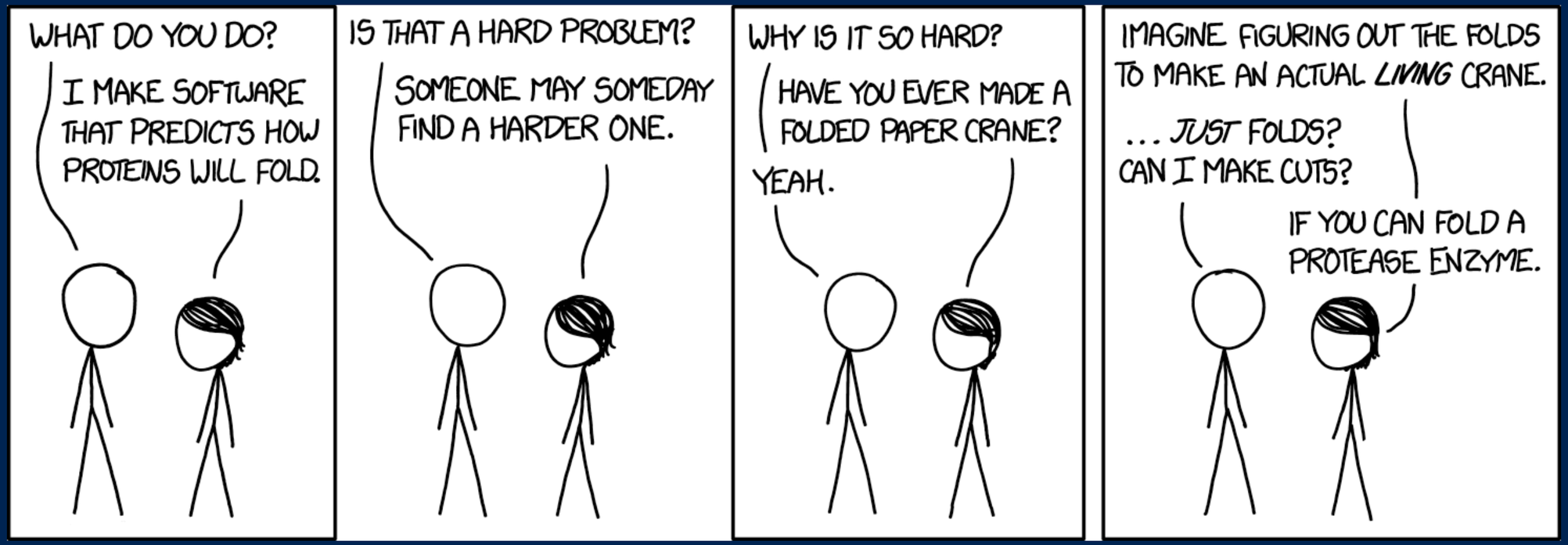
alphafold

brettkoonce.com/talks

april 15, 2019


WHAT DO YOU DO?
I MAKE SOFTWARE THAT PREDICTS HOW PROTEINS WILL FOLD.
IS THAT A HARD PROBLEM?
SOMEONE MAY SOMEDAY FIND A HARDER ONE.
WHY IS IT SO HARD?
HAVE YOU EVER MADE A FOLDED PAPER CRANE?
YEAH.
IMAGINE FIGURING OUT THE FOLDS TO MAKE AN ACTUAL *LIVING* CRANE.
... *JUST* FOLDS? CAN I MAKE CUTS?
IF YOU CAN FOLD A PROTEASE ENZYME.

