on solvent and temperature, either N- or C-coordination occurs without loss of dinitrogen, as studied by Hansmann and co-workers. In the reaction with organometallic Au(I) precursors, Severin and colleagues observed the coordination of one molecule of diazoolefin via the central carbon atom when THF was used as the solvent. In contrast, the team led by Hansmann noted the formation of a dinuclear Au(I) complex as well as an azo-bridged dimer of the mesoionic N-heterocyclic olefin when diethyl ether was the solvent. Whether this divergent reactivity is related to the N-heterocyclic backbone of the diazoolefin or arises from the change in solvent remains unclear and requires further investigation. Additional studies by Severin and co-workers on the formation of metal complexes of diazoolefins revealed that they all feature a surprisingly stable diazo functional group, which remains untouched for complexes based on Pd(II), Rh(I) or Al(III).

As described in 1987 by Bott, such (mesoionic) N-heterocyclic diazoolefins are remarkably stable. To evaluate this property, the use of these diazoolefins in organic synthesis was studied (Fig. 1d). In a few initial applications, the Severin group highlighted the reaction with different electrophiles in [2+3] cycloaddition reactions of dimethyl acetylenedicarboxylate, maleimides, carbon disulfide or tetracyanoethylene, and the Hansmann group describes the reaction with a quinone methide. In all reactions the diazo functional group remains untouched and nitrogen is not released. In the reaction with organic isocyanides, a formal substitution occurs. It is only under photochemical conditions — that recently attracted the interest of organic chemists for sustainable, metal-free carbene transfer reactions[8] — that the diazo functional group is cleaved to release a free vinylidene that reacts with a pendant aromatic ring.

Both of these studies open up new exciting discoveries based on the chemistry of diazoolefins. But many questions remain regarding more generalized structural requirements that allow the synthesis of diazoolefins, coordination properties of diazoolefins in metal complexes and the reactivity of diazoolefins in vinylidene transfer reactions for applications in organic synthesis to name just a few examples. ❐

Claire Empel and Rene M. Koenigs ✉
*Institute of Organic Chemistry, RWTH Aachen University, Aachen, Germany.*
*Twitter: @EmpelClaire*
*Twitter: @ReneKoenigs*
✉*e-mail: rene.koenigs@rwth-aachen.de*

Published online: 27 October 2021
https://doi.org/10.1038/s41557-021-00811-1

References
1. Bott, K. *Chem. Ber.* **120**, 1867–1871 (1987).
2. Antoni, P. W., Golz, C., Holstein, J. J., Pantazis, D. A. & Hansmann, M. M. *Nat. Chem.* **13**, 587–593 (2021).
3. Varava, P., Dong, Z., Scopelliti, R., Fadaei-Tirani, F. & Severin, K. *Nat. Chem.* https://doi.org/10.1038/s41557-021-00790-3 (2021).
4. Roy, M. M. D. & Rivard, E. *Acc. Chem. Res.* **50**, 2017–2025 (2017).
5. Eymann, L. Y. M. et al. *J. Am. Chem. Soc.* **141**, 17112–17116 (2019).
6. Wendinger, D. & Tykwinski, R. R. *Acc. Chem. Res.* **50**, 1468–1479 (2017).
7. Klein, S., Tonner, R. & Frenking, G. *Chem. Eur. J.* **16**, 10160–10170 (2010).
8. Yang, Z., Stivanin, M. L., Jurberg, I. D. & Koenigs, R. M. *Chem. Soc. Rev.* **49**, 6833–6847 (2020).

Acknowledgements
C.E. thanks the Fonds der Chemischen Industrie for a Kekulé scholarship.

Competing interests
The authors declare no competing interests.

Check for updates

MACHINE LEARNING

# Fast track to structural biology

Machine learning algorithms are fast surpassing human abilities in multiple tasks, from image recognition to medical diagnostics. Now, machine learning algorithms have been shown to be capable of accurately predicting the folded structures of proteins.

Cecilia Clementi

In 1997, an IBM computer called Deep Blue won a game of chess against Garry Kasparov, the world chess champion at that time. Deep Blue was a specialized computer, built by IBM for this purpose and mostly relied on brute force computing power, by evaluating 200 million positions per second. In the following almost 25 years, the advent of modern machine learning has produced several milestones in competitions of the 'man against the machine' kind. In 2015 the company DeepMind reported that the program AlphaGo was able to defeat several masters at the game of Go. This fact made a news splash as, besides the intricacies of the game itself, AlphaGo is not a specialized machine but a computer program that can be run on relatively standard hardware. Since then, there has been a rapid succession of additional programs able to outperform humans in different tasks, from complex games to medical diagnostics.

