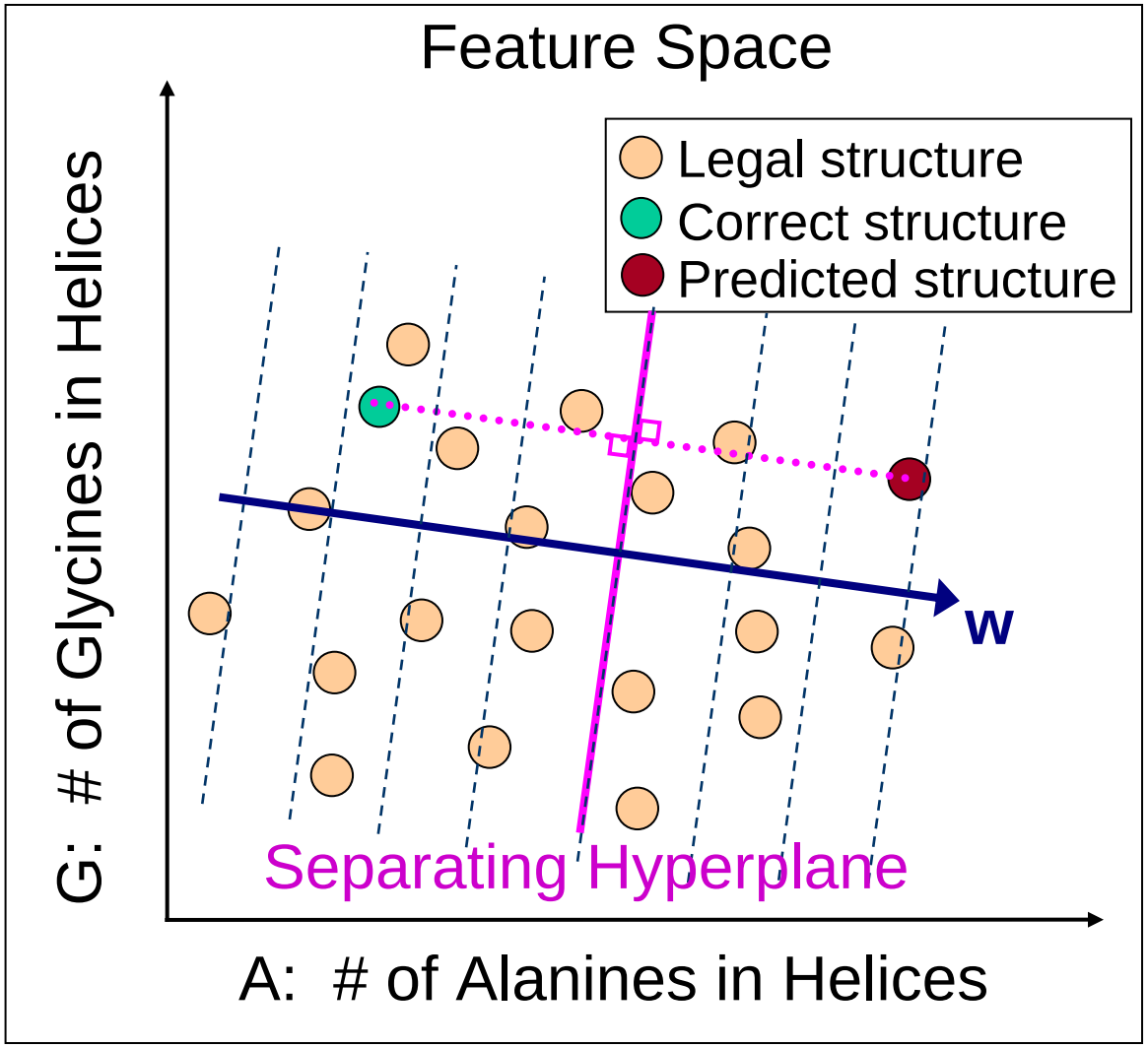
1. Predict stucture

Learning the Parameters



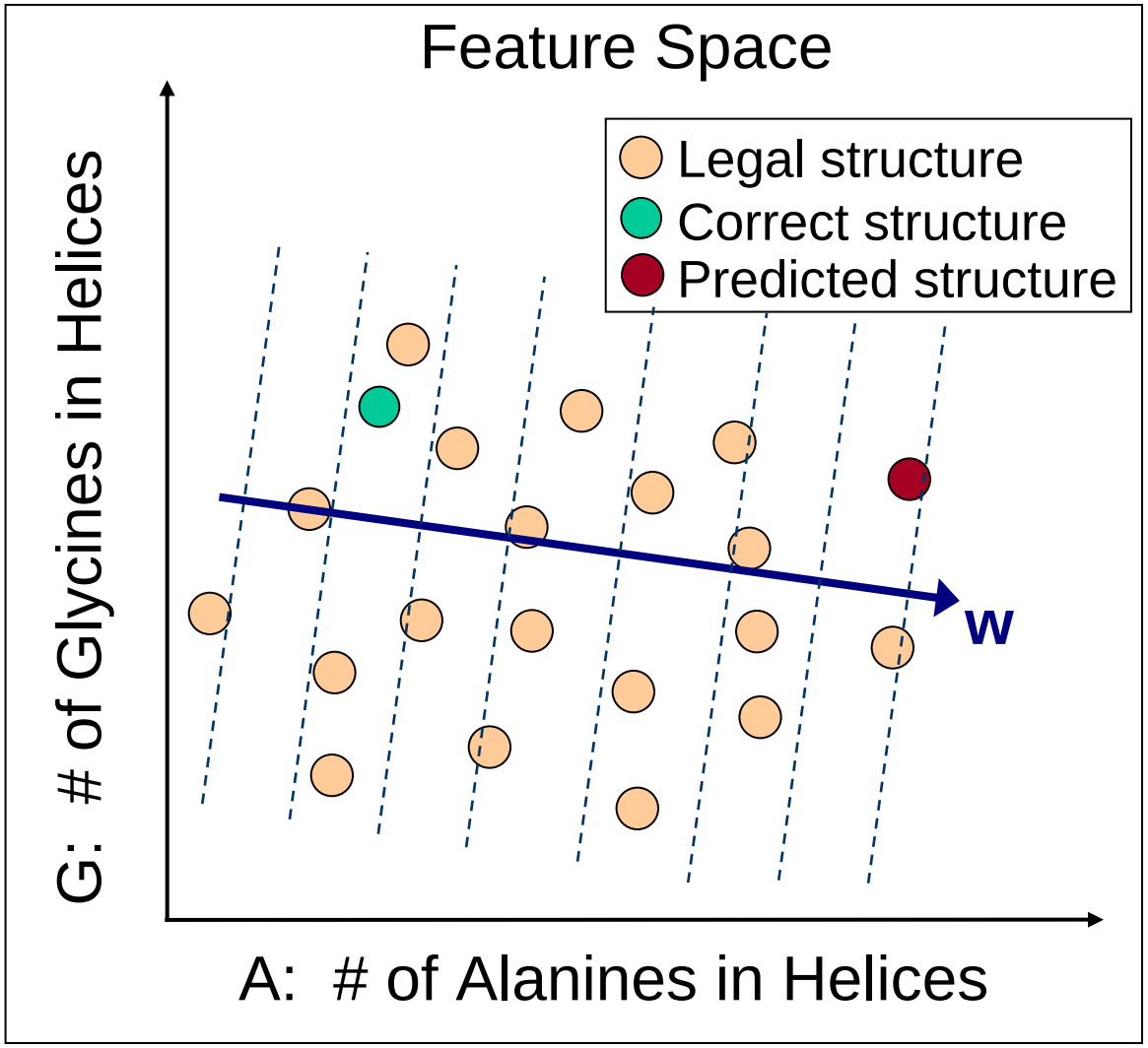
1. Predict stucture
2. Refine parameters

