## Title
From systems to structure — using genetic data to model protein structures



## Journal
Nature Reviews Genetics, 23(6)

## ISSN
1471-0056

## Authors
Braberg, Hannes
Echeverria, Ignacia
Kaake, Robyn M
et al.

## Publication Date
2022-06-01

## DOI
10.1038/s41576-021-00441-w

Peer reviewed

# From systems to structure — using genetic data to model protein structures

*Hannes Braberg*[1,2], *Ignacia Echeverria*[1,2,3], *Robyn M. Kaake*[1,2,4], *Andrej Sali*[2,3,5] *and Nevan J. Krogan*[1,2,4,6]✉

Abstract | Understanding the effects of genetic variation is a fundamental problem in biology that requires methods to analyse both physical and functional consequences of sequence changes at systems-wide and mechanistic scales. To achieve a systems view, protein interaction networks map which proteins physically interact, while genetic interaction networks inform on the phenotypic consequences of perturbing these protein interactions. Until recently, understanding the molecular mechanisms that underlie these interactions often required biophysical methods to determine the structures of the proteins involved. The past decade has seen the emergence of new approaches based on coevolution, deep mutational scanning and genome-scale genetic or chemical–genetic interaction mapping that enable modelling of the structures of individual proteins or protein complexes. Here, we review the emerging use of large-scale genetic datasets and deep learning approaches to model protein structures and their interactions, and discuss the integration of structural data from different sources.

[1]*Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA, USA.*

[2]*Quantitative Biosciences Institute, University of California, San Francisco, San Francisco, CA, USA.*

[3]*Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA.*

[4]*Gladstone Institutes, San Francisco, CA, USA.*

[5]*Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, USA.*

[6]*Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.*

✉*e-mail: nevan.krogan@ucsf.edu*

https://doi.org/10.1038/s41576-021-00441-w

Deciphering the functional consequence of genetic variation within and across populations is a fundamental question of biology. To address this, a combination of techniques to interrogate changes on both systems-wide and mechanistic scales is required (FIG. 1). Systems-wide approaches provide a high-level view and generate networks that describe how different proteins or genes relate to each other or to environmental perturbations. Such networks have proved highly informative, enabling functional annotations of proteins and conveying information on the architectures of entire biological systems[1,2]. Protein–protein interaction (PPI) networks describe which proteins interact[3–5] (FIG. 1a). Experimental methods to determine PPIs include affinity purification–mass spectrometry (AP–MS)[6,7], yeast two-hybrid (Y2H) screening[8] and protein fractionation[9]. AP–MS and protein fractionation identify proteins that form complexes together in a cell type of interest, whereas Y2H uses a yeast reporter system to identify binary interactions. PPI networks describe proteins that are in physical contact but lack the resolution to discern mechanism, which often requires knowledge of the structures of the proteins and the complexes they form. Typically, high-resolution protein structures are determined using biophysical approaches, such as X-ray crystallography[10], cryogenic electron microscopy (cryo-EM)[11] and NMR spectroscopy[12] (FIG. 1b). These methods are key for elucidating protein mechanisms and designing drugs that bind to active sites or disrupt PPIs. However, traditional structural biology methods are often time-consuming and rely on purification of the relevant proteins, which is not always feasible. Furthermore, they take place in vitro, which can introduce artefacts and may not always reflect biologically relevant protein conformations.

PPI mapping and traditional structural biology are centred on proteins and their physical attributes. Genetic methods provide a functional context by means of measuring the phenotypic consequences of perturbing proteins or PPI networks. The characterization of genetic interactions[13], which describes how mutations in different genes affect one another, has proved a particularly useful complement to PPI networks. Systematic mapping of genetic interactions enables the generation of functional interaction networks, shedding light on the biological purpose of the PPIs[14,15] (FIG. 1c, left panel). Until recently, systematic genetic analyses were applied only at a whole-gene or protein level, relying on traditional structural biology for deciphering mechanistic actions. Over the past decade, developments in genetic interaction mapping and the related field of coevolution, which studies how protein residues evolve together, have allowed structural biology to be tackled on a genetic basis. By identifying pairs of residues that are related

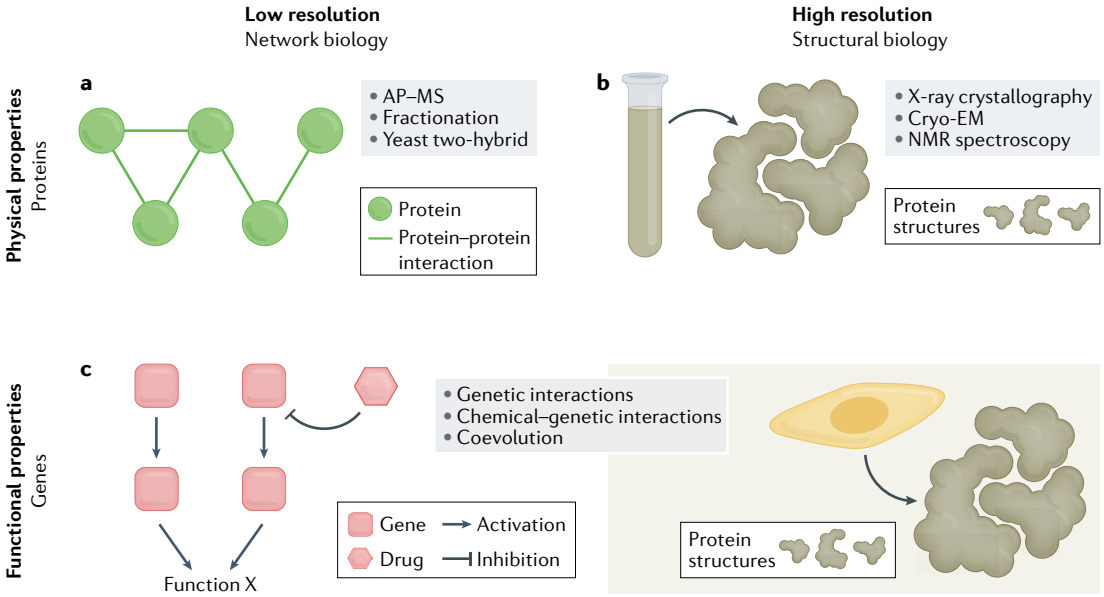



Fig. 1 | **Readouts, scale and resolution.** A complete understanding of cellular processes requires measurements of physical and functional properties at a low-resolution, systems-wide scale and at high resolution of individual components. **a** | Protein–protein interaction networks describe which proteins bind to each other and are generated using methods such as affinity purification–mass spectrometry (AP–MS), protein fractionation and yeast two-hybrid screening. **b** | High-resolution structures of proteins and their complexes are determined using biophysical methods, such as X-ray crystallography, cryogenic electron microscopy (cryo-EM) and NMR spectroscopy, that typically take place in vitro. **c** | Functional interaction networks (left panel) describe how different genes or proteins or regions thereof affect the function of each other, or how they respond to drugs. Functional connections are determined using methods such as genetic or chemical–genetic interaction mapping. Improvements in these methods and the related field of coevolution have recently enabled the structures of proteins and their complexes to be determined (right panel).

