

# MODELLING OF PROTEIN C $\alpha$ TRACES THROUGH STRUCTURAL CONSTRAINTS PREDICTED BY MACHINE LEARNING

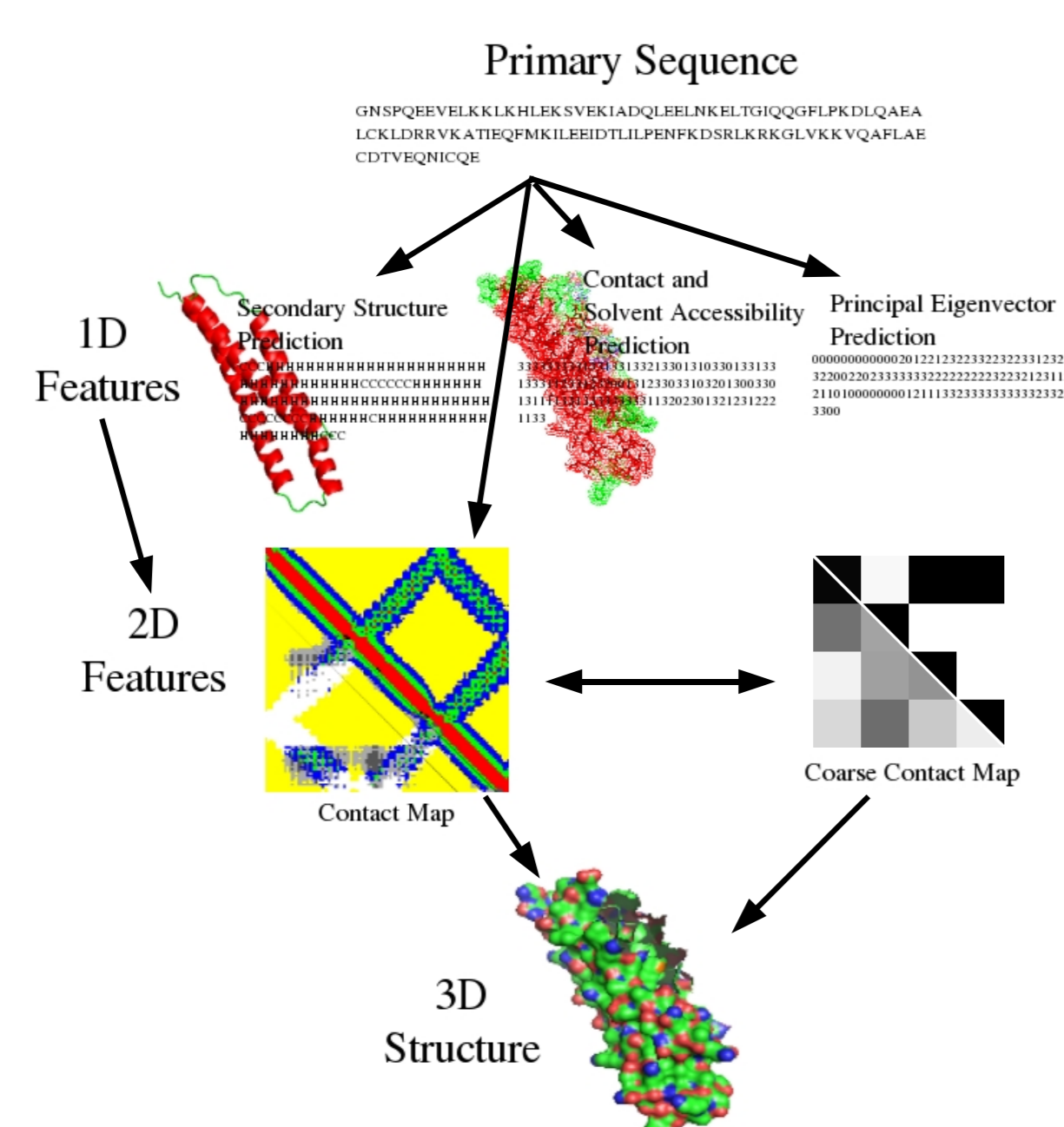
Davide Baù, Alberto J. Martin Martin, Catherine Mooney, Alessandro Vullo, Ian Walsh, Gianluca Pollastri  
{davide.bau|alberto.j.martin|catherine.mooney|alessandro.vullo|ian.walsh|gianluca.pollastri}@ucd.ie



School of Computer Science and Informatics  
University College Dublin  
Belfield, Dublin 4, Ireland



Distill is a suite of servers for the prediction of protein structural features: secondary structure; relative solvent accessibility; contact density; backbone structural motifs; residue contact maps at 6, 8 and 12 Angstrom; coarse protein topology. The servers are based on large-scale ensembles of recursive neural networks and trained on large, up-to-date, non-redundant subsets of the Protein Data Bank. Together with structural feature predictions, Distill includes a server for prediction of C $\alpha$  traces.