It is expected that similar advances could also be made in scientific research, and that new machine-learning-based programs could upend established approaches to scientific problems. We have recently witnessed this happening in structural biology with the unveiling of the DeepMind AlphaFold2 program at the CASP14 competition[1] and later reported in *Nature*[2]. AlphaFold2 has proved capable of determining the correct fold of many proteins, "to a level of accuracy comparable to that achieved with expensive and time-consuming lab experiments" according to the organizers of the CASP14 competition[1].

Knowing a protein's structure is often important for understanding the protein's function, or at least a starting point towards it. For decades, the use of experimental techniques such as X-ray crystallography and NMR (more recently also cryo-EM) have been the most reliable methods for the determination of protein structures. The importance of these methods is

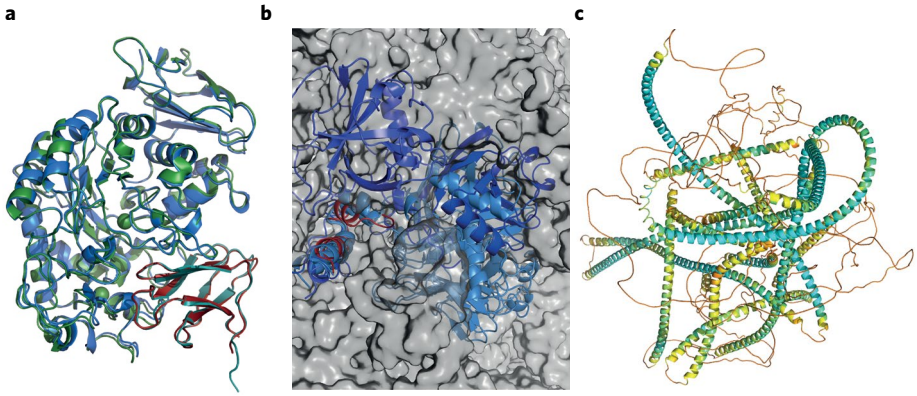


**Fig. 1 | AlphaFold2 prediction of protein structures in three different cases. a**, Prediction for the structure of the monomers in the heterodimeric protein with PDB code 1BVN. The prediction is shown in sky blue for the larger monomer and in teal for the shorter monomer and compared to the PDB structure (green and red, respectively). The agreement is very good. **b**, Prediction for the protein with PDB code 6ZMO. The AlphaFold2 top predicted structures (shown in different shades of blue) do not align in the C-terminal domain (PDB structure shown in red), which was experimentally determined only as part of a larger complex. After alignment, the predicted structures have spatial clashes with the other components in the complex (grey surface). This illustrates the limitation of only considering the monomer in isolation for prediction. **c**, One of the predicted structures for the protein with UniProt code A0A0B4K7K9 (no PDB structure available). The AlphaFold2 pretrained models cannot give consistent predictions for this target, the top scoring structures are all very different from each other and the final relaxation step cannot eliminate the spatial clashes. The color code for this structure reflects the AlphaFold2 predicted measure of error (pLDDT) and indicates that most of the structure has significant uncertainty. This protein is expected to be multimeric, partially disordered and complexed with other proteins.

underscored by the fact that several Nobel Prizes have been awarded for the successful determination of the structures of complex proteins, starting from the Chemistry Nobel Prize of 1962 to Max Perutz and John Kendrew who first demonstrated that X-ray can be used for this purpose.

AlphaFold2 is not the first software designed to predict protein structures on a computer. This field of research has a long history, and a large number of codes exist. Since 1994, the performance of the different structure prediction codes from different research groups is evaluated every two years in the Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition. Teams taking part in the CASP challenge are given the sequences of about 100 proteins with no known structure, and they have a few months to submit their predictions. In the meantime, the structures for the same set of proteins are experimentally solved and are later used to assess the predictions that were submitted by the different CASP teams. Since the first CASP edition, the ability of computer models to predict protein structures has steadily improved over the years, with some eminent successes. In CASP13 in 2018, DeepMind entered the competition for the first time with the AlphaFold code and achieved an overall result significantly better than everybody else. Still, it could not get everything right. But in the latest CASP14 competition, the newly redesigned AlphaFold2 could predict the structures of nearly all protein targets at atomic resolution, achieving an accuracy that prompted the CASP organizers to consider the protein structure prediction problem 'solved'.

The AlphaFold2 source code has been made public and discussed in a very recent paper[2], and several people in the community have started to test it, confirming its impressive performance. At the same time, the group of David Baker from the University of Washington was able to 'reverse engineer' AlphaFold2 to find its most important components (from the general idea alone, before the code was released) and extend them to improve their structure prediction code (into the new RoseTTAFold)[3], significantly boosting its performance, rendering it almost comparable to AlphaFold2 (ref. [4]).

These recent developments indicate that accurate protein structure prediction is now not only possible, but also readily available. The AlphaFold2 and RoseTTAFold source codes and associated web-servers provide powerful new tools as well as a trove of new ideas on how machine learning can be used in science, which the community can capitalize on even beyond protein structure prediction.