through genetic interactions or coevolution, these methods are providing high-resolution functional information sufficient to model the structures of proteins and their complexes (FIG. 1c, right panel).

In this Review, we describe the fundamentals of coevolution and genetic interaction mapping, and outline how these methods have evolved over the past decades. We discuss how technical advances and the growth of protein sequence databases have enabled the application of these methods to inform structural modelling of proteins and protein complexes. We also describe chemical–genetic interaction mapping, which is closely related to genetic interaction mapping and has similarly been used for structural modelling. We list applications of these methods and discuss emerging approaches that will enable expansion into new systems. For brevity, we do not discuss traditional structural biology methods (reviewed in[16–19]).

## Coevolution and deep learning approaches

The genetic material of all living organisms evolves over time. This evolution takes place in the form of alterations to the DNA sequence, often as single base substitutions. Coevolution analysis is based on the principle that amino acid residues in a protein, or in two interacting proteins, mutate and evolve together when they reside in the same functional region[20]. For example, in a single protein, spatially proximal amino acid residues that are essential to a specific function are likely to evolve together over time. Similarly, with two interacting proteins, if one protein evolves in the binding interface, the other protein can develop complementary changes in the interface to avoid disruption of the interaction site. This evolutionary phenomenon was observed more than three decades ago[20], and its application to predicting residue–residue contacts was made feasible a few years later with the growth of protein sequence databases and increases in computational power[21–25].

***Modelling protein structures using coevolution.*** Accurate identification of residue–residue contacts is crucial for coevolution-based protein structure modelling. Residue–residue contacts are predicted by generating a multiple sequence alignment of a protein family and identifying correlations in amino acid changes for pairs of residue positions across the alignment. Early methods used local statistical models to determine covariation between residue pairs, relying on the assumption that each correlated residue pair is independent of all other pairs[21–23,26,27]. Thus, while computationally efficient, these approaches failed to accurately represent real proteins, in which each residue can interact with many others. As a result, the local approaches were not able to distinguish direct from indirect correlations between residue pairs. Direct correlations reflect true residue–residue contacts, whereas indirect correlations arise for pairs that coevolve without being in contact. Indirect correlations can arise, for example, between residues that are evolutionarily constrained through a network path of direct contacts[28]. Accurate structure prediction requires

Multiple sequence alignment
An alignment of the sequences from multiple proteins. The multiple sequence alignment defines how the residue positions in each protein relate to those of the other proteins.

Protein family
A group of evolutionarily related proteins. The members of a protein family will typically have similar sequences and/or structures and related functions.

that only direct correlations be considered. Hence, the local statistical models were sufficient to predict contacts but lacked the resolving power necessary to model entire protein structures. During the past decade, local models have been replaced by global models, which recognize that correlated pairs are dependent on each other and furthermore incorporate the conservation of individual residues[29–33]. Global models enable the distinction of directly coupled residue pairs from those that should be excluded from the analysis because they are indirectly coupled. Crucially, these technical advancements have been accompanied by the rapid growth of protein sequence databases such as UniProt[34], increasing the coverage of sequence space across the members of protein families and making possible the systematic comparison of evolutionary changes at residue level in prokaryotes. Together, these developments paved the way for using coevolution to model the structures of monomeric proteins. The first successful determination of protein folds using coevolution was achieved by EVfold[35,36], followed by other methods, such as DCA-fold[37], FILM3 (REF.[38]) and GREMLIN[39] (FIG. 2a).

