ScanNet uncovers binding motifs in protein structures with deep learning

Determining the functional properties of a protein from its structure is challenging. This study presents an interpretable deep learning model that directly learns function-bearing structural motifs from raw data, allowing accurate mapping of protein binding sites and antibody epitopes onto a protein structure.

This is a summary of:
Tubiana, J., Schneidman-Duhovny, D. & Wolfson, H. J. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat. Methods* https://doi.org/10.1038/s41592-022-01490-7 (2022)

Published online:
1 June 2022

The problem

Although experimental and in silico protein structure determination techniques are rapidly improving[1], determining the function of a protein from its structure remains a challenge. Identifying structural features such as protein–protein binding sites — which are important therapeutic targets — or epitopes on viral proteins, which bind to antibodies and drive the immune response to viral pathogens, can give insights into protein function. To identify such structures, researchers have so far relied on two classes of methods: comparative approaches, which involve comparing a query protein with similar, previously annotated proteins; and machine learning approaches that consider specific properties of the sequence and structure of the protein, such as electrostatic charge, hydrophobicity, molecular surface curvature or solvent accessibility. However, these approaches are limited in terms of accuracy, throughput and coverage.

The solution

In proteins, functionality is borne by local structural motifs — spatially contiguous sets of amino acid arranged in a specific fashion. Examples of such motifs include catalytic triads responsible for enzymatic activity or the zinc fingers that are a signature of nucleic acids binding sites. Therefore, definition and identification of such motifs are the two cornerstones of structure-based function prediction. We developed a neural network architecture called ScanNet (Spatio-Chemical Arrangement of Neighbors Network) to achieve these two goals with a high level of interpretability. ScanNet extracts local atomic and amino acid neighborhoods from structural data and passes them through trainable motif-detecting filters.

We trained ScanNet on large data sets of annotated protein structures to detect protein–protein and protein–antibody binding sites, and found it to be significantly more accurate than previous approaches based on comparative modeling or feature-based machine learning. In particular, ScanNet could accurately predict the epitopes of the SARS-CoV-2 spike protein (Fig. 1). The filters learned by the network could be readily visualized and interpreted: we found simple patterns such as hydrogen bonds, secondary structure elements and exposed hydrophobic residues, as well as more complex ones such as hotspot 'O-rings' (Fig. 1). The complex representation learned was a posteriori found to correlate with many known physicochemical features, such as solvent accessibility or electrostatic potential. Taken together, our experiments suggest that ScanNet successfully learned some of the fundamental physicochemical principles underlying protein–protein interactions.

The implications

Our findings have multiple short-term applications. Rapid prediction of protein–protein binding sites without knowledge of specific binding partners could facilitate the systematic design of targeted protein binders, which are practical reagents for experimental investigation and potential therapeutics[2]. Further, predicting the antibody epitope distribution of a given antigen could have broad impact in the field of computational immunology. Indeed, we recently leveraged ScanNet to investigate the impact of SARS-CoV-2 variant mutations on the humoral response and, in a recent preprint article, showed that the receptor-binding domain of the spike protein — the main region targeted by antibodies — has significantly reduced antigenicity in the Omicron variant compared to previous strains[3]. Our finding was confirmed by controlled immunization experiments and could explain the high breakthrough infection rates and low efficiency of Omicron-targeted vaccines[4]. Another potential of ScanNet is the design of non-immunogenic therapeutic proteins: by predicting overall antigenicity levels, ScanNet could facilitate the identification of candidate proteins at high risk of inducing an adverse immune response.

There are three main limitations to the presented work. First, ScanNet relies on the availability of a defined structure for the target protein, and a substantial fraction of the human proteome is disordered, consisting of proteins that do not adopt a well-defined protein structure in isolation. For such disordered proteins, sequence-based methods for function prediction remain the most suited. Second, the model is currently limited in its ability to discover additional function-bearing structural motifs by the prediction task and the amount of labeling data available. Self-supervised learning is a promising future direction towards a complete dictionary of function-bearing motifs found in nature. Finally, ScanNet is yet to be adapted for prediction of interactions between specific protein binding partners, a problem of paramount interest to the scientific community.