Distill's modelling scheme (<http://distill.ucd.ie>). Distill is composed of: a set of state-of-the-art predictors of protein features (secondary structure, relative solvent accessibility, residue contact maps, contact maps between secondary structure elements, structural motifs, contact density)<sup>1-6</sup> based on Machine Learning and trained on large, non-redundant subsets of the PDB; a simple **optimisation algorithm** that searches the space of protein C $\alpha$  trace configurations under the guidance of a potential based on these predicted features. Distill's modelling scheme is fast: on a small cluster of state-of-the-art PCs it can solve protein coordinates on a genomic scale in the order of days.

## Multiclass Contact Maps

A multiclass contact map is an quantised version of the distance matrix (distances among the atoms in a protein), defined as follow:

- For a protein of  $N$  residues, the multiclass contact map is a  $N \times N$  symmetric matrix  $S$ , whose elements  $S_{ij}$  are arrays of length  $c = \text{number of classes}$
- Two residues  $i$  and  $j$  are said to be in the class  $c_x$  whenever the mutual distance between their C $\alpha$  atoms is within the distance boundaries defined for that class.

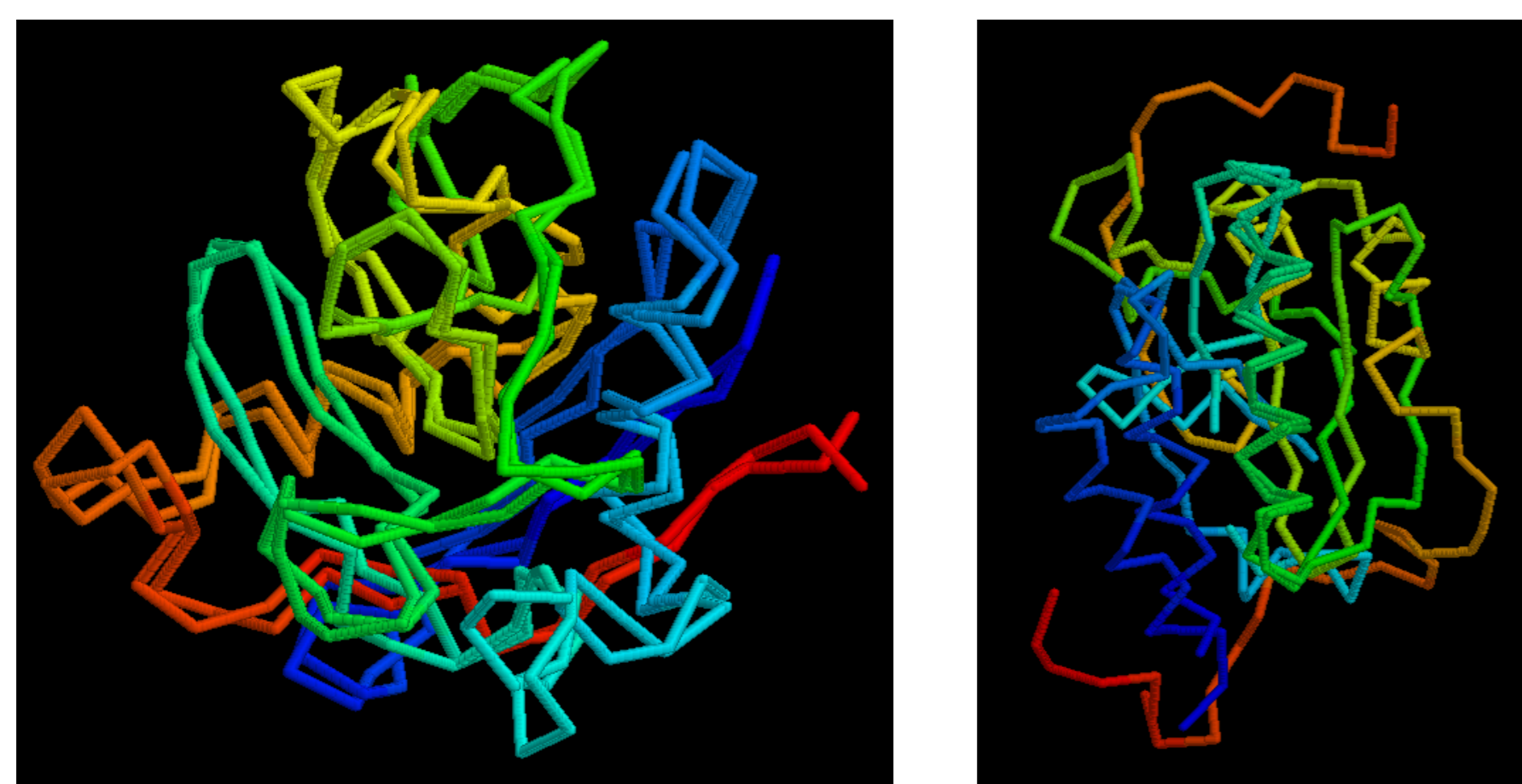
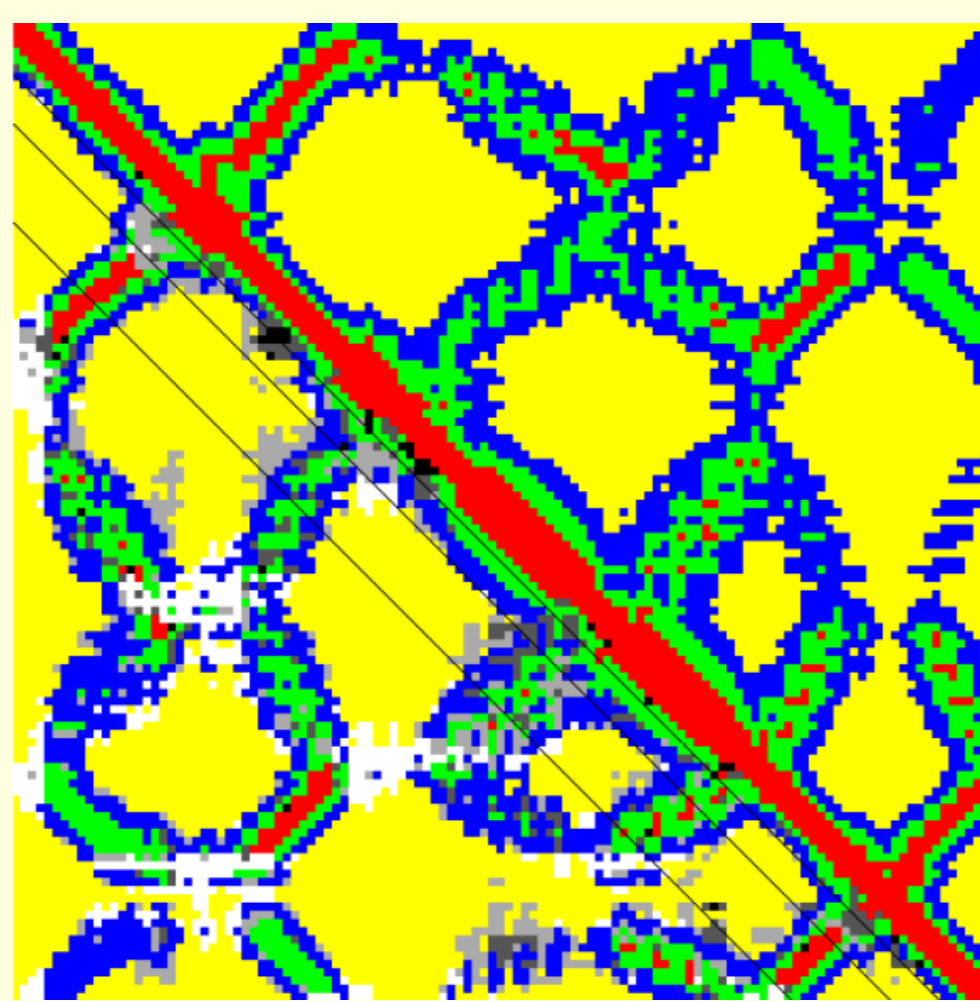


Fig. 1. Examples of reconstruction. Real structure superimposed to the predicted one for CASP targets (CM and FR\_A-NF respectively) T0290 (left, 173 amino acids, TM score = 0.8846) and T0354 (right, 130 amino acids, TM score = 0.3084)