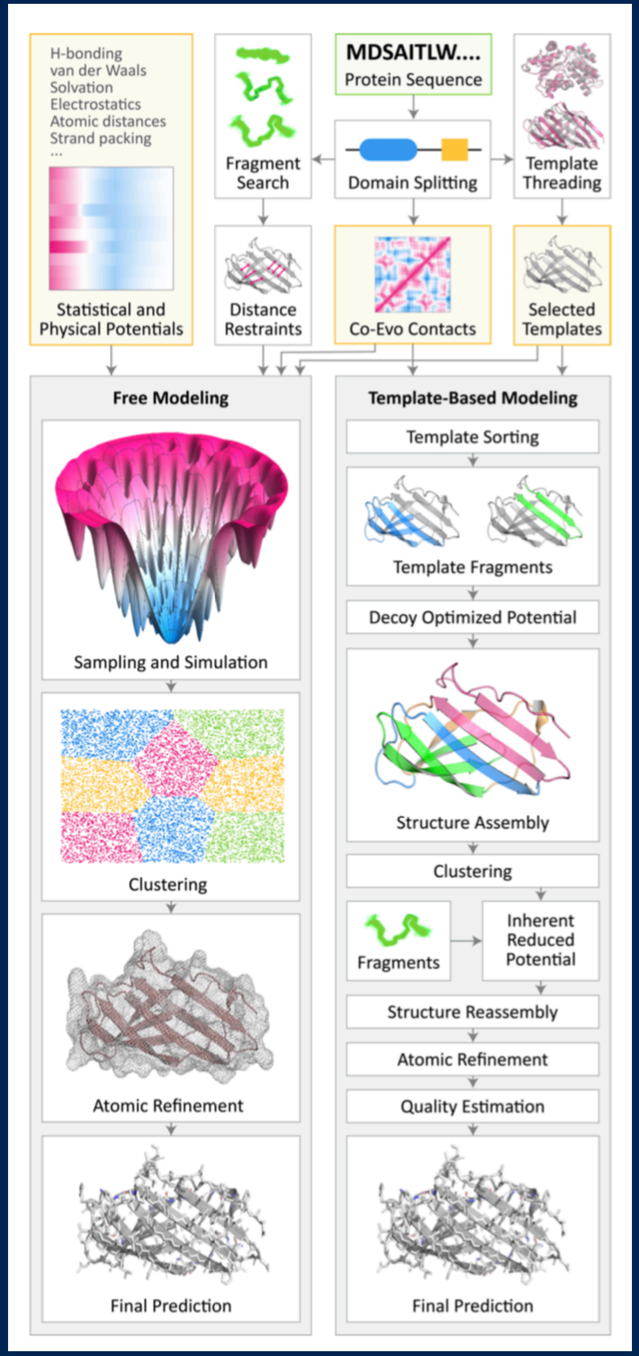
overview

- protein modeling, casp
- alphafold
- draw, torsion backbone, simulated annealing
- coevolutionary residues, scoring networks
- energy models

protein modeling

- sequence —> ?? —> model
- model —> predict drug interactions, de novo proteins
- experimental processes to do, but $$$
- ideal: input —> computer —> model —> science —> profit!

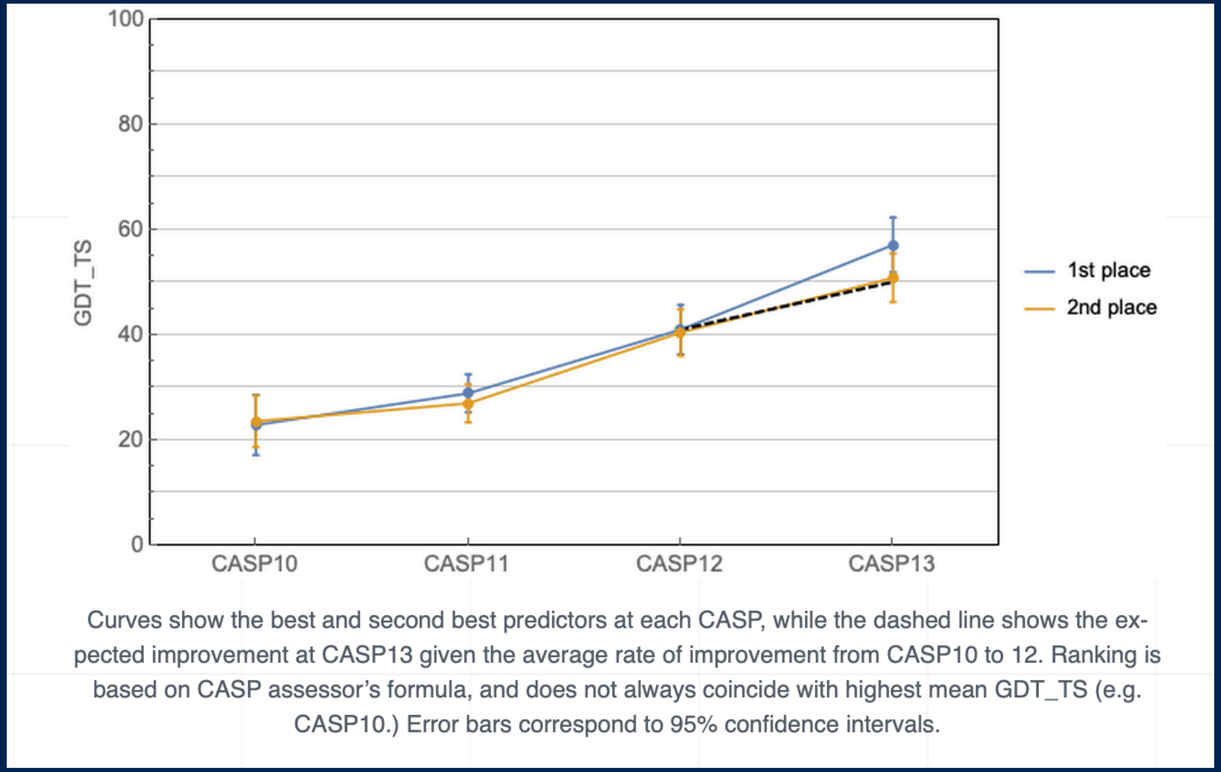
ab initio



homology

casp

- every two years
- competition, groups from around world
- given known sequences —> models —> recent undisclosed protein —> results



Curves show the best and second best predictors at each CASP, while the dashed line shows the expected improvement at CASP13 given the average rate of improvement from CASP10 to 12. Ranking is based on CASP assessor's formula, and does not always coincide with highest mean GDT_TS (e.g. CASP10.) Error bars correspond to 95% confidence intervals.