A few weeks ago, DeepMind publicized the AlphaFold2 predictions for the protein structures of the entire human proteome and of the proteomes of 20 additional organisms in a publicly accessible database (https://alphafold.ebi.ac.uk/)[5]. An additional and important output of AlphaFold2 is the local (residue-level) uncertainty associated with the prediction of the structure, which has been shown to be well calibrated. All the protein structures in the database come with associated uncertainties. It was also shown that predicted protein structures with large regions with high levels of uncertainty tend to be more flexible, partially disordered, or able to fold only in the presence of ligands or in complexes.

While AlphaFold2 excels in the prediction of the protein structures of single proteins, it is not particularly good at predicting the structure of protein complexes (see Fig. 1), although some 'hacks' have been reported. On the other hand, RoseTTAFold appears able to obtain reliable protein–protein complex models[4]. The prediction of the structure of proteins in complexes and/or in the presence of other molecules is presumably the next challenge in the development of these approaches, and it can have important direct implications in drug design and protein engineering applications. But how much can the methods be extended to cover more complicated (and biologically relevant) scenarios? Can changes in the structure upon changes in the physical environment (temperature, pH, and so on) also be readily predicted by similar algorithms? What about the structure of protein aggregates (for example, amyloids) upon changes in concentration? Biological function is an intrinsically dynamical concept as multiple biological processes are activated or regulated by the relative population of ensembles of protein configurations, and the changes in such populations in response to environmental changes. How a protein finds its folded structure is often very important to understand its function (or malfunction). It is not possible (yet?) to obtain this information from a AlphaFold-like software. Will it be in the (imminent) future? Maybe, it will prove much more challenging than the prediction of single protein structures.

The prediction of a protein structure as obtained by X-ray is a well-defined problem, and also has a very quantifiable measure of success. Machine learning algorithms are traditionally developed and tested on

tasks with well-defined benchmarks, and their performance is usually measured by how much they can improve the state of the art based on such benchmarks. It is then natural to use machine learning in a context like CASP, where the problem at hand is well-defined and different algorithms are compared on the same set of proteins. However, many interesting outstanding questions in science have no benchmarks. The computational study of a protein configurational landscape and its modulation in a biological environment require more complex and often indirect experimental verification. Understanding the essential physical ingredients shaping such a landscape enable researchers to manipulate it and predict changes. Much has been achieved in this respect in the last few decades with the formulation of the energy landscape theory of protein folding. Hopefully machine learning will also be able to help in the development of physics-based theories, but a qualitative leap is required towards this end. Even if AlphaFold2 and RoseTTAFold can produce a tremendous boost in the field of structural and computational biology, the field as a whole is clearly not solved by them and several additional step-changes are required to be able to fully predict (and manipulate) the function of proteins in practical applications. As in any area of science, the solution of a problem while answering some questions poses new challenges.

Good commentaries on the AlphaFold2 code itself are already available[6]. However, I think it's worth noting that one of the strengths of the approach is that it tightly integrates domain knowledge and state of the art machine learning tools into an elegant software solution. One important take-home message here is that a lot of prior knowledge informed the design of the code: the algorithm is not the prototypical black box that magically transforms sequences into structures, but it combines decades of important results. I believe this is a lesson for the application of machine learning in science in general: it does not substitute theory or knowledge but it extends them to a level that was not possible before and creates 'fast tracks' for them to run on[7]. ❐

Cecilia Clementi [1,2]✉

[1]*Department of Physics, The Free University of Berlin, Berlin, Germany.* [2]*Center for Theoretical Biological Physics, Department of Chemistry, Department of Physics, and Department of Chemical and Biomolecular Engineering, Rice University, Houston, TX, USA.*

✉*e-mail: cecilia.clementi@fu-berlin.de*

Published online: 27 October 2021
https://doi.org/10.1038/s41557-021-00814-y

## References

1. Artificial intelligence solution to a 50-year-old science challenge could 'revolutionise' medical research. *Protein Structure Prediction Center* (30 November 2020); https://predictioncenter.org/casp14/doc/CASP14_press_release.html
2. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
3. Baek, M. et al. RoseTTAFold: the first release of RoseTTAFold. *Zenodo* https://zenodo.org/record/5068265 (2021).
4. Baek, M. et al. *Science* **373**, 871–876 (2021).
5. Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021).
6. AlQuraishi, M. The AlphaFold2 method paper: A fount of good ideas. *Wordpress* https://moalquraishi.wordpress.com/2021/07/25/the-alphafold2-method-paper-a-fount-of-good-ideas/ (2021).
7. Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C. *Annu. Rev. Phys. Chem.* **71**, 361–390 (2020).

## Competing interests

The author declares no competing interests.