Learning the Parameters



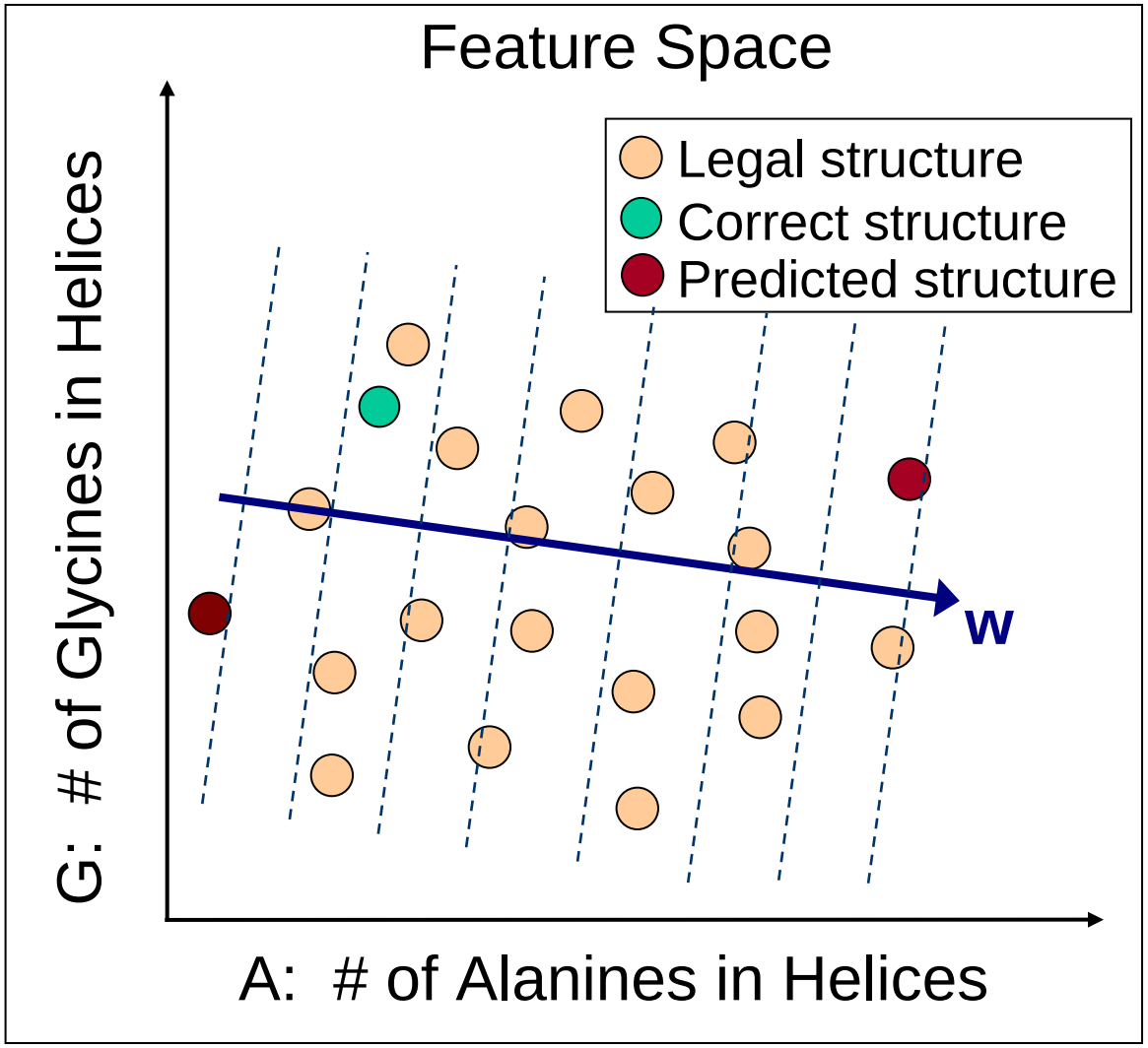
1. Predict stucture
2. Refine parameters

Learning the Parameters



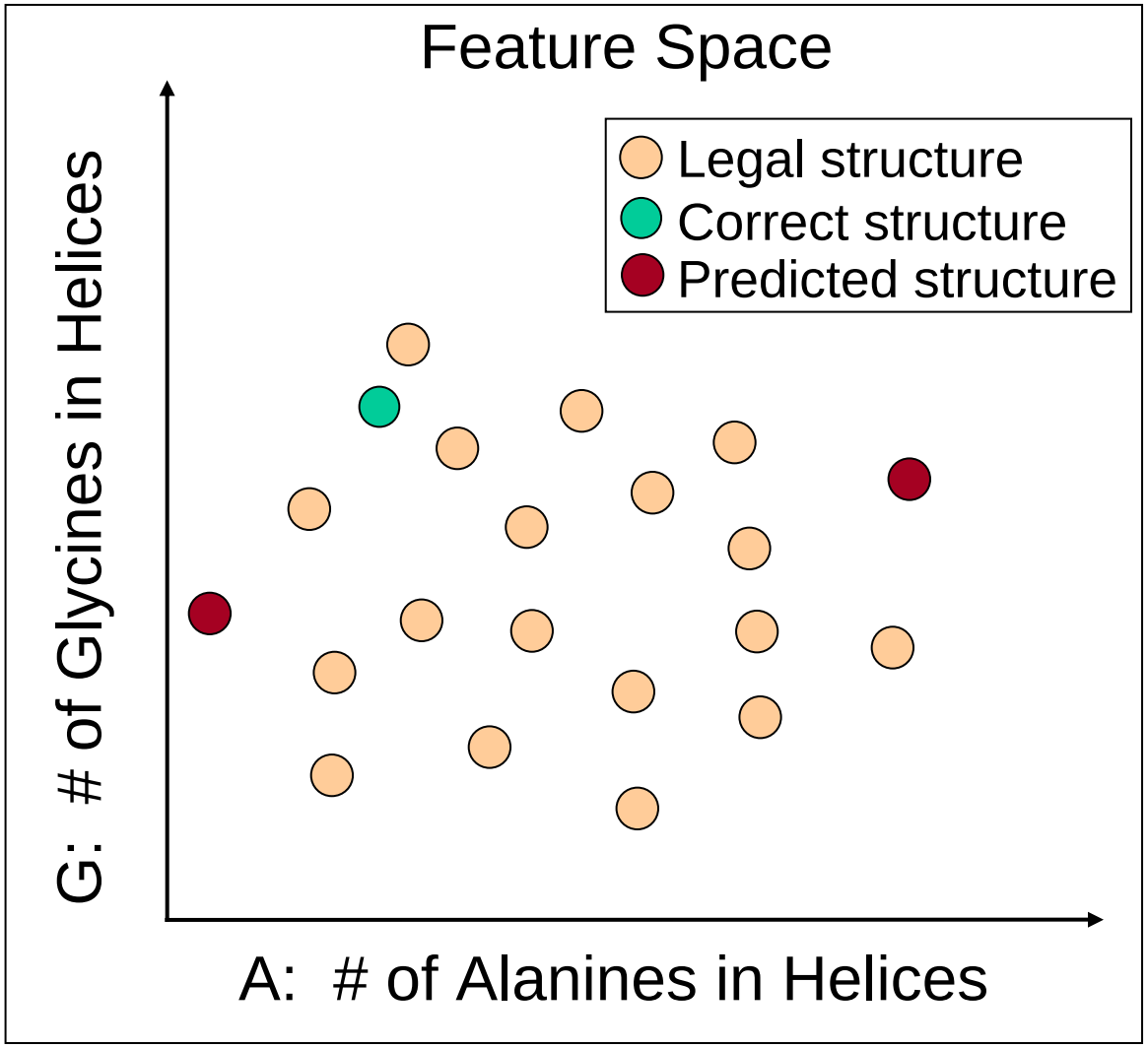
1. Predict stucture
2. Refine parameters