alphafold

- nn 1: draw model to generate fragments
- simulated annealing to combine
- nn 2a: inter-residue distances
- nn 2b: scoring network
- relaxation, nn 3: final protein scoring

draw

- vae + attention model
- backbone + torsion angles

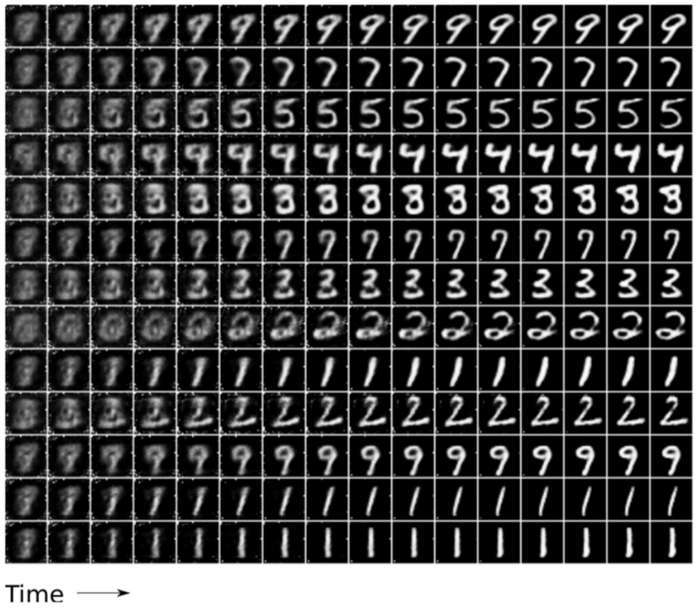


Figure 7. **MNIST generation sequences for DRAW without attention.** Notice how the network first generates a very blurry image that is subsequently refined.

simulated annealing

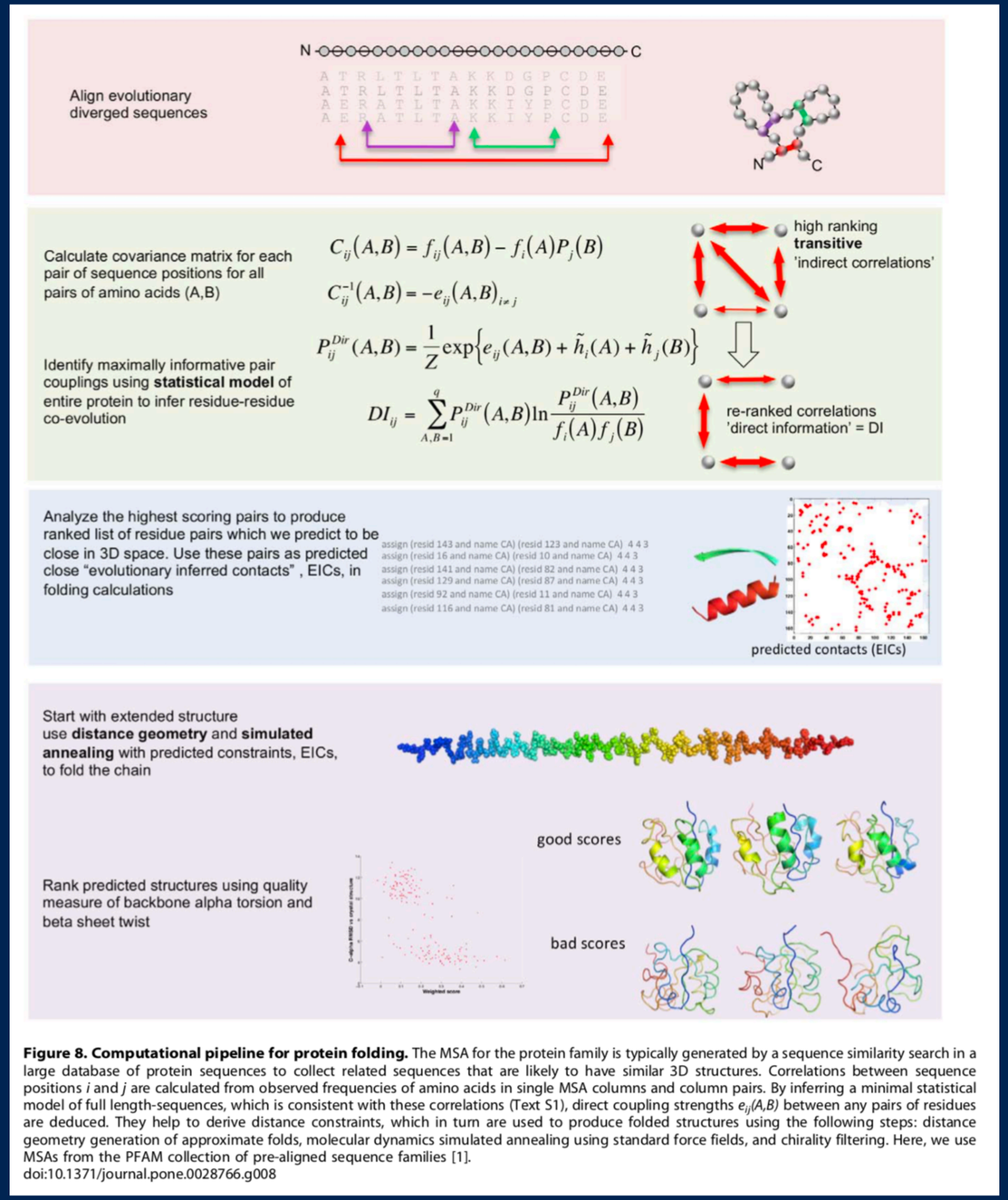
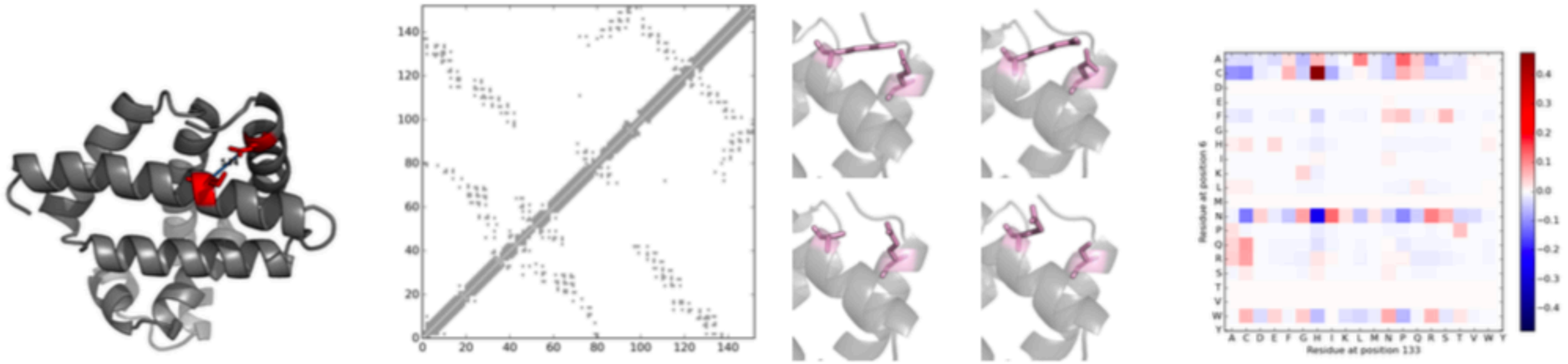