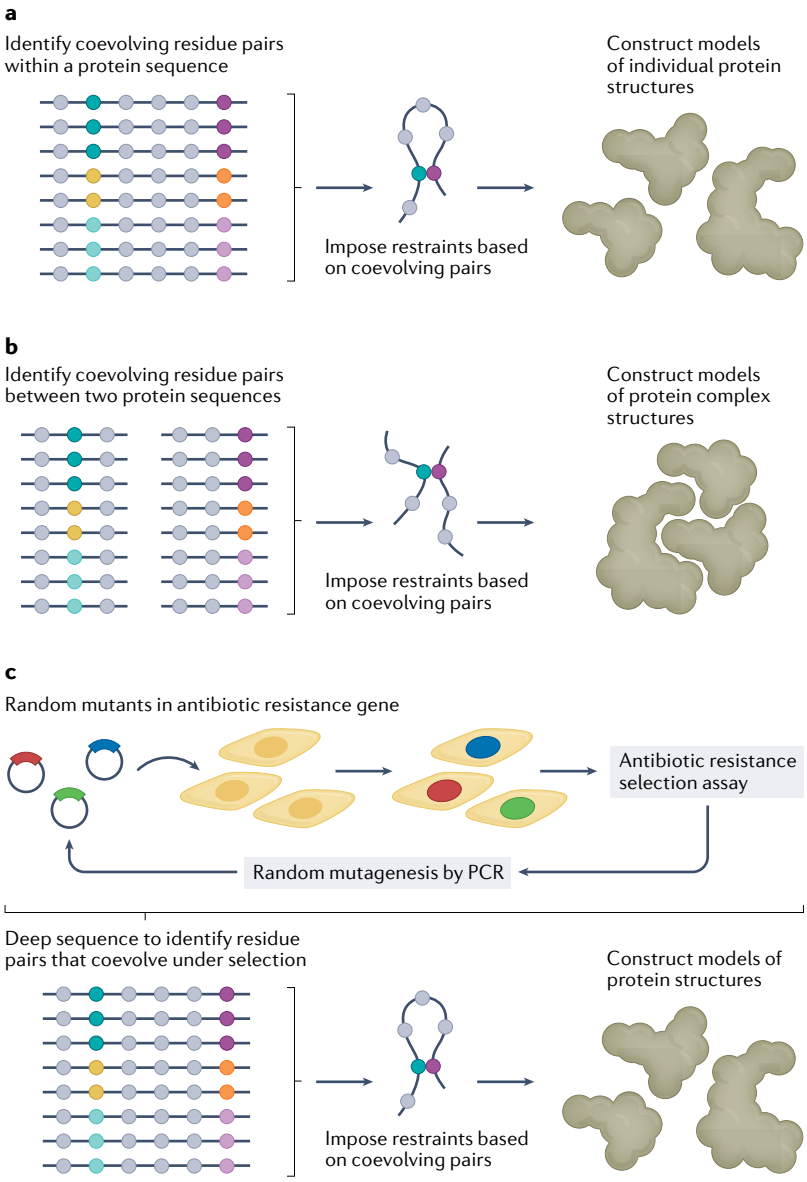



Fig. 2 | **Structural modelling of proteins and their complexes using coevolution.** Coevolution methods identify pairs of amino acid residues within or between proteins that have evolved together. Such pairs are often close in space and can be used to derive spatial restraints for structural modelling. **a** | To identify coevolving residue pairs in a protein, a multiple sequence alignment of its protein family is first generated. Pairs of sequence positions whose residue types change in a correlated fashion across the sequence alignment are coevolving and are likely to be close in space. Spatial restraints are generated based on predicted contacts and used for modelling the protein structures. **b** | Similar to part **a**, but coevolving residue pairs are here identified across the sequence alignments of an interacting pair of proteins. Here, the predicted residue contacts are thus between two different proteins, and the resulting restraints are used for modelling protein complexes instead of individual proteins. **c** | Random mutagenesis is carried out on an antibiotic resistance gene, and plasmids harbouring the gene variants are transformed into cells, followed by selection for functional copies of the gene. Surviving variants are again exposed to random mutagenesis and reintroduced into the assay. After a sufficient number of cycles, variants are deep sequenced to identify coevolving residue pairs and structural modelling is carried out as in part **a**. Filled circles represent sequence positions and the colours represent different residue types (grey denotes any residue type).

***Modelling of protein complexes and prediction of PPIs using coevolution.*** The same coevolution principles used to determine residue–residue contacts within a protein can be used to determine residue–residue contacts between proteins. However, a key challenge lies in the identification of orthologues to generate the paired multiple sequence alignments required for quantifying coevolution among residues between two proteins. Only organisms that contain both interacting proteins can be used for the multiple sequence alignments, and the interacting pairs must be correctly paired in each species, which is particularly difficult if the proteins have paralogues that perform other cellular functions[32,40–43]. To enable prediction of PPIs and modelling of their interfaces (FIG. 2b), most studies have limited their scope to protein pairs that are likely to interact based on specific criteria. For example, several efforts have focused on protein pairs encoded close to each other in conserved genomic locations (for example, on the same operon)[40,41], or pairs of protein families with members known to interact[42,44]. Although these studies demonstrated that coevolution could in principle be used for the systematic identification of PPIs, the challenges of scaling to unbiased and proteome-wide predictions made this unfeasible in practice. Furthermore, coevolution methods are computationally costly, and applying them to identify PPIs requires the combinatorial pairing of all possible interaction partners. A recent effort tackled these challenges via a combination of techniques to systematically identify PPIs in *Escherichia coli* and *Mycobacterium tuberculosis* using coevolution[45]. Hundreds of previously uncharacterized PPIs were discovered by quantifying the coevolution of residue pairs across several millions of protein pairs in both organisms. The high computational requirements were managed via a multistep protocol incorporating a faster pre-screen using local models[26], followed by global models[32,39] and structural modelling to home in on the highest confidence interactors. This study showed that coevolution is highly effective for PPI prediction in binary complexes, but less so in higher-order complexes or those that contain nucleic acids[45].

Orthologues
Evolutionarily related genes in different species. The proteins encoded by orthologous genes are typically responsible for the same function in the respective organisms.

Paralogues
Genes with similar sequences that originated via a duplication event within a genome. Paralogues belong to the same species and their encoded proteins are typically not involved in the same function.

Neural network
A category of machine learning that is inspired by the human brain and is central to deep learning algorithms.

Homology modelling
A method for determining the structure of a protein on the basis of sequence similarity with another protein of known structure by satisfying spatial restraints.

***Experimental evolution.*** Coevolution has proved powerful for determining the structures of proteins and their complexes. However, the requirement of large protein families with sufficient diversity and the obfuscating effects of paralogues impose limitations on the applicability of the approach. An experimental method (3Dseq)[46] was recently developed with the aim of using protein sequence variation generated in a laboratory to determine coevolving residues and subsequent application of computational coevolution methods for structure modelling. The approach relies on iterative generation of mutations in a given gene using error-prone PCR and exposure to a medium that selects functional variants of the gene (FIG. 2c). Selected populations are deep sequenced, and coevolving residue pairs are identified by comparison throughout the population, allowing inference of residue couplings and structural modelling using the same principles as for natural coevolution. The method was applied to two antibiotic resistance proteins from *Pseudomonas* — β-lactamase PSE1 and acetyltransferase AAC6 — expressed in *E. coli*, with functional selection by ampicillin for PSE1 and kanamycin for AAC6, resulting in accurate high-resolution models of both structures[46]. As 3Dseq does not rely on natural variation, it is particularly well suited to proteins that lack the large number of family members required for natural coevolution modelling and should provide an avenue for tackling eukaryotic systems.

***Deep learning-based approaches.*** In addition to experimental evolution, numerous computational developments have refined and extended the coevolution field. Improved statistical models[30,39,47] have increased accuracy and decreased the required number of aligned protein sequences. Incorporation of metagenome sequencing datasets has provided a means of increasing the sequence space accessed by multiple sequence alignments[48]. Finally, several new methods, such as RaptorX[49], ComplexContact[50] and DeepCov[51], use deep learning to extract and integrate additional protein sequence features with the coevolution data for contact prediction. Although these advances increased the accuracy of modelling and enabled systematic studies across prokaryotic proteomes, the technology has, in most cases, not been applied to eukaryotic proteins and complexes.