Learning the Parameters



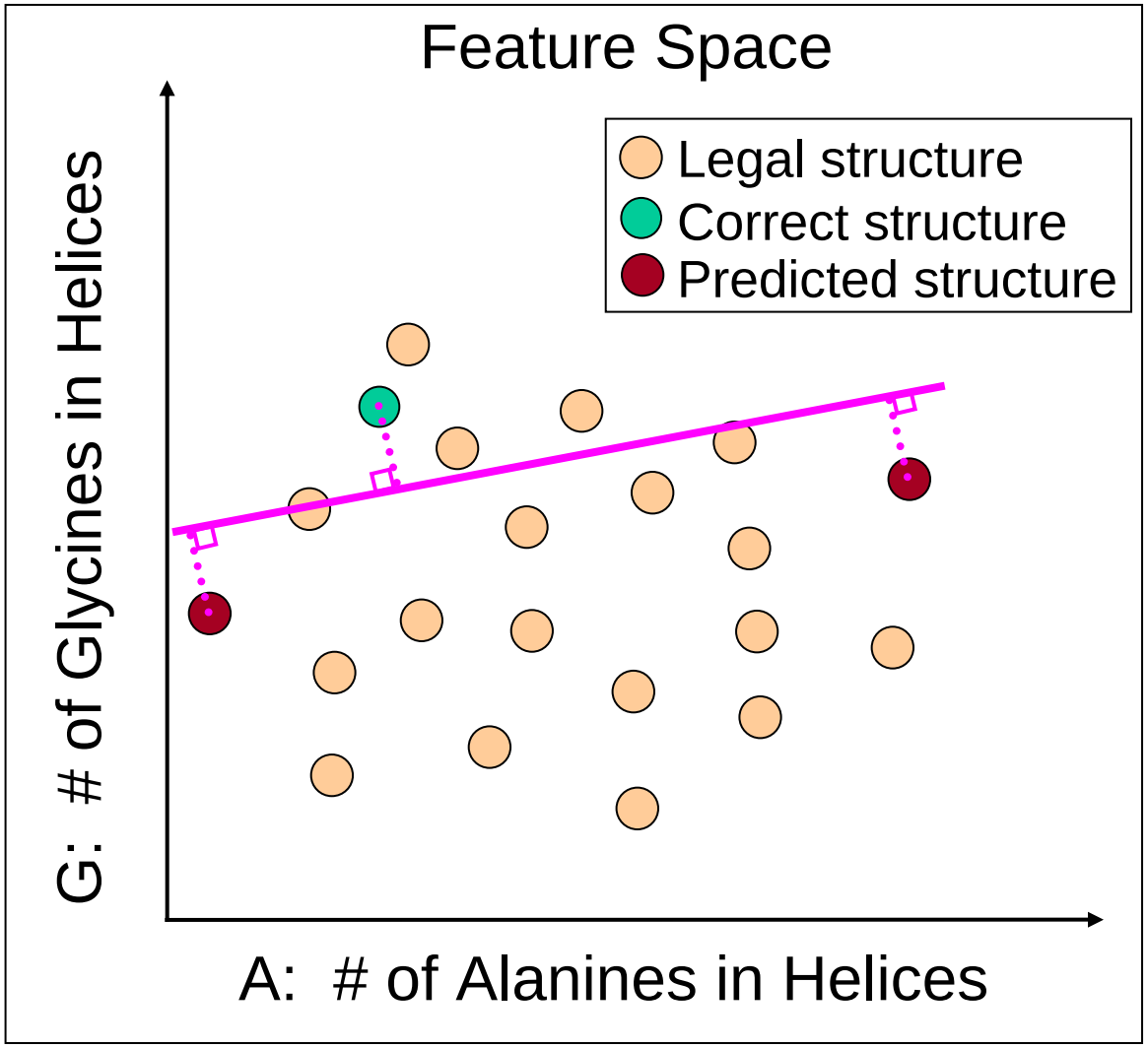
1. Predict stucture
2. Refine parameters
3. Predict structure

Learning the Parameters



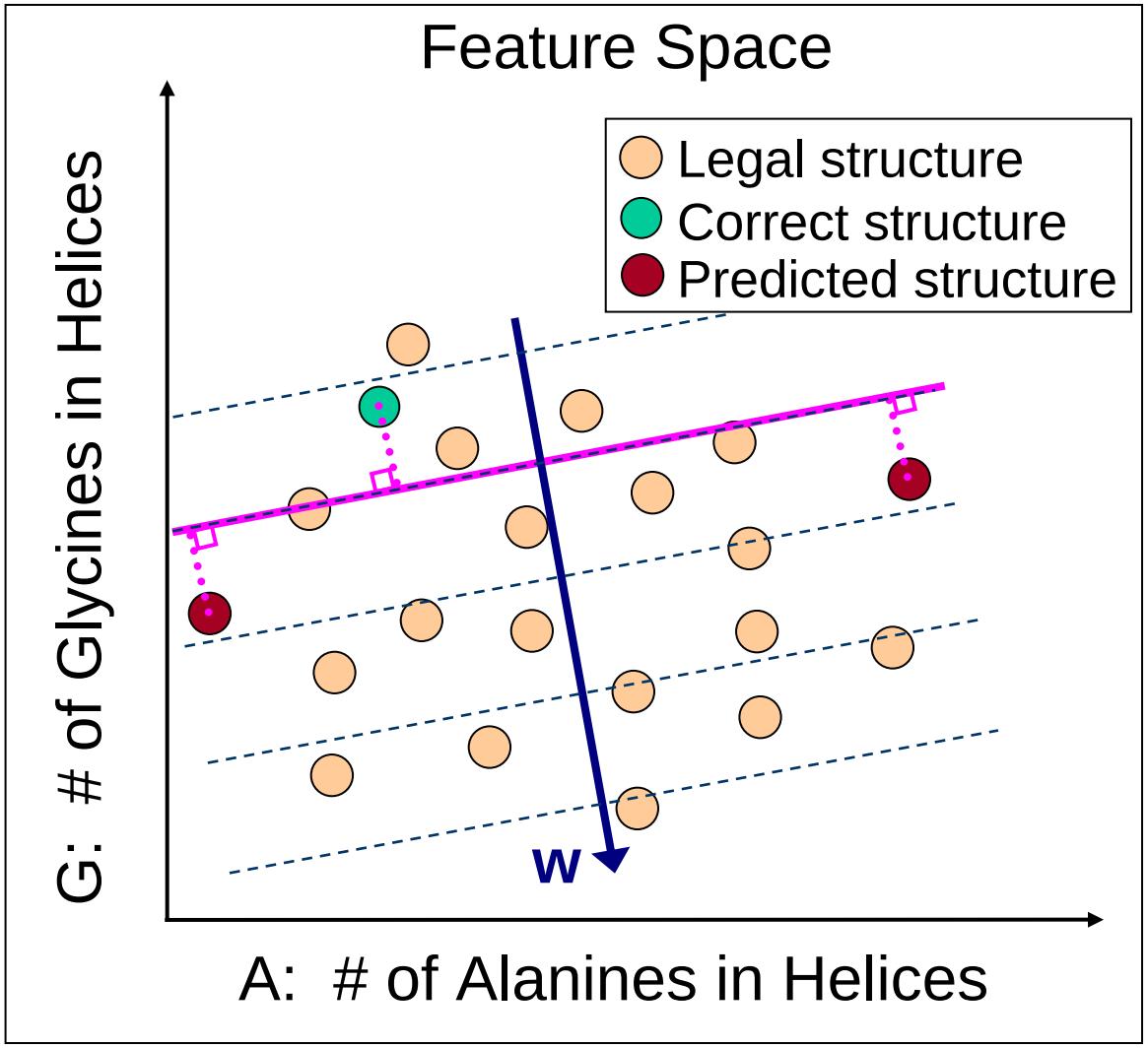
1. Predict stucture
2. Refine parameters
3. Predict structure