## Acknowledgements

This work is supported by Science Foundation Ireland grants 04/BR/CS0353 and 05/RFP/CMS0029, a UCD President's Award 2004, and an Embark Fellowship to AV from the Irish Research Council for Science, Engineering and Technology

1. G. Pollastri, A. McLysaght. "Porter: a new, accurate server for protein secondary structure prediction" *Bioinformatics*, 21(8):1719-20, 2005
2. A. Vullo, I. Walsh, G. Pollastri. "A two-stage approach for improved prediction of residue contact maps" *BMC Bioinformatics*, 7:180, 2006
3. D. Baù, G. Pollastri, A. Vullo. "Distill: a machine learning approach to ab initio protein structure prediction" in *Analysis of Biological Data: A Soft Computing Approach*, S. Bandyopadhyay, U. Maulik and J. T. L. Wang eds., World Scientific, in press
4. C. Mooney, A. Vullo, G. Pollastri. "Protein Structural Motif Prediction in Multidimensional  $\phi - \psi$  Space leads to improved Secondary Structure Prediction" *Journal of Computational Biology*, in press
5. A. Vullo, O. Bortolami, G. Pollastri, S. Tosatto. "Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines" *Nucleic Acids Research*, 34:W164-W168
6. G. Pollastri, A. Vullo, P. Frasconi, P. Baldi. "Modular DAG-RNN Architectures for Assembling Coarse Protein Structures" *Journal of Computational Biology*, 13:3, 631-650, 2006
7. M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2:295-306, 1997
8. J. Platt (2000) Probabilistic outputs of support vector machines and comparison to regularised likelihood methods. MIT press, Cambridge, MA
9. S. Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L.M. Iakoucheva, M.S. Cortese, J.D. Lawson, C.J. Brown, J.G. Sikes et al. (2005) Disprot: a database of protein disorder, *Bioinformatics*, 21, 137-140

## Abstract

All structural feature predictors are based on single- or dual-layer Recursive Neural Network architectures for Directed Acyclic Graphs (DAG RNNs). One-dimensional feature predictors (i.e. those mapping the primary sequence into a sequence of the same length) are based on 1D DAG RNNs, while contact and distance map predictors are based on 2D DAG RNNs. Secondary structure, solvent accessibility and distance map predictors are provided structural information about PDB templates as a further input, when templates are available.

## Reconstruction Algorithm

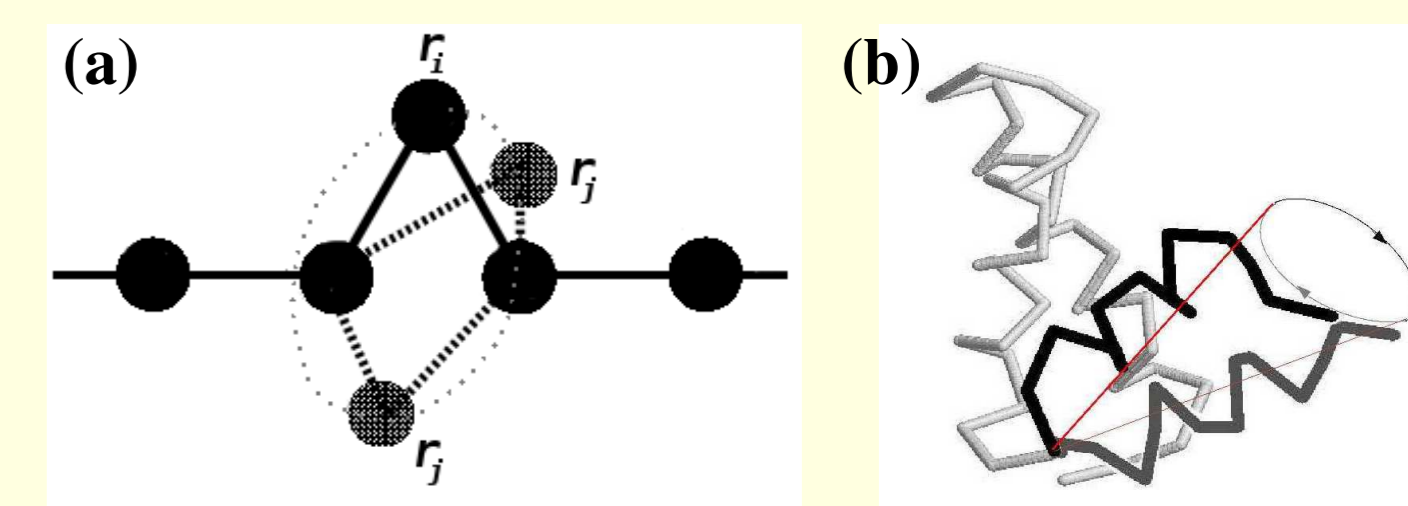
**Goal:** Quickly reconstruct draft protein structures for relatively short sequences ( $L \leq 200$ )  
Proteins are described at a coarse level as their backbone C $\alpha$  trace