Recent advances in deep learning have led to a revolutionary development in the form of the neural network-based AlphaFold[52], which enables regular prediction of protein structures at near experimental accuracy, in prokaryotes as well as eukaryotes. The AlphaFold (version 2) engine makes use of constraints on protein structure derived from evolution, physics and geometry. During training, AlphaFold parses experimental protein structures deposited in the protein databank (PDB)[53], as well as clustered protein sequence databases, such as BFD[52] and UniRef90 (REF.[54]), learning rules to govern the modelling of structure from sequence. The neural network takes as input a multiple sequence alignment of a given protein and its family members to extract evolutionary information for individual residues as well as on a pairwise basis. Incorporation with components learnt from the PDB enables the final structure prediction[52].

AlphaFold has proved remarkably effective for determining the structures of individual proteins and their complexes. The AlphaFold model, trained on single protein chains, was showcased on nearly the entire human proteome, resulting in confident structure predictions for 58% of all residues[55]. In comparison, experimental efforts over the past several decades have together resulted in structural coverage of 17% of human protein residues[55]. Similarly, a study across 11 different proteomes found that AlphaFold added structure determination for on average 25 percentage points of additional residues over existing experimental structures or those that could be derived by homology modelling[56]. Interestingly, despite being trained on single proteins, AlphaFold proved capable of modelling the structures of protein complexes[56–58]. Most recently, AlphaFold-Multimer has been released, featuring a model trained on multimeric protein structures, which clearly outperforms the standard AlphaFold for modelling protein complex structures[59].

Inspired by the performance of AlphaFold, the RoseTTAFold[60] software was developed using similar ideas. The accuracy of RoseTTAFold is generally somewhat lower than that of AlphaFold, but the predictions are faster and require less computational power[60]. RoseTTAFold provided early evidence that this technology can model protein complexes in addition to individual proteins[60]. Recently, the respective strengths of RoseTTAFold and AlphaFold were combined to not only model but also identify protein complexes[61]. The high speed of RoseTTAFold was leveraged to examine more than 4 million paired multiple sequence alignments to generate a set of approximately 5,500 potential PPIs in *Saccharomyces cerevisiae* (budding yeast). AlphaFold was then applied to this smaller set to identify higher-confidence candidate protein complexes and model their structures[61]. Importantly, like all technologies discussed in this Review, these methods rely on data generated from experimental approaches and should be viewed as powerful complements to these[62], rather than as replacements.

## Genetic and chemical–genetic interactions

A complementary approach to coevolution and deep learning-based methods leverages the measurement of genetic interactions, providing a means for structural modelling using sets of intentionally designed mutations.

For most organisms, such as *Homo sapiens*, budding yeast or *E. coli*, any given gene is typically directly functionally related to only a small number of other genes. Thus, when deleting or otherwise perturbing two different genes, the cellular response will most often reflect the combined effect of the two as independent contributions. Genetic interactions arise between genes for which the response deviates from this expectation, indicating that the genes are functionally related. Genetic interactions can be measured by multiple phenotypic readouts, but often centre around cell replication and survival as this can be informative for most systems, including unicellular organisms and human cancer cells. Positive genetic interactions arise when the cell is either no sicker (epistatic) or healthier (buffering) than

Subunits
Single proteins in the context of a protein complex.

the sickest single mutant. This may indicate factors that operate in the same pathway or are subunits of the same non-essential complex[63]. Conversely, negative genetic interactions (synthetic sick or lethal) occur when mutations in two genes lead to a more severe growth defect than expected. This may reflect factors that function in parallel pathways or are non-essential subunits of the same essential protein complex (FIG. 3a).

Chemical–genetic interactions, similar to genetic interactions, describe how the presence or absence of a drug or environmental perturbation affects the phenotype of a single genetic mutation. Here, a positive interaction reflects that drug treatment has a lesser effect on the mutant phenotype than expected, which could indicate that the drug inhibits pathways in which the mutated gene functions. By contrast, negative chemical–genetic interactions arise when the effect of a mutation in the presence of a drug is more severe than expected, potentially indicating that the drug inhibits a parallel pathway (FIG. 3b). Notably, the relationships that form the basis of genetic and chemical–genetic interactions are often more complex than the illustrative examples provided here.

***Systematic analysis of genetic and chemical–genetic interactions.*** Early work on concepts that underlie genetic interactions focused on small numbers of genes that were already known to affect a given phenotype of interest[13]. In the early 2000s, the creation of gene deletion libraries in budding yeast and advances in high-throughput technologies paved the way for systematic mapping of genetic and chemical–genetic interactions[64]. A key development was introduced by synthetic genetic array (SGA), which enabled the rapid crossing of a set of test mutants across a deletion library in a plate-based format, providing an efficient means of identifying synthetic lethal interactions[15]. A different method, diploid-based synthetic lethal analysis with microarrays (dSLAM), relied on barcoded yeast mutants grown in a pooled competitive format, where microarrays were used to quantify the amounts of the different single and double mutants[65]. These methods were primarily developed to identify negative genetic interactions. The ability to capture positive genetic interactions was introduced by epistatic miniarray profile (E-MAP), which expanded on SGA to provide quantitative measurements of the entire spectrum of genetic interactions in a high-throughput format[66,67]. This approach enables the generation of a continuous genetic interaction profile for each test mutant, consisting of its scores across all deletion library mutants; these profiles can be used to group together proteins that are functionally related or belong to the same complex[14,67–70] (FIG. 3c). In parallel with these developments, related methods were designed for determining chemical–genetic interactions, following a similar format but using a library of chemical perturbations in place of the deletion library[71,72] (FIG. 3c). Chemical–genetic interaction mapping relies on methods similar to those of genetic interaction mapping but is considerably less complex, as it simply relies on the addition of drugs to the plates or pools of single mutants[65,71–74].