Learning the Parameters



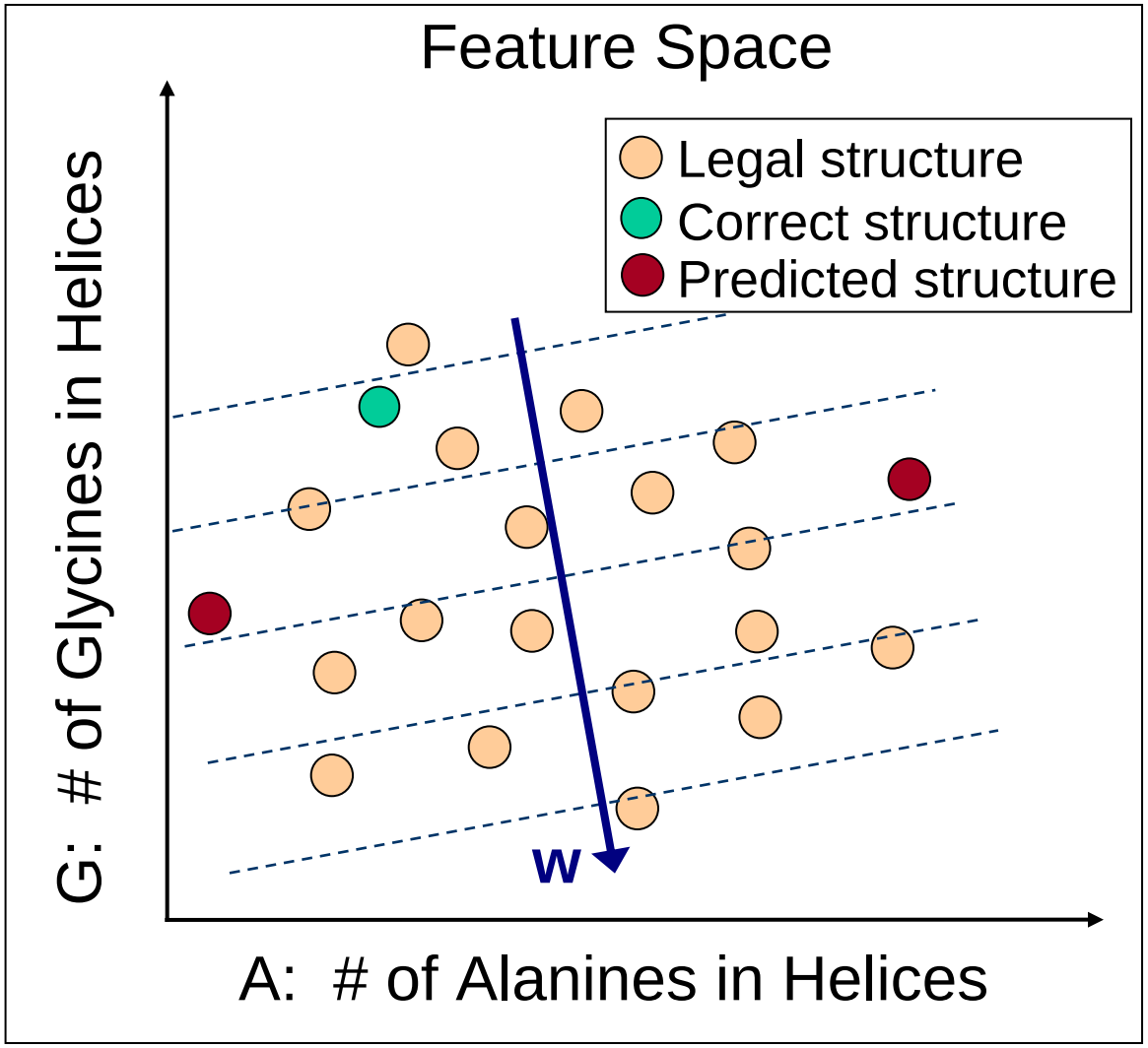
1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters

Learning the Parameters



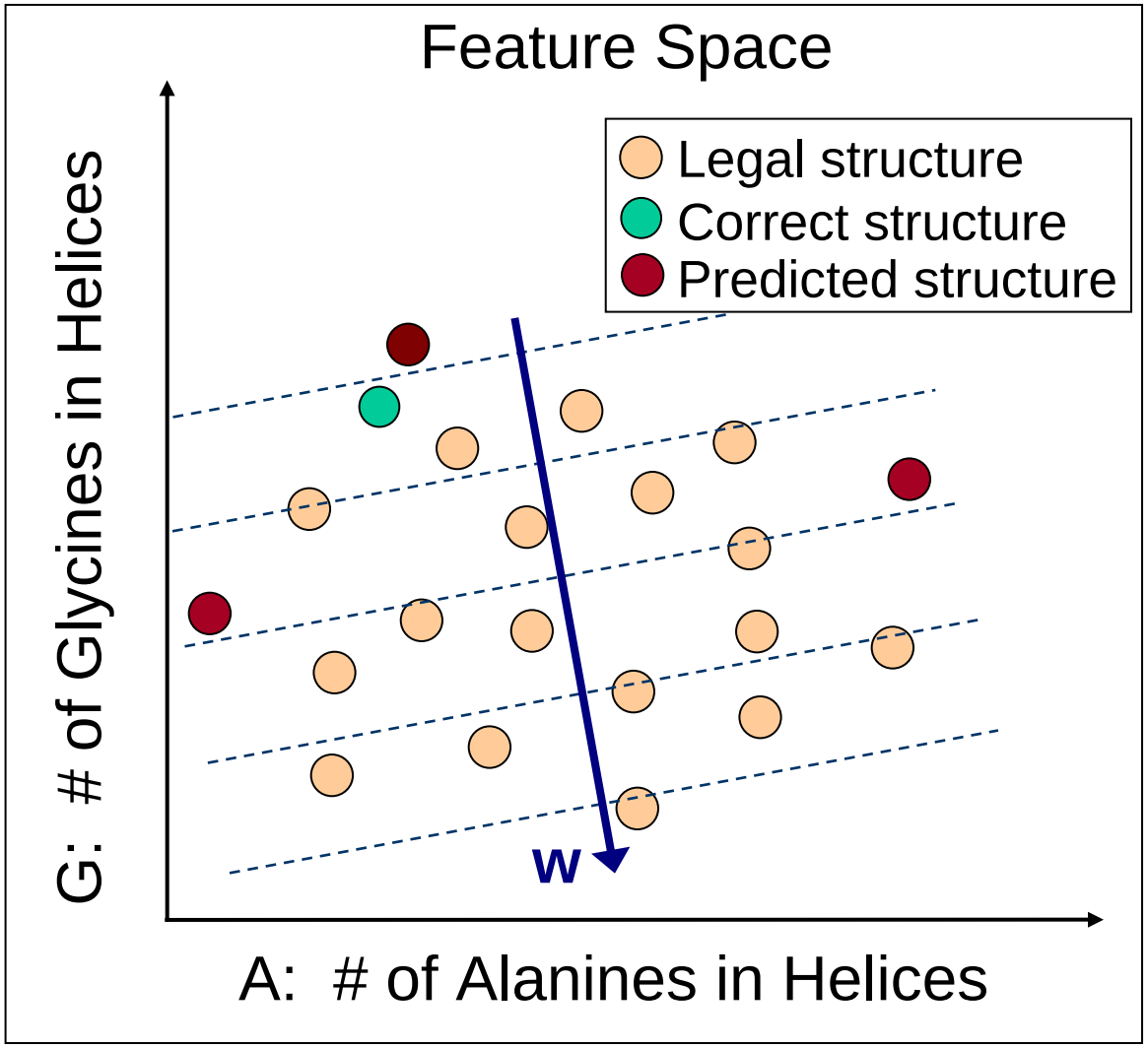
1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters

Learning the Parameters



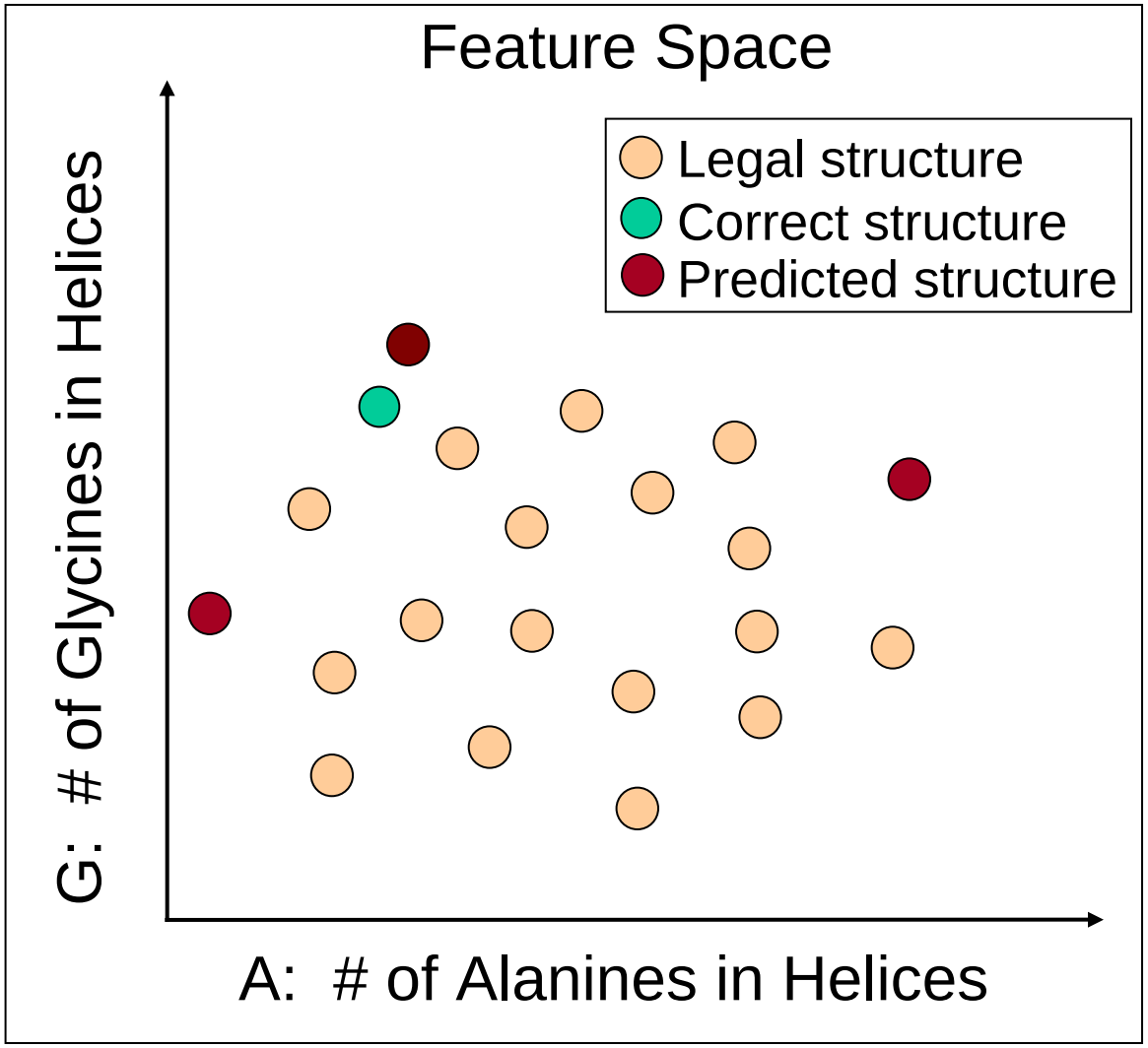
1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters

Learning the Parameters



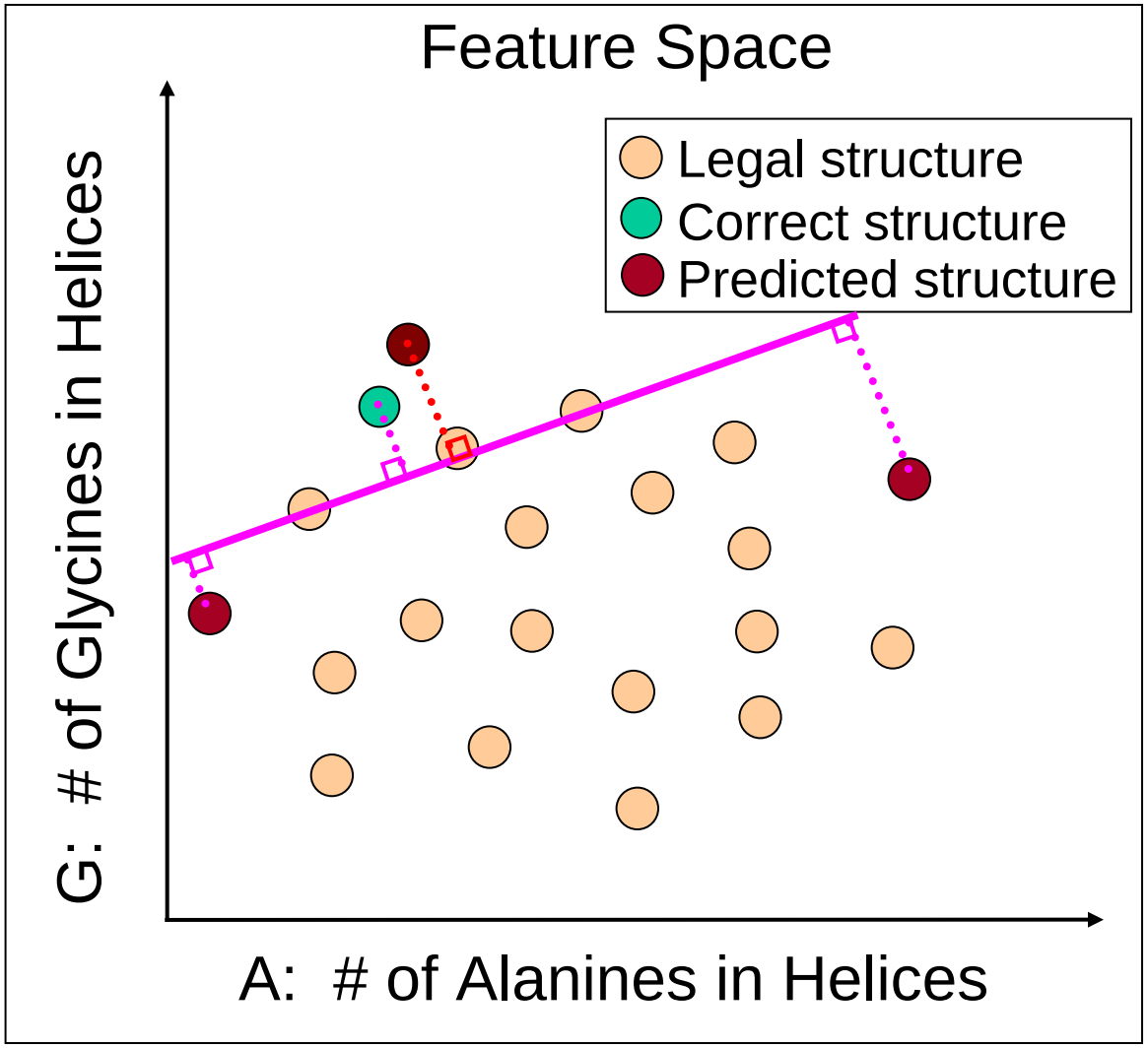
1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters
5. Predict structure