Figure 8. Computational pipeline for protein folding. The MSA for the protein family is typically generated by a sequence similarity search in a large database of protein sequences to collect related sequences that are likely to have similar 3D structures. Correlations between sequence positions *i* and *j* are calculated from observed frequencies of amino acids in single MSA columns and column pairs. By inferring a minimal statistical model of full length-sequences, which is consistent with these correlations (Text S1), direct coupling strengths $e_{ij}(A,B)$ between any pairs of residues are deduced. They help to derive distance constraints, which in turn are used to produce folded structures using the following steps: distance geometry generation of approximate folds, molecular dynamics simulated annealing using standard force fields, and chirality filtering. Here, we use MSAs from the PFAM collection of pre-aligned sequence families [1].

doi:10.1371/journal.pone.0028766.g008

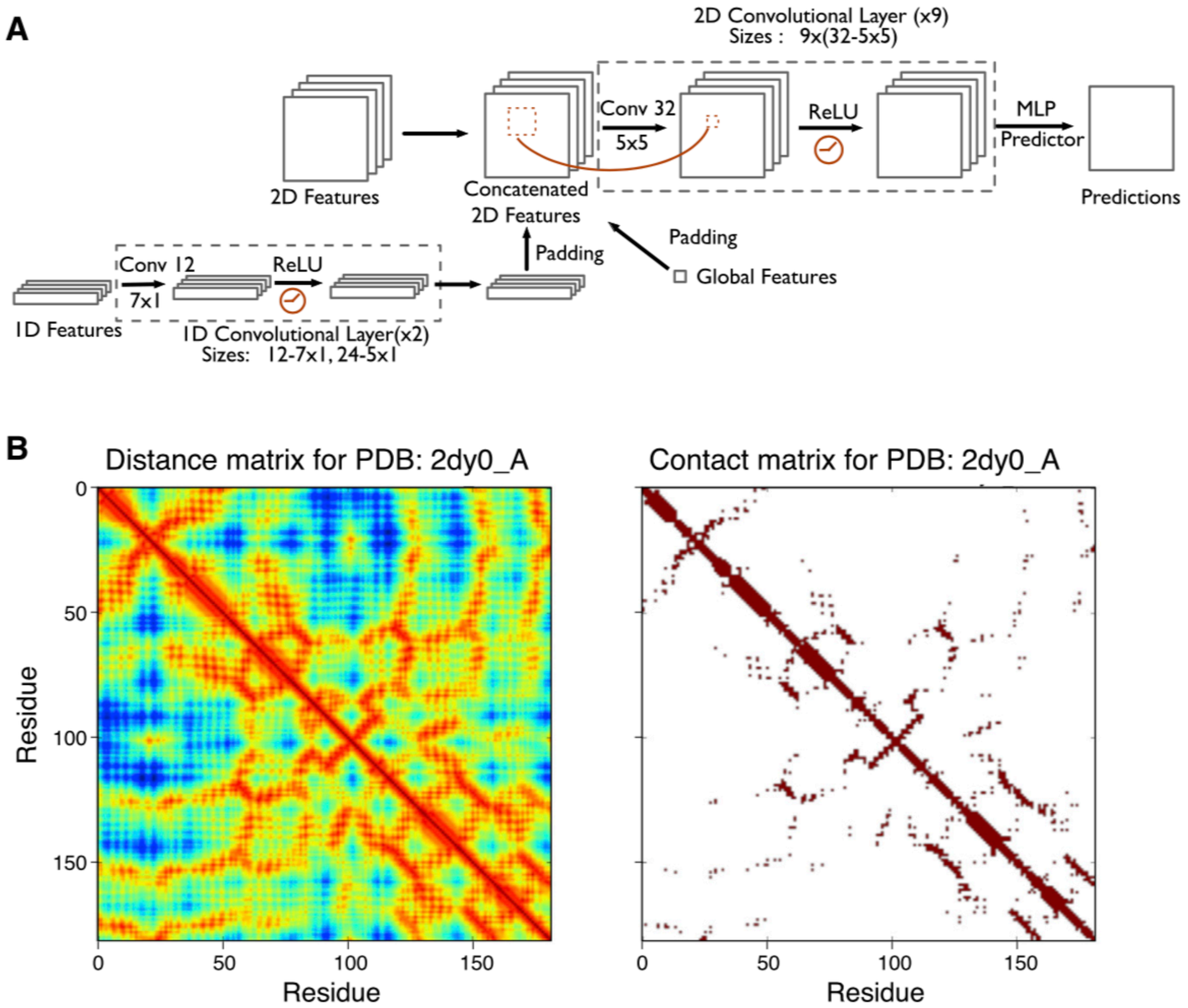
coevolution stats



(a) Tertiary structure (b) Contact map (c) Variants of contact (d) Co-evolution statistics

Figure 1: Oxymyoglobin (a) and its contact between amino acid residue 6 and 133. Helix–helix contacts correspond to "checkerboard" patterns in the contact map (b). Various variants of the contact 6/133 encountered in nature (native pose in upper left, remaining poses are theoretical models) (c) are reflected in the co-evolution statistics (d).

deep contact



accurate de novo prediction of protein contact maps

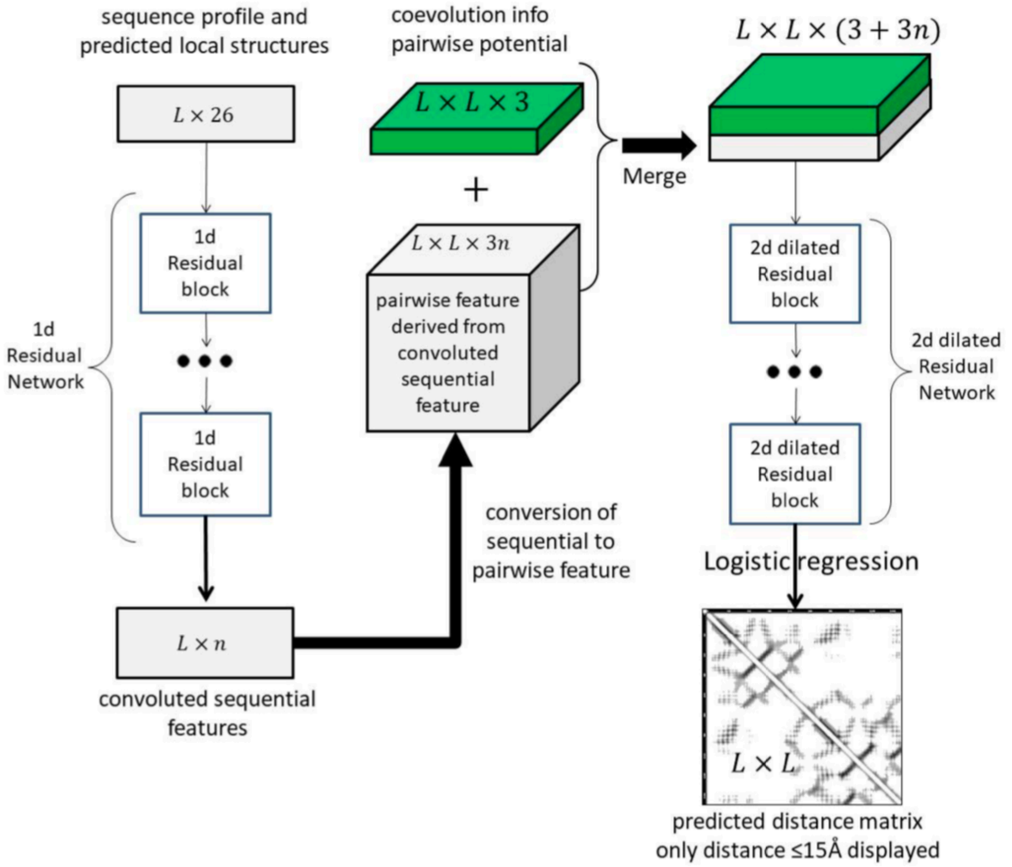


Figure S1. The overall deep network architecture for the prediction of protein distance matrix. The left column is a 1D deep residual neural network that transforms sequential features (e.g., sequence profile and predicted secondary structure). The right column is a 2D deep dilated residual neural network that transforms pairwise features. The middle column converts the convoluted sequential features to pairwise features and combine them with the original pairwise features. The picture is adapted from Figure 1 in the paper at https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324 .