Systematic genetic and chemical–genetic interaction mapping (for example, chemical–genetic miniarray profile (CG-MAP)) have proved highly effective for organizing genes on the basis of function on both local and global levels[14,67–71,74–76]. The technologies have been adapted to different model systems, including *Caenorhabditis elegans*[77], *E. coli*[75,76], *Schizosaccharomyces pombe*[78]

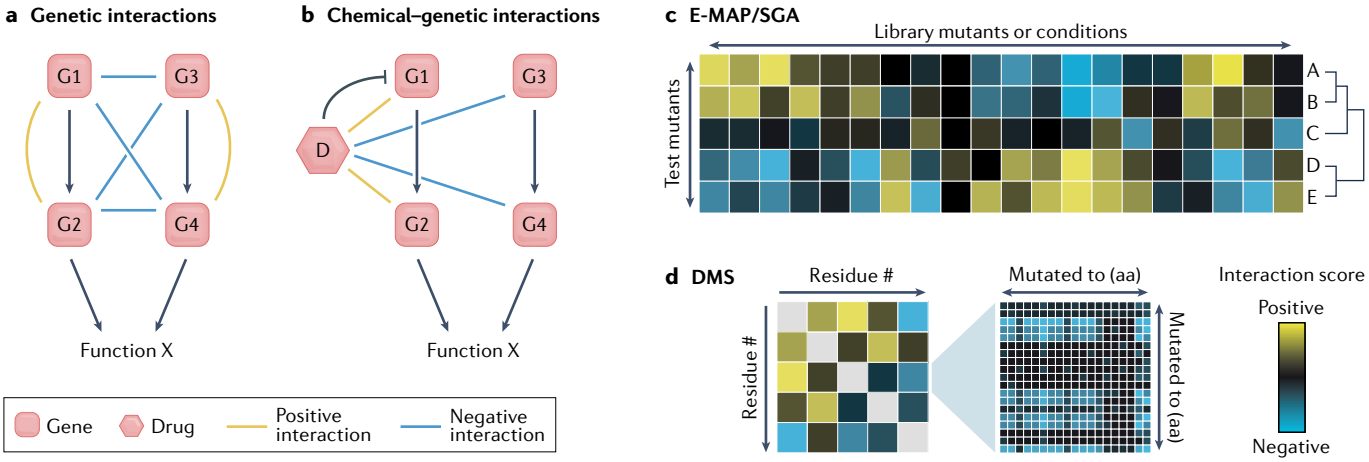



Fig. 3 | **Mapping of genetic and chemical–genetic interactions.** Genetic and chemical–genetic interactions describe the functional relationships between pairs of mutations or between a mutation and a drug, respectively. **a** | A positive genetic interaction between two gene deletions may indicate that the gene products operate in the same pathway (G1–G2 or G3–G4), whereas a negative interaction can arise if the products of the deleted genes belong to parallel pathways (for example, G1–G3). **b** | Positive interactions between a drug (D) and a gene deletion can indicate an antagonistic relationship (for example, D–G1), whereas a negative interaction may indicate that the gene product belongs to a parallel pathway of the drug target (for example, D–G3). **c** | The epistatic miniarray profile (E-MAP) and synthetic genetic array (SGA) approaches allow for high-throughput measurements of genetic or chemical–genetic interactions between a set of test mutants (*y*-axis) and a genome-scale library (*x*-axis). Each row constitutes the genetic interaction profile for a test mutant (A–E), and clustering these by similarity (tree on right) provides a functional organization of the mutants. **d** | Deep mutational scanning (DMS) can be used to measure genetic interactions between all pairwise combinations of point mutations in a gene. For each pair of residue positions (left), all possible combinations of amino acids (aa) are measured (right), which can be used to generate a composite genetic interaction score for the position pair. Depictions in parts **c**,**d** are illustrative subsets of much larger interaction maps.

Knockdowns
Genes whose expression has been reduced.

Complex haploinsufficiencies
Negative genetic interactions observed in cells that are hemizygous for two different genes. The phenotype of the two hemizygous loci combined is more severe than expected if the genes were unrelated.

Hemizygous
A diploid cell is hemizygous for a gene if it harbours only one functional allele of the gene.

Allostery
A process whereby an active site in a protein (enzyme) is regulated by the binding of a molecule to a different site (typically distal in space).

and *Drosophila melanogaster* cell lines[79]. More recently, advances in RNA interference (RNAi) and CRISPR–Cas9 (REF.[80]) genome editing have enabled expansion into mammalian cells[81–85].

***Genetic interactions of point mutants.*** Most genetic interaction maps have focused on whole-gene deletions or knockdowns. However, early studies in budding yeast investigated the genetic interaction profiles for limited numbers of point mutants. For example, alanine scan mutations of the actin gene *ACT1* were screened for genetic interactions with more than 200 genes that had been shown to exhibit complex haploinsufficiencies in a strain hemizygous for *ACT1* (REF.[86]). The screen revealed that alanine mutations in close proximity on the actin surface shared many interactions (that is, exhibited similar genetic interaction profiles), suggesting that they may be disrupting the same PPI binding interfaces[86]. Similarly, an early budding yeast E-MAP that focused on chromatin biology included three alleles of the *POL30* gene[14], which encodes the multifunctional protein PCNA that functions in DNA replication and repair and in chromatin assembly. The *pol30-79* point mutant allele gave rise to a genetic interaction profile similar to that of *pol30-DAmP* (a gene knockdown allele), suggesting a destabilizing effect on the protein. The genetic interaction profiles of these mutants were consistent with a defective DNA replication and repair system[14,63,87]. By contrast, the *pol30-8* allele, which perturbs a different region of PCNA, exhibited genetic interactions relating to defects in chromatin assembly. Interestingly, this allele has been shown to diminish the PPI between PCNA and chromatin assembly factor 1 (CAF1)[88]. These results indicated that genetic interactions provide a high level of resolution and allow the dissection of multifunctional proteins into regions that are functionally and physically connected to other factors. Spurred by these findings, the E-MAP technology was extended to screen entire libraries of point mutations in a set of related proteins to generate point mutant E-MAPs (pE-MAPs)[89,90]. Quantitative SGA screens have also included large numbers of point mutations; however, these have generally been chosen on the basis of their phenotype as temperature-sensitive alleles of essential genes, rather than systematic mutations of a specific protein or complex[68,69].

Concurrently with pE-MAP, a complementary approach termed deep mutational scanning (DMS) was developed[91]. DMS set out to tackle the problem of identifying the most informative mutations to study in a protein, without the requirement of preselecting residues of interest. To this end, the method allows for a comprehensive screen of point mutations in a protein or protein domain. DMS relies on the rapid synthesis of large numbers of mutations in a gene, in conjunction with a genotype–phenotype coupled selection assay. In its most basic form, DMS quantifies the effects of individual point mutations on a specific function, via the chosen selection assay. However, it can also be applied to pairs of point mutations to quantify genetic interactions[91] (FIG. 3d).