Learning the Parameters



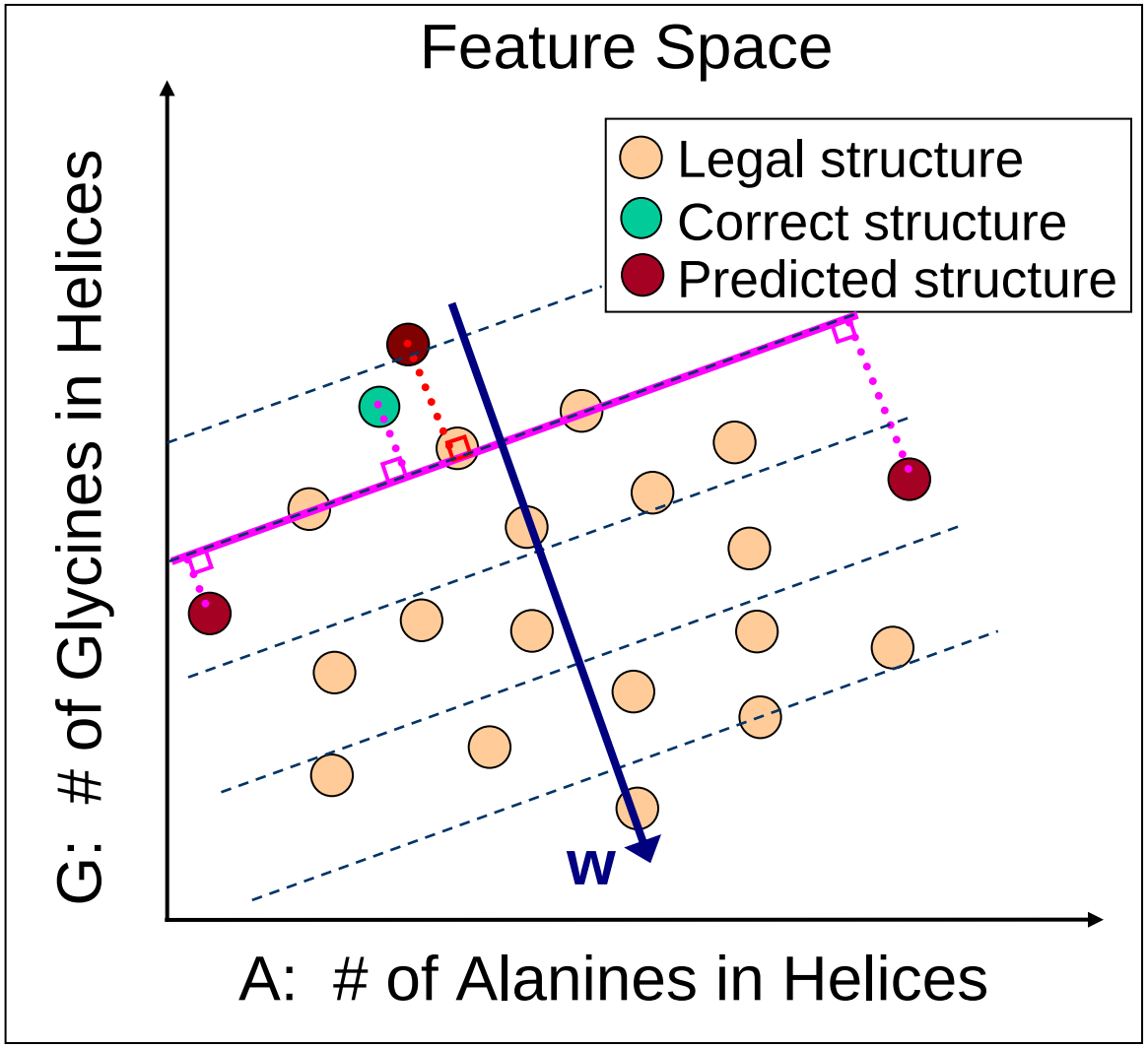
1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters
5. Predict structure

Learning the Parameters



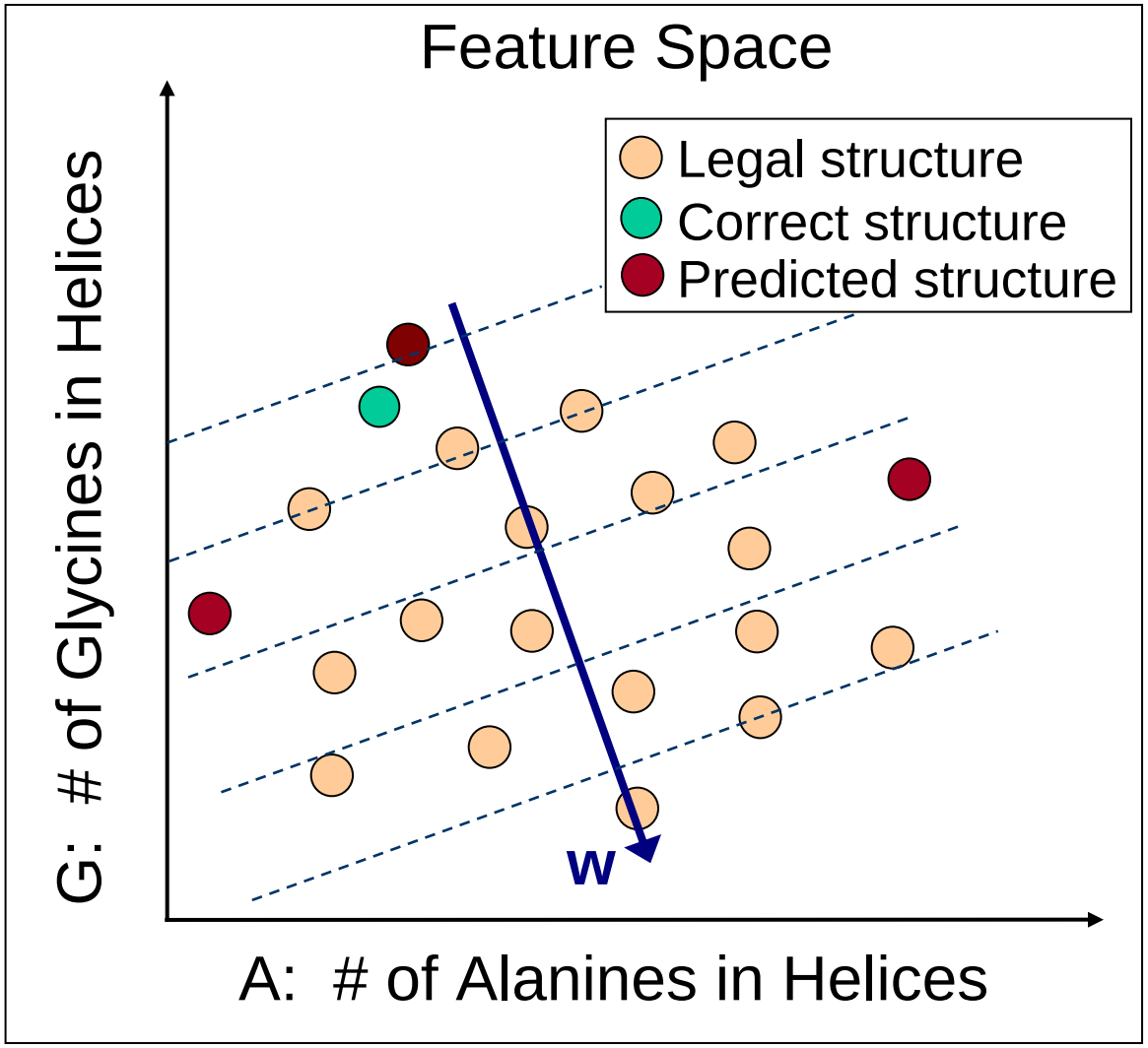
1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters
5. Predict structure
6. Refine parameters