scoring network

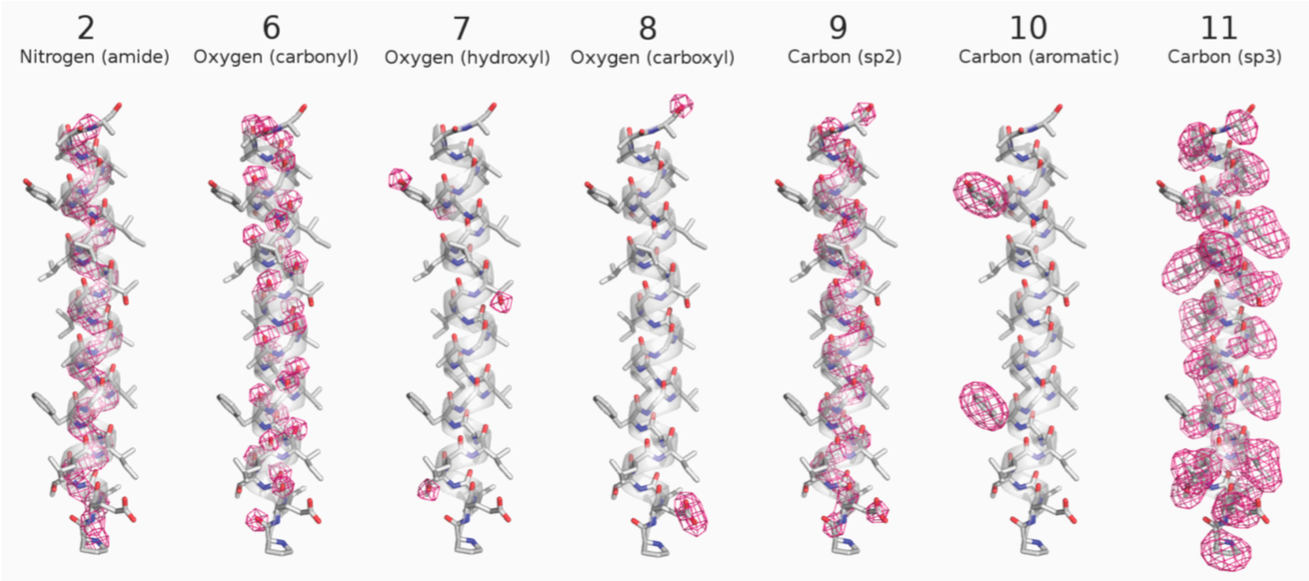


Figure 2. Representation of a protein structure (PDB code 5eh6) using atomic densities. The density maps are calculated according to Eq. 1 and rendered using Pymol [33] with an isosurface level of 0.5.

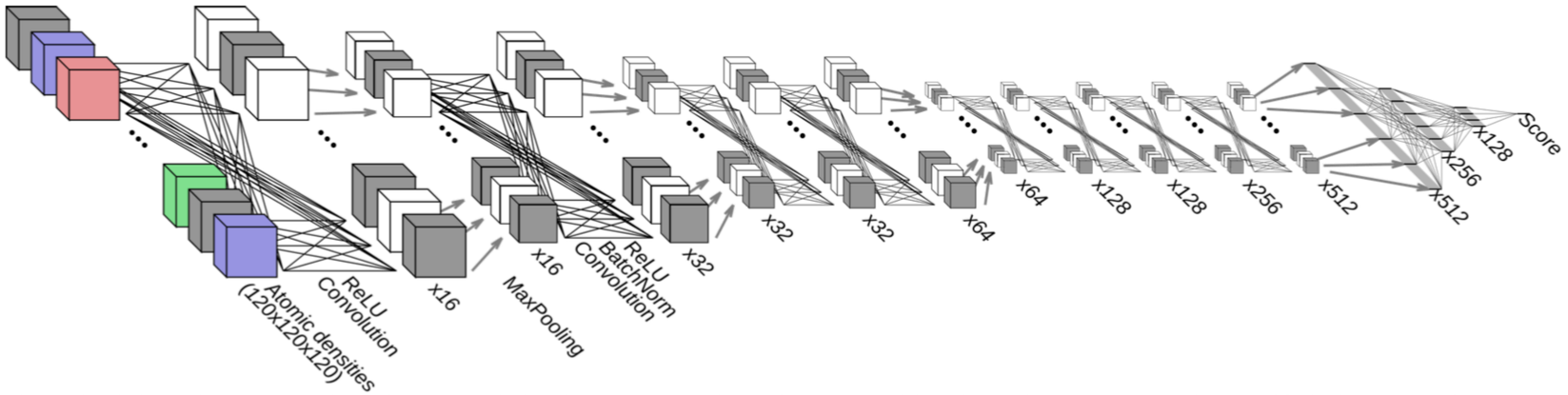


Figure 3. Schematic representation of the convolutional neural network architecture used in this work. Unless otherwise specified, line connections across boxes denote the consecutive application of a 3D convolutional layer ("Convolution"), a batch normalization layer ("BatchNorm"), and a ReLU layer. Grey arrows between boxes denote maximum pooling layers ("MaxPooling"). Labels "$\times M$" denote the number of 3D grids and the number of filters used in the corresponding convolutional layer. The grey stripes denote one-dimensional vectors and crossed lines between them stand for fully-connected layers with ReLU nonlinearities. Details of the model can be found in Table S3 of Supplementary Information.

relaxation

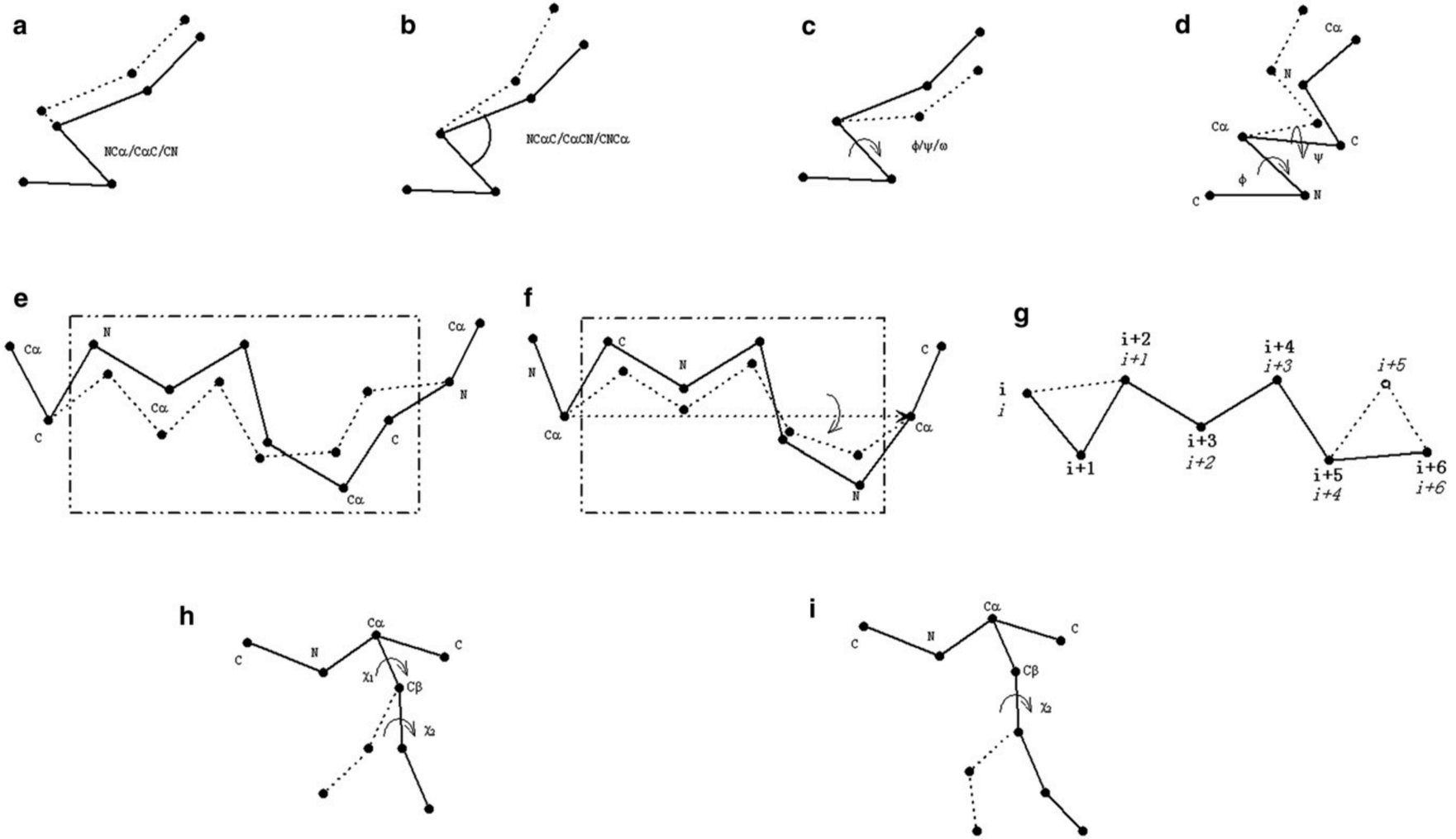


FIGURE 3 Illustration of movements used for the main-chain simulation (*a–g*) and full-atomic simulation (*a–i*). New positions of atoms after movements are connected by dash lines. New residue numbers after the shift in *g* are in italic type.

refinement

nemo

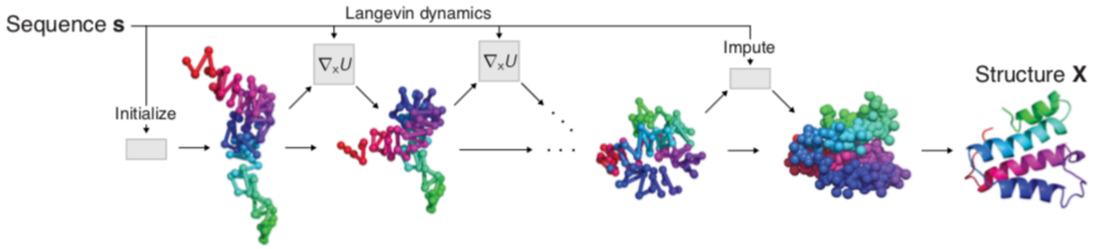


Figure 1: **An unrolled simulator as a model for protein structure.** NEMO combines a neural energy function for coarse protein structure, a stochastic simulator based on Langevin dynamics, and an atomic imputation network to build atomic coordinate output from sequence information. It is trained end-to-end by backpropagating through the *unrolled* folding simulation.

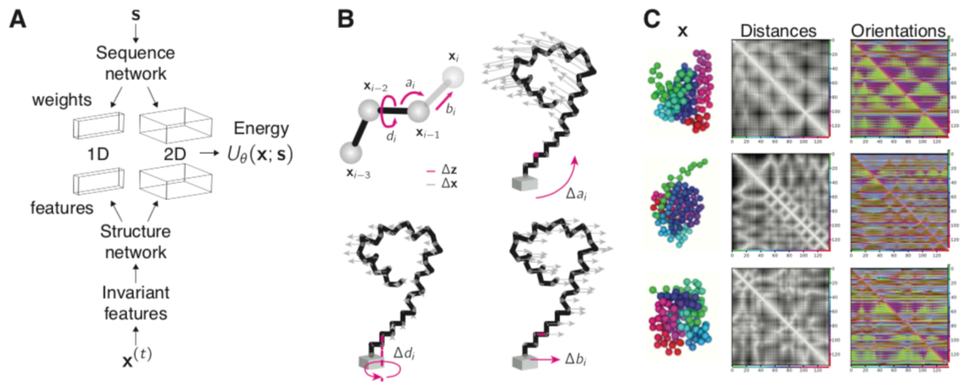


Figure 2: **A neural energy function models coarse grained structure and is sampled by internal coordinate dynamics.** (A) The energy function is formulated as a Markov Random Field with structure-based features and sequence-based weights computed by neural networks (Figure 6). (B) To rapidly sample low-energy configurations, the Langevin dynamics simulator leverages both (i) an internal coordinate parameterization, which is more effective for global rearrangements, and (ii) a Cartesian parameterization, which is more effective for localized structural refinement. (C) The base features of the structure network are rotationally and translationally invariant internal coordinates (not shown), pairwise distances, and pairwise orientations.

3d mnist demo

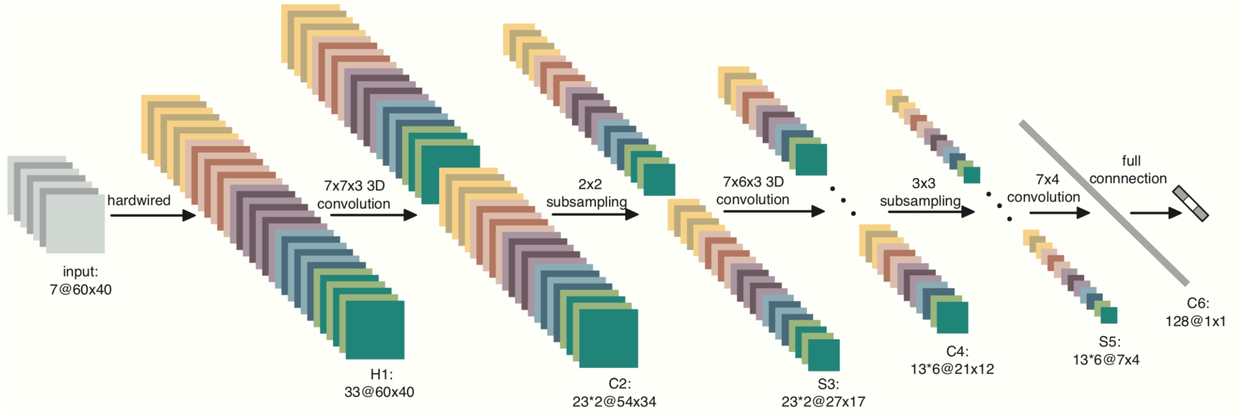


Figure 3. A 3D CNN architecture for human action recognition. This architecture consists of 1 hardwired layer, 3 convolution layers, 2 subsampling layers, and 1 full connection layer. Detailed descriptions are given in the text.

- **medium.com/shashwats-blog/3d-mnist-b922a3d07334**

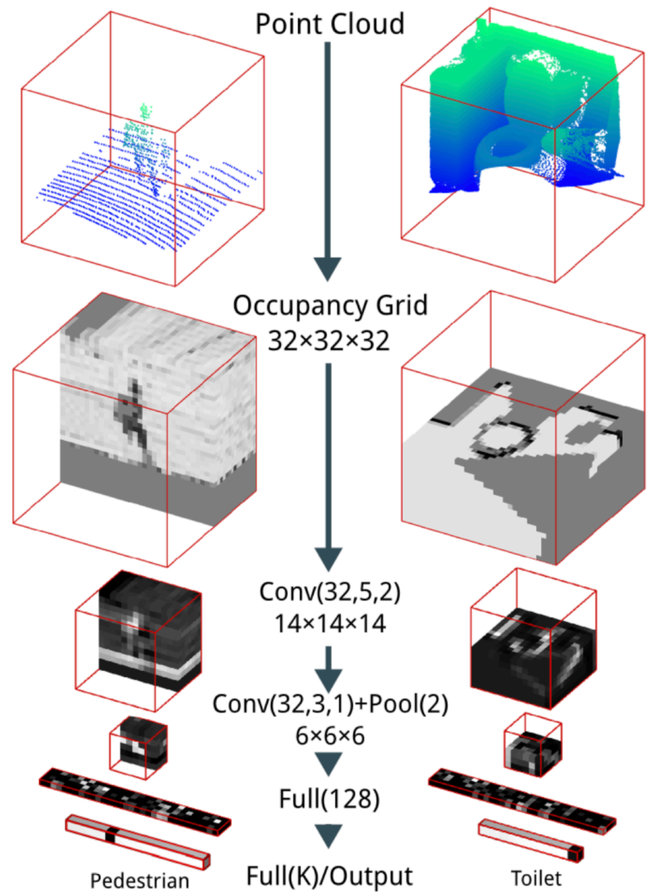


Fig. 1. The VoxNet Architecture. $Conv(f, d, s)$ indicates f filters of size d and at stride s, $Pool(m)$ indicates pooling with area m, and $Full(n)$ indicates fully connected layer with n outputs. We show inputs, example feature maps, and predicted outputs for two instances from our experiments. The point cloud on the left is from LiDAR and is part of the Sydney Urban Objects dataset [4]. The point cloud on the right is from RGBD and is part of NYUv2 [5]. We use cross sections for visualization purposes.

conclusion

- protein modeling, alphafold overview
- traditional approaches, big data, new techniques
- end to end pipelines
- field needs more eyeballs!

thanks for coming!

links

- moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened
- youtube.com/watch?v=HOVdHAnC8LI
- youtube.com/watch?v=R20_s8XPw8U