The development of pE-MAP and DMS enabled the systematic study of the relationship between genetic interactions and residue distances in a protein structure. The first pE-MAP covered 53 budding yeast point mutants in RNA polymerase II (RNAPII), crossed against a library of 1,200 deletion and knockdown mutants[89]. This study revealed that pairs of residues that exhibited similar genetic interaction profiles were typically close in space, whether they resided in the same or different RNAPII subunits[89,90]. Several early DMS studies revealed similar patterns for the pairwise genetic interactions between point mutants[92–94]. For example, a screen of double mutants of 75 residues in the RRM2 domain of the budding yeast PAB1 protein showed that both positive and negative genetic interactions were enriched at shorter distances between the mutated residues[92]. These findings were supported in a screen of genetic interactions for all pairs of mutations in 55 residues of the IgG binding domain of streptococcal protein G (GB1)[93]. In some proteins, such as those regulated by allostery, these trends can differ. For example, a recent pE-MAP screen of the molecular switch Gsp1/Ran revealed that the genetic interaction profiles of interface mutations reflected their biophysical effects on the switch cycle kinetics, instead of their interface locations[95]. These studies highlight how genetic interactions ultimately report on mechanism and showcase the complementarity of this technology to traditional structural biology approaches.

***Modelling the structures of proteins and their complexes using genetic and chemical–genetic interactions.*** Similar to coevolution, genetic interaction data have been used for structural modelling of proteins and their complexes. The key challenge remains how to derive spatial restraints between pairs of residues that can be used for modelling. pE-MAP and DMS provide complementary strengths for this purpose. For example, DMS can provide comprehensive genetic interaction measurements of all possible residue–residue combinations in a protein. Indeed, these fine-grained data can be used to model the secondary structure and tertiary structure of small proteins or domains[96–98] (FIG. 4a,b). Two groups[96,97] examined genetic interaction data from DMS scans of GB1 (REF.[93]), the RRM2 domain of the budding yeast PAB1 protein[92], the human YAP65 WW domain[99] and the heterodimer FOS–JUN[100]. The authors set out to use the genetic interaction data from each of these studies to predict structural contacts between residue pairs in the respective protein domains and to test whether the contacts could be used for structure determination[96,97]. The GB1 dataset was the most comprehensive and covered nearly all possible mutation pairs across 55 residues, which allowed the determination of residue contacts and accurate modelling of both secondary and tertiary structure of the domain[96,97]. The RRM2 and WW domain datasets covered only a fraction of the possible double mutants and were sequenced less deeply. Although contact prediction was possible with these datasets, the secondary structure predictions were not accurate. The fold of a 22–24 residue section of the WW domain could be modelled; however, the RRM2 domain fold could not[96,97]. The data for the FOS–JUN dimer covered a stretch of 32 residues on each monomer and enabled contact predictions across