Learning the Parameters



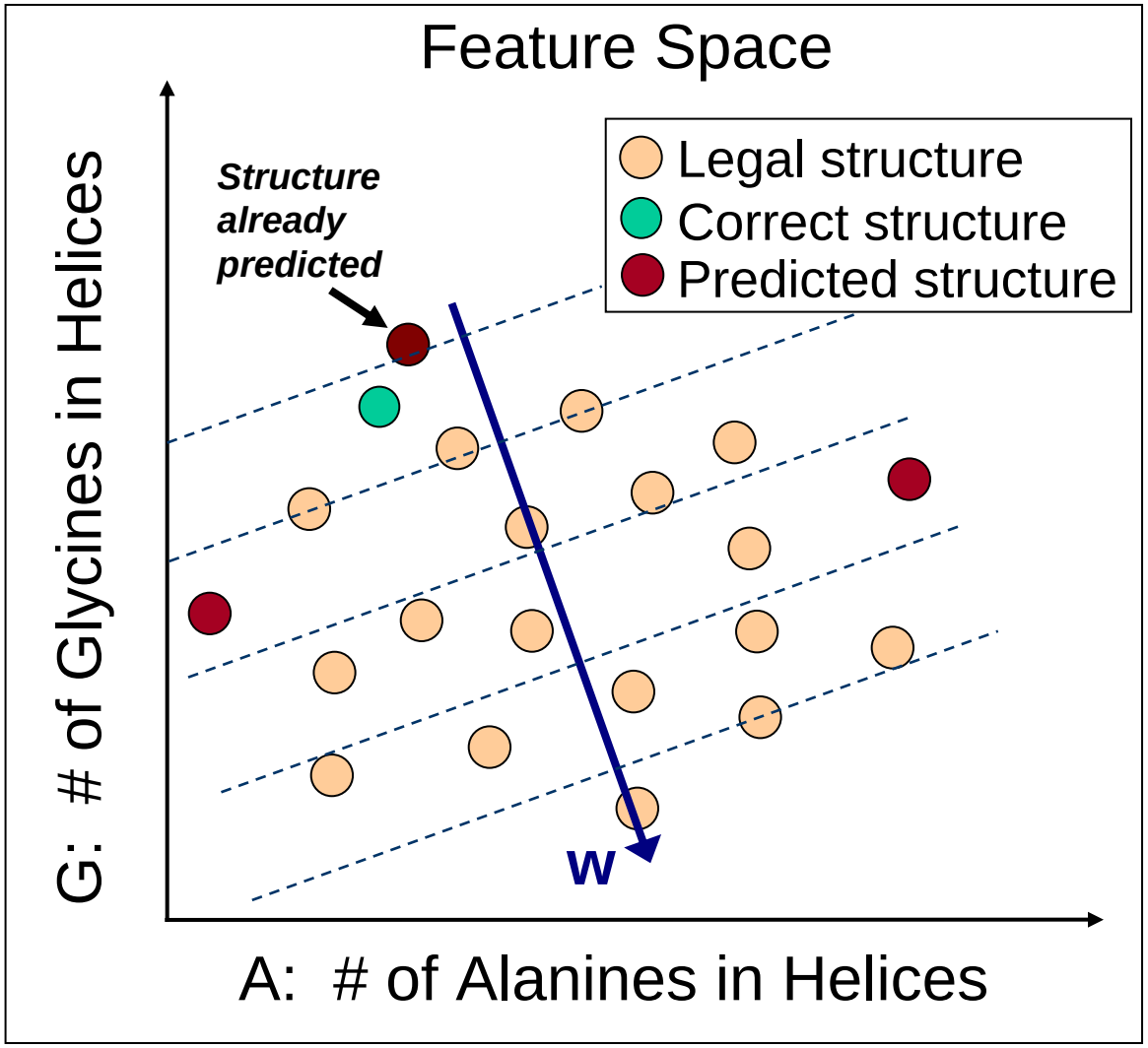
1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters
5. Predict structure
6. Refine parameters

Learning the Parameters



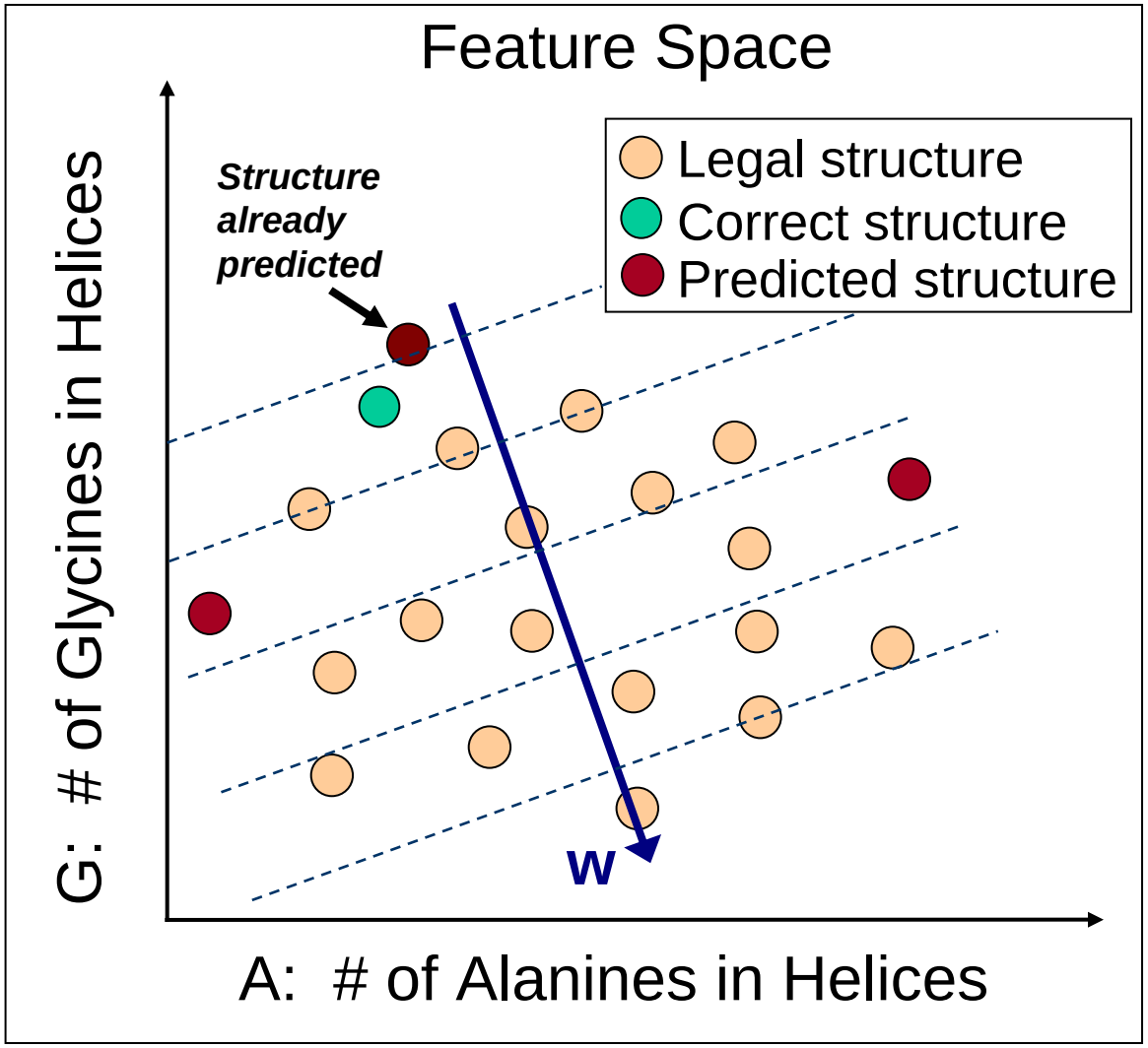
1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters
5. Predict structure
6. Refine parameters