### Bootstrap:

- Generation of an initial physically realisable configuration with a self-avoiding random walk and explicit modelling of predicted helices. A random structure is generated by adding C $\alpha$  positions one after the other the whole backbone is represented

### Search:<sup>7</sup>:

- Refinement of the initial bootstrapped structure by global optimisation of a pseudo-energy function using local moves and a simulated annealing protocol
- A randomly chosen point  $i$  is displaced (crankshaft move) (a)
- Secondary structure elements are displaced as a whole, without modifying their geometry (b)



## Accepted Moves

A new set of coordinates  $S^{(t+1)}$  is accepted as the best next candidate with probability  $p = \min(1, e^{\Delta C/T^{(t)}})$  defined by the annealing protocol, where  $\Delta C = C(S^{(t)}, \mathcal{M}) - C(S^{(t+1)}, \mathcal{M})$ ,  $\mathcal{M}$  is the configurational space of physically realisable protein models and  $T^{(t)}$  is the temperature at stage  $t$  of the schedule.

## Spritz<sup>5</sup>: Long Disordered Regions (LD) Predictor

Implemented with probabilistic soft-margin SVM<sup>8</sup>. We use a Gaussian kernel which is less biased than the linear one but also more prone to overfitting The predictor is trained from a subset of completely disordered sequences from DISPROT release 1.2<sup>9</sup> by removing sequences containing errors of annotations and then using the most up to date GI accession numbers. The final set is completed by incorporating an equal sized subset of chains classified as having no disordered fragments and derived from the PDBselect25, March 2002 release.

## Shandy: Domain Predictor

1. Classification: proteins are predicted as single or multidomain (currently implemented with a hard threshold of 160 amino acids)
2. If the protein is considered to be a multidomain one, a 1D-DAG RNN is used to predict residues as boundary or non-boundary
3. Predictions in (2) are smoothed and the location of domain boundaries determined

## Ranker

An index of reliability (estimation of the TM score by an artificial neural network trained in 5-fold cross-validation) of the model is provided in the remark fields of the PDB files, based on:

- the degree to which the model enforces the predicted constraints
- the size and estimated secondary structure composition of the query
- the absence or presence (and, in the latter case, degree) of sequence similarity between the query entries in the PDB

## CASP7 Results

Targets	Ranked 1 <sup>st</sup>	Best submitted
TM score		
CM easy	0.65840	0.66548
CM hard	0.40505	0.42811
FR_H	0.24443	0.25443
FR_A-NF*	0.25573	0.27730
GDT		
CM easy	0.50866	0.51550
CM hard	0.30451	0.32338
FR_H	0.19810	0.20754
FR_A-NF*	0.25632	0.27619

Table 1. TM score and GDT for CASP7 targets. *Ranked 1<sup>st</sup>*: predicted structure ranked as first by our ranker. *Best submitted*: actual best submitted reconstruction (due to a glitch, we used the 2005 version of the PDB database for CASP predictions).

Targets	Ranked 1 <sup>st</sup>	Best submitted
TM score		
CM easy	0.68833	0.69497
CM hard	0.43766	0.45888
FR_H	0.25102	0.25543
FR_A-NF	0.25723	0.28248
GDT		
CM easy	0.53464	0.54169
CM hard	0.33330	0.35040
FR_H	0.20289	0.20546
FR_A-NF	0.25631	0.27903

Table 2. Same as Table 1, but using an updated version of the PDB database (last updated before the beginning of CASP experiments).

\*Four of our FR\_A-NF target predictions have been ranked in the top 10.

## Conclusions

- The pipeline is able to generate topologically correct folds in several cases
- Although Distill's 3D models are often still crude, nonetheless they are often informative
- Distill's modelling scheme is fast: on a small cluster of state-of-the-art PCs it can solve protein coordinates on a genomic scale in the order of days

## Current Research Directions

- Improve cost function and reconstruction strategy
  - Learning to adapt the energy function or distance of a candidate decoy to the native one
  - Include more physical constraints
- Improve contact map prediction
  - Better  $\beta$ -sheets interactions, long-range contacts predictions
  - Remove spurious contacts