Jérôme Tubiana
Tel Aviv University, Tel Aviv, Israel

EXPERT OPINION

"The approach is novel and integrates several clever ideas that result in a network architecture that is well suited for protein structure. It has a structural inductive bias by virtue of a hierarchical processing of atoms and residues. The network fully encodes the geometry of the structure. I suspect these ideas will be more broadly applicable to other learning problems in structural biology, and the approach will be of great interest to the community." **David Koes, University of Pittsburgh, Pittsburgh, PA, USA**

FIGURE

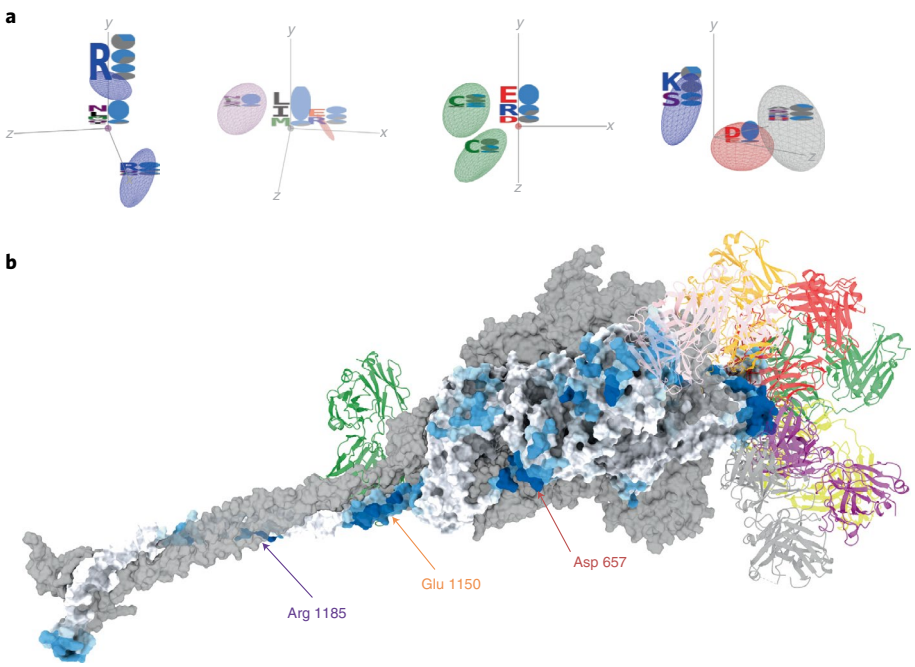

Fig. 1 | Illustration of the ScanNet model. a, Selected visualizations of spatiochemical patterns learned at the amino acid scale for the prediction of antibody binding sites. Patterns are defined by the presence of specific sets of amino acids (shown as colored letters with height proportional to their frequency) at prescribed locations (depicted as Gaussian ellipsoids) and in prescribed conformation (fully solvent-exposed or buried, respectively shown as blue or gray circles). **b**, Antibody binding site predictions overlaid on the SARS-CoV-2 spike protein trimer (surface representation, colored by probability: white, low; dark blue, high). Arrows highlight predicted epitopes outside the main immunogenic regions. Representative antibodies are shown in cartoon form. © 2022, Tubiana, J. et al.

BEHIND THE PAPER

Like many other groups, we identified deep learning as a promising opportunity for providing scientific breakthroughs. However, I vastly underestimated the difficulty of applying deep learning to protein structures. One challenge is caused by the peculiar nature of protein structures: in our first attempts, we did not even manage to recognize α-helices and β-sheets, an a priori trivial task! Benchmarking the software was a major hurdle, with no established data sets, multiple inference settings, and dozens of baseline methods developed over the past two decades. Achieving a high level of interpretability was especially difficult, and we considered giving up on this objective; however, we realized we needed this aspect to clearly demonstrate that deep learning can learn structural motifs or physics. Altogether, this took much longer and expected! I am especially grateful to Prof. Wolfson and my funders, the Safra Center for Bioinformatics and the Human Frontier Science Program Organization, for their patient support. **J.T.**

REFERENCES

1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
The AlphaFold paper presents breakthrough results in protein structure prediction.

2. Cao, L. et al. Design of protein binding proteins from target structure alone. *Nature* https://doi.org/10.1038/s41586-022-04654-9 (2022).
This article presents a hybrid computational–experimental protocol for generic design of protein binders.

3. Tubiana, J. et al. Reduced antigenicity of Omicron lowers host serologic response. Preprint at https://doi.org/10.1101/2022.02.15.480546 (2022).
In this preprint article, we leverage ScanNet to monitor the impact of mutations on SARS-CoV-2 antigenicity.

4. Servellita, V. et al. Neutralizing immunity in vaccine breakthrough infections from the SARS-CoV-2 Omicron and Delta variants. *Cell* https://doi.org/10.1016/j.cell.2022.03.019 (2022).
This article reports analysis of sera from SARS-CoV-2 vaccine breakthrough infection for variants, showing that immune response is especially weak for Omicron.

FROM THE EDITOR

"Deep learning has taken protein structure prediction capabilities to new heights. When the ScanNet paper from Tubiana et al. was submitted, it stood out to me because of its performance for protein binding site prediction while maintaining interpretability of underlying chemical principles of binding. My hope is that the method will help further elucidate the core principles underlying protein binding and function." **Arunima Singh**, **Senior Editor**, ***Nature Methods***