Learning the Parameters



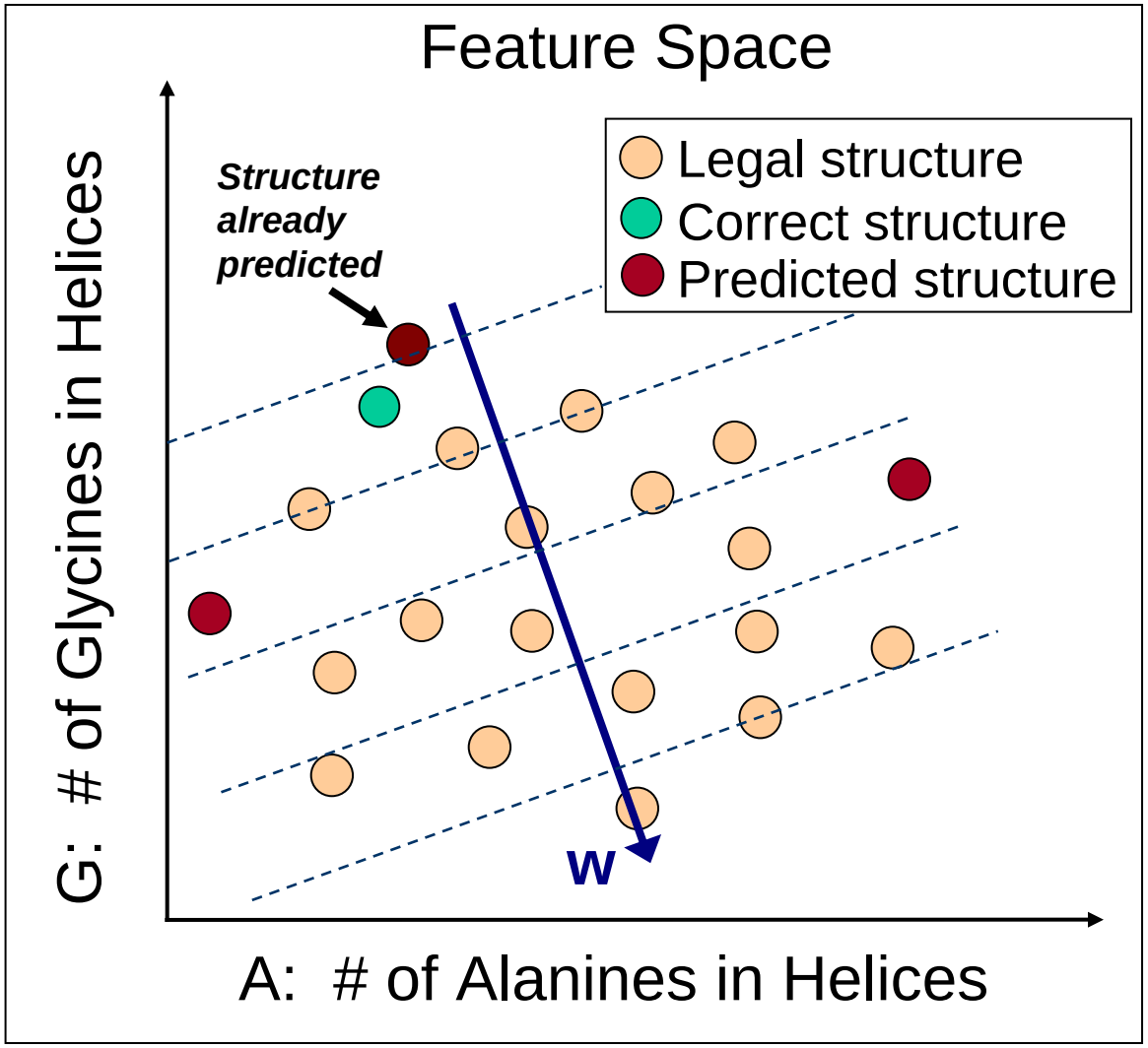
1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters
5. Predict structure
6. Refine parameters
7. Predict structure

Learning the Parameters



1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters
5. Predict structure
6. Refine parameters
7. Predict structure
8. Terminate

Learning the Parameters



1. Predict stucture
2. Refine parameters
3. Predict structure
4. Refine parameters
5. Predict structure
6. Refine parameters
7. Predict structure
8. Terminate

Details in paper:

- How to converge faster
- Early termination condition
- [Tsochantaridis et al., ICML'02]

Experimental Methodology

- **Data set: 300 non-homologous all-alpha proteins**
  - From EVA's sequence-unique subset of the PDB, July 2005
  - Only consider alpha helices ("H" symbol in DSSP)
- **Randomly split into 150 training, 150 test proteins**

Results

Metric	Value	Explanation
Q_α	77.6%	percent of residues correctly predicted
SOV_α	73.4%	segment overlap measure [Zemla'99]

- **Comparison to others**
  - **Best HMM method to date** that does not utilize alignment info
    - Offers 3.5% (Q_α), 0.2% (SOV_α) over previous best [Nguyen02]
  - Lags behind neural networks; e.g., Porter overall SOV = 76.6%
  - However, we could likely gain 6-8% from alignment profiles
- **Caveats**
  - Moving beyond all-alpha proteins, we could suffer 3% [Rost93]
  - By considering 3/10 helices, we could decrease 2% [Jones99]

Conclusions

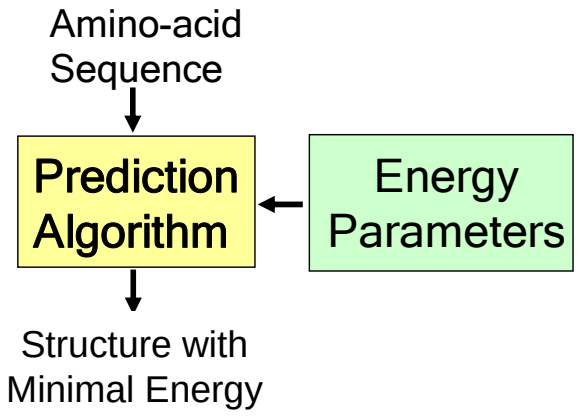
- **Represents first step toward learning biophysical parameters for energy minimization techniques**
  - Iterative, demand-driven learning process using SVMs
- **Promising results on alpha-helix prediction**
  - 77.6% among best Q_α for methods without alignment info
- **Future work: super-secondary structure**
  - Will predict full “contact maps” rather than 3-state labels
  - For beta sheets, replace HMMs by multi-tape grammars

http://protein.csail.mit.edu/

Extra Slides

Prediction Algorithm

- **Parameters represent energetic benefit of a given feature in a protein structure**
  - Features are fixed, chosen by designer
  - Example features:
    - Number of prolines in an alpha helix
    - Number of coils shorter than 2 residues



- **Energy (structure) = $\sum_{\text{features} \in \text{structure}}$ Energy (feature)**
- **Minimal-energy structure found with dynamic prog.**
  - Idea: consider all structures, exploiting overlapping problems
  - Implemented as HMM using Viterbi algorithm

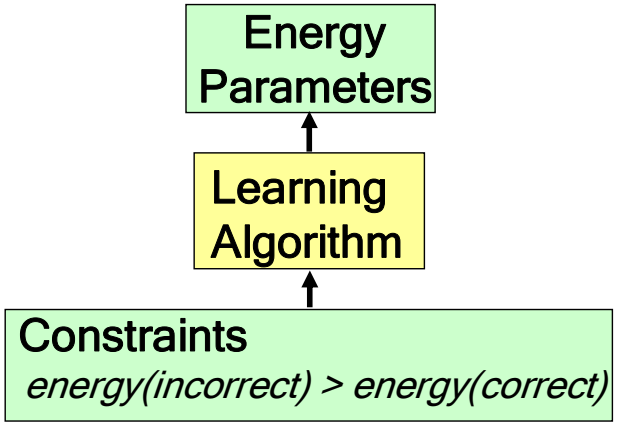
Learning Algorithm

- **Constraints have form:**

  For all incorrectly predicted structures S_i,

  in future selection of the parameters **w**:

  $\text{Energy}_w(S_i) > \text{Energy}_w(\text{correct structure})$



  Constraints are linear in the energy parameters.

- **If feasible, could solve with linear programming**
- **In general, solve with Support Vector Machines (SVMs)**
  - $\text{Energy}(S_i) \geq \text{Energy (correct structure)} + 1 - \xi_i \qquad (\xi_i \geq 0)$
  - Find parameters **w** minimizing $\frac{1}{2} \|\mathbf{w}\|^2 + C/n \sum_{i=1}^{n} \xi_i$

  ➡ Provides general solution using soft-margin criterion