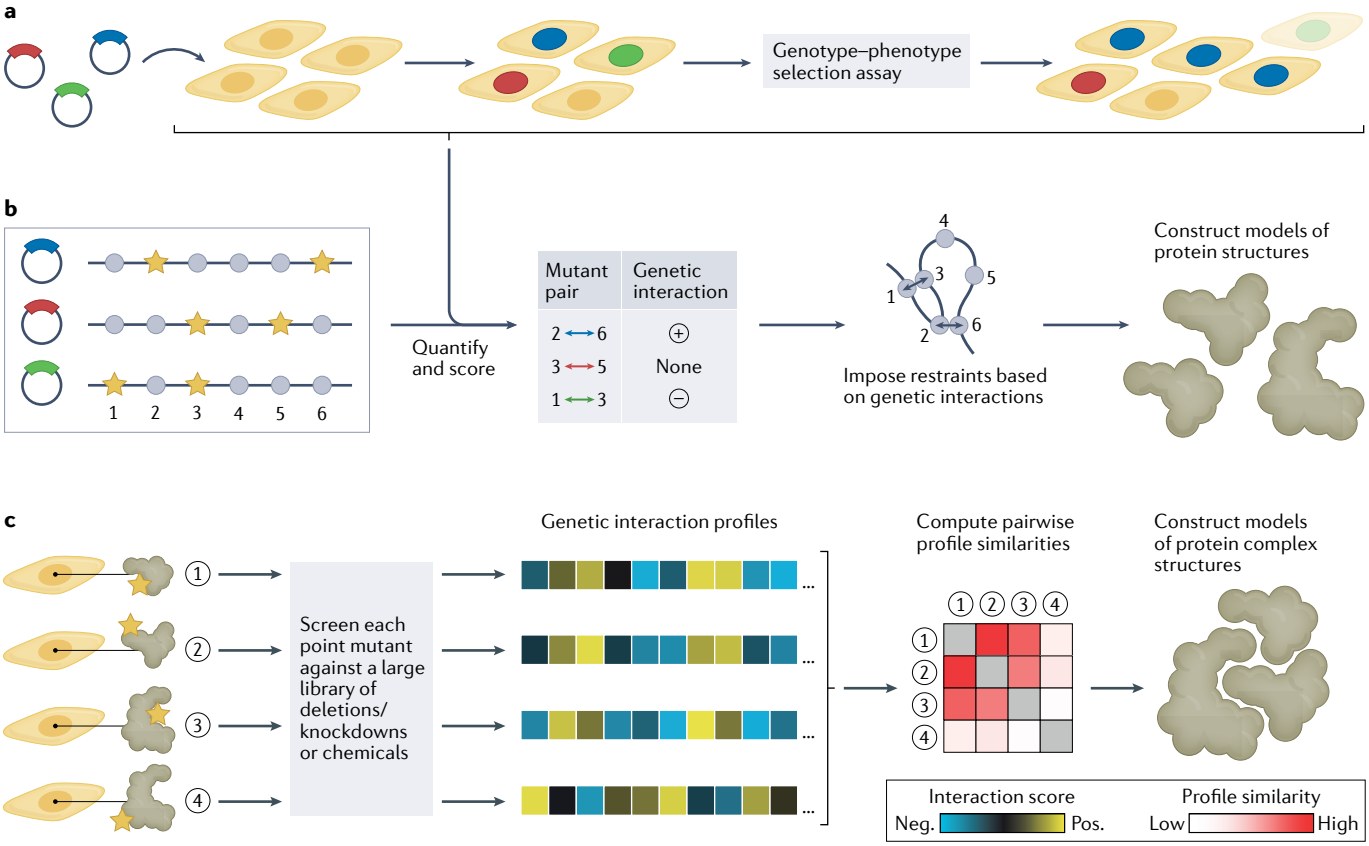



Fig. 4 | **Structural modelling of proteins and their complexes using genetic and chemical–genetic interactions. a** | Deep mutational scanning (DMS) relies on the rapid synthesis of mutated variants (blue, red or green) of a gene, which are cloned into vectors and introduced into an assay (here, cell-based) that competitively selects for variants with particular traits. The composition of variants is determined via deep sequencing before and after selection, allowing for identification of variants that are enriched or depleted by the selection. **b** | When using DMS to measure genetic interactions, each gene variant contains two point mutations (stars). The selection assay identifies mutant pairs that are enriched (positive genetic interaction) or depleted (negative genetic interaction) compared with an expectation from the quantities of each single mutant. Likely residue contacts are identified based on the genetic interactions and used for modelling the structure of the protein. **c** | The point mutant epistatic miniarray profile (pE-MAP) approach relies on in vivo screening of a set of point mutants in two or more interacting proteins against a large library of gene deletions and/or knockdowns (pE-MAP) or chemicals (chemical–genetic miniarray profile (CG-MAP)). The resulting genetic (or chemical–genetic) interaction profiles often consist of more than 1,000 genetic interactions for each point mutant. Pairwise comparison of the profiles provides measures of genetic similarity between all pairs of tested point mutants. High similarity between a pair of point mutants indicates a likely contact between the mutated residues. The structure of the protein complex is modelled using this relationship for pairs of residues that reside in different subunits of the complex.

the interface[96,97]. The predicted contacts were then incorporated into a protein docking of the two monomers as spatial restraints, greatly improving the accuracy of the models compared with docking without DMS-derived restraints[96]. Finally, one of the studies also predicted contacts in an RNA molecule[96,101], the twister ribozyme from *Oryza sativa*, suggesting that DMS could be used for RNA structure prediction. Interestingly, although the two studies[96,97] harnessed different ranges of the genetic interaction data and used different interaction metrics for computing contact predictions, they nonetheless arrived at similar results. This suggests that the approach is robust and highlights the massive information content of DMS data. Accordingly, both groups showed that sparser data subsets still allowed modelling of the GB1 structure at an accuracy similar to that achieved when using the complete dataset. These findings highlight the potential of DMS as a structural biology tool, and other studies have further applied it to successfully reveal structural features of intrinsically disordered proteins[102,103].

Whereas DMS is well suited for modelling the structures of small proteins and domains, the pE-MAP approach is more appropriate for determining structures of protein assemblies. pE-MAP has lower coverage than DMS but enables comparison of genetic interactions across residues in any number of interacting proteins in a single screen, which facilitates the modelling of interactions. Additionally, pE-MAP provides systems-wide cellular information for every mutated residue via its genetic interaction profile with thousands of other mutants in different pathways and processes. A recent study harnessed these traits to use pE-MAP and chemical–genetic interaction data to determine the structures of protein complexes[104] (FIG. 4c). Using a technique termed integrative structure determination[105] (BOX 1), the authors modelled the structures of three protein complexes: histones H3 and H4 in budding yeast; subunits Rpb1 and Rpb2 of RNAPII in budding yeast, and subunits RpoB and RpoC of bacterial RNA polymerase (RNAP) in *E. coli*. The histone pE-MAP included

Box 1 | **Integrative structure determination**

Integrative structure determination is a powerful tool to determine the structures of macromolecular assemblies[105,131] by providing a framework to combine information from varied experimental approaches, bioinformatics tools and prior knowledge. Integrative modelling aims to maximize the completeness, accuracy and precision of the resulting model by computing an ensemble of structural models that are consistent with all the input information. The integrative modelling approach has been successful in determining the architecture of large macromolecular assemblies[132,133], describing the structural heterogeneity of flexible protein complexes[134,135] and rationalizing the effect of pathogenic mutations[132,136]. The integrative modelling workflow iterates through the following four stages (see the figure).

### Gathering information

A large variety of experimental and computational information can be used for integrative modelling including X-ray crystallography, NMR spectroscopy, electron microscopy, chemical cross-linking mass spectrometry, small-angle scattering and affinity purification–mass spectrometry. Evolutionary residue–residue couplings computed from natural variation[40,41,137] or from experimental evolution[46] can also be used for modelling and are often complementary to experimental methods. Recently it has also been demonstrated that genetic interactions measured using the point mutant epistatic miniarray (pE-MAP) platform[104] and deep mutational scanning[96,97,102,103,138] (DMS) can be used for integrative modelling of small proteins and protein complexes.

### Representing the system and translating information into spatial restraints

A structural model of a macromolecular assembly is defined by the conformations and relative positions and orientations of its components (for example, atoms, residues, domains and subunits). Thus, the representation is defined by all the structural variables that need to be determined on the basis of input information. This includes, for example, the components of the system (including the copy number), the coordinates of the components and whether multiple states need to be modelled. The scoring function consists of a series of terms that encode the spatial restraints that quantify the degree of a match between the structural models and the input information. For example, pE-MAP data were converted into a Bayesian data likelihood that provides an upper bound on the distance spanned by the mutated residues and objectively interprets the noise in the experimental data[104]. Similarly, data from DMS experiments and coevolution analysis are converted into upper-bound or harmonic distance restraints between the residues[40,41,96,97,102,139]. The scoring function also accounts for the physicochemical properties of proteins via terms such as excluded volume and sequence connectivity[140].

### Structural sampling

Structural models are computed by sampling the conformations and/or the configuration of the components; this is often achieved by using Monte Carlo-based methods for stochastic sampling. The result is an ensemble (that is, the model) of predicted structures that agree with the input information within acceptable tolerances.

### Validating the model

Validation of the model is essential to quantify its uncertainty and to assess the degree of consistency between the model and the information used and not used to compute it[141,142]. To this end, the validation protocol includes five steps whose outputs are an estimate of the model precision (quantified by the variability between the models in the ensemble), one or more representative structures and their uncertainties, and mapping of the known information into the structures in the model.

This protocol (that is, stages 2, 3 and 4) can be scripted using the open-source Integrative Modelling Platform (IMP) package[143].

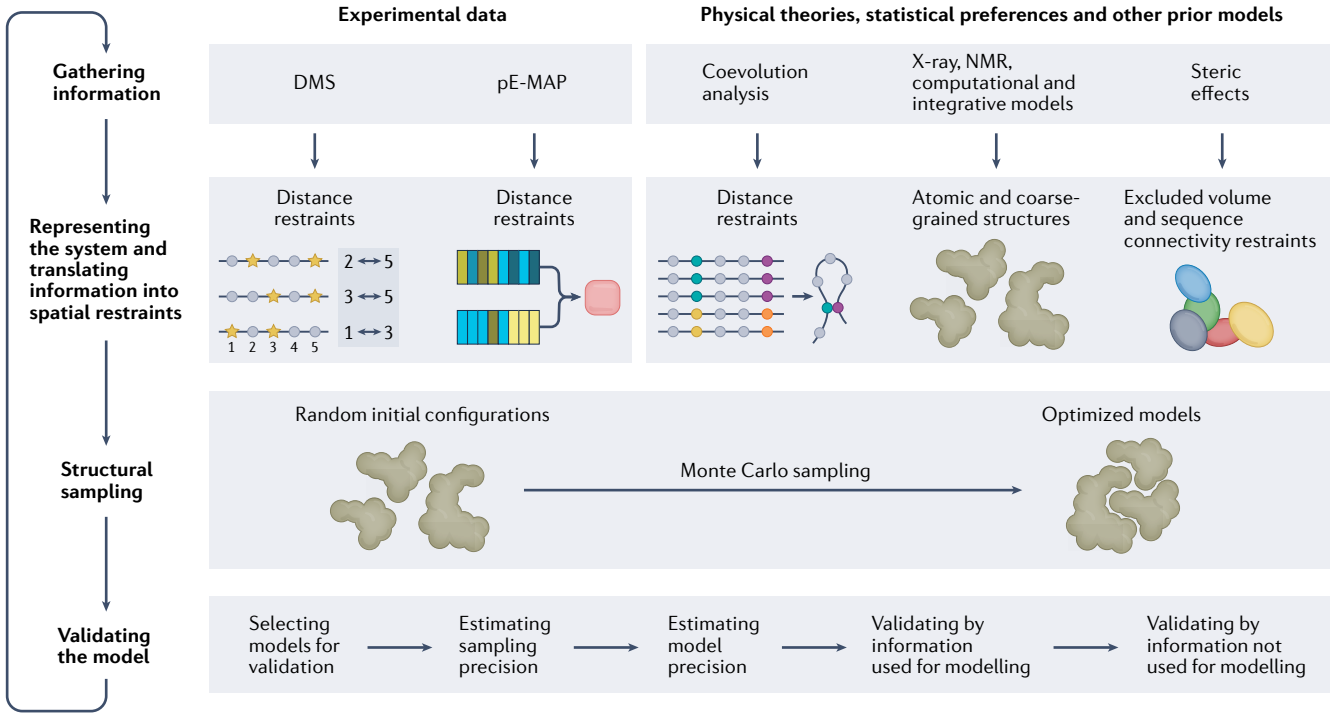



Figure adapted with permission from REF.[104], AAAS.

a comprehensive alanine scan as well as context-specific mutations, resulting in a map of 350 histone mutants crossed against 1,370 deletion or knockdown mutants[104]. Distance restraints between H3–H4 residue pairs were devised using the similarity of genetic interaction profiles between the corresponding mutations. These restraints were then applied to arrange the structures of the H3 and H4 subunits, capturing the interface of

their interaction and obtaining an accurate structure of the H3–H4 complex. The RNAPII dataset provided an opportunity to test the performance of the approach on a system that differs vastly from that of the histones. Specifically, Rpb1 and Rpb2 are much larger than the histones (1,200–1,700 residues versus 100–140 residues) and the RNAPII pE-MAP is much sparser, with 53 point mutants crossed against 1,200 deletion or knockdown mutants[89]. In addition, the authors split Rpb1 into two domains for the structural modelling to test the applicability to a higher-order system. The model of this three-body complex proved accurate, suggesting that the approach is generalizable and can effectively harness the contents of sparse datasets. Extending the use of the approach to chemical–genetic interactions, the authors accurately modelled the RpoB–RpoC complex of bacterial RNAP using a CG-MAP of 44 point mutants subjected to 83 different environmental stresses[106]. This showed transferability of the approach to chemical–genetic interaction maps in spite of the reduced size of the interaction profiles in this dataset. Finally, in a comparison of integrative structure determination using cross-linking mass spectrometry (XL-MS) data and pE-MAP data, the authors found that the two performed similarly, but crucially led to higher accuracy models when combined[104]. Thus, a key value of the methods described in this Review is that their data types are typically orthogonal to those traditionally used in structural biology, allowing data integration that results in improved models[105] (BOX 1).

## Emerging approaches

A key promise and challenge for the methods discussed in this Review is the expansion into new systems, scales and organisms. The continued success of this field will rely on the effective integration of complementary data types to best make use of available methods (FIG. 1). In particular, the integration of experimental data with those from computational coevolution and deep learning models should prove valuable. Such efforts will likely benefit from a fine-grained interpretation of the scale and resolution represented by each data type. For example, it has been shown that residue–residue contacts derived from coevolution are more accurate when compared with experimentally determined side chain contacts than with more commonly used backbone contacts[107]. This finding suggests that the dominant effect observed in coevolution reflects side chain interactions, and could be harnessed to generate more precise models when computationally feasible.

To better complement computational methods, there is a need to increase the speed and coverage of experimental genetic approaches. Advances in CRISPR–Cas9 genome editing (BOX 2) are setting the stage for such developments. For example, chemical–genetic interaction mapping is primed for modelling PPIs on a proteome-wide scale in yeast, using a recent method to efficiently generate point mutations while surveying their drug sensitivities in a multiplexed fashion[108] (BOX 2). