Say it with proteins: an alphabet of crystal structures

From manual searching of the Protein Data Bank (PDB), I have curated a set of protein crystal structures corresponding to the capital letters of the Roman alphabet (**Fig. 1**). In choosing structures, I aimed to include a range of different structural motifs and to exclude nucleic acids or proteins solved while bound to nucleic acids. Sometimes these letter shapes seem to be incidental, and sometimes the shape is key to the protein's biological function. For example, the specific shape is likely to be important for *L* (from elongation factor P), which mimics the shape of tRNA; for the sinuous *W* (from DNA-binding domain from BurrH), which tracks DNA's major groove for modular sequence recognition; and for proteins with holes that enclose DNA (*A*, from DNA gyrase) or puncture the membrane (*O*, from the toxin cytolysin A). PDB accession codes and descriptions of function for all proteins are provided in **Supplementary Table 1**. This set may be useful for outreach and teaching, by drawing attention to the diversity of protein structures attained by natural selection. It is conceivable that the set may also have value in bionanotechnology and synthetic biology, in which at times molecular assembly needs a specific shape more than a specific function.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nsmb.3011).

ACKNOWLEDGMENTS
Funding was provided by Worcester College Oxford. I thank E. Lowe (University of Oxford) for the diffraction image.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

Mark Howarth
Department of Biochemistry, Oxford University, Oxford, UK.

e-mail: mark.howarth@bioch.ox.ac.uk

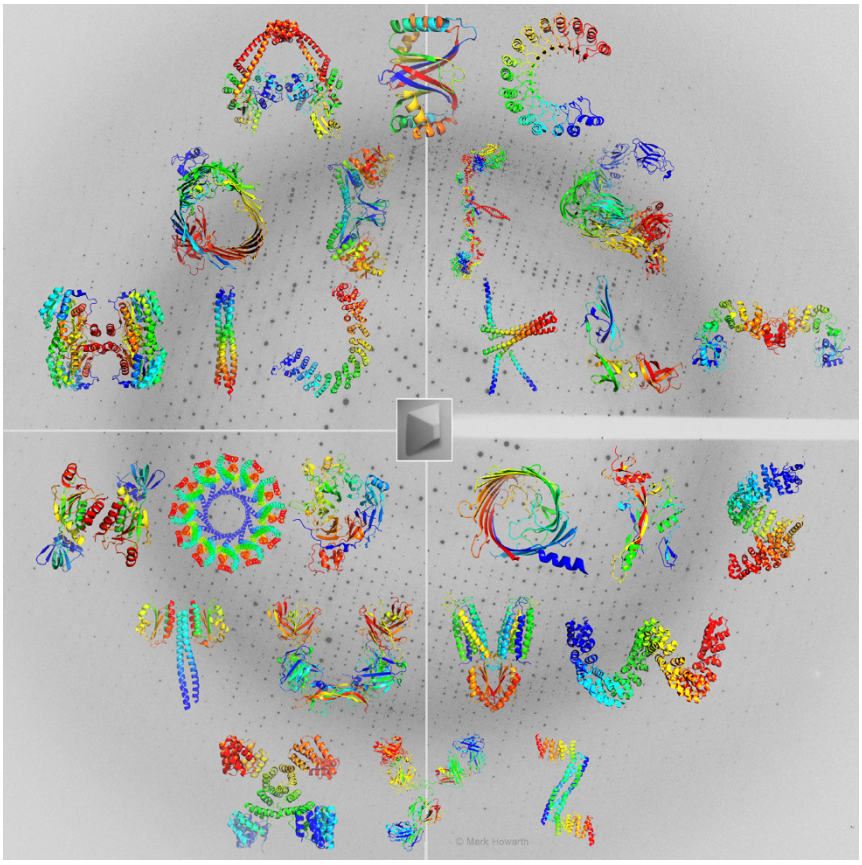

Figure 1 A protein alphabet. Selected protein crystal structures from the PDB in cartoon format and alphabetical order, overlaid on a diffraction image (provided by E. Lowe), with a central bright-field image of a protein crystal. Proteins are colored with the chainbows format, with the N terminus of each chain in blue through to the C terminus in red. Proteins are shown in monomeric (*C*, *D*, *G*, *J*, *L*, *P*, *Q*, *W*), homodimeric (*A*, *B*, *E*, *M*, *N*, *T*, *V*, *X*), heterodimeric (*R*, *S*), homotrimeric (*I*), heterotrimeric (*Z*), homotetrameric (*H*, *K*), heterotetrameric (*Y*), heterohexameric (*F*, *U*) or homododecameric (*O*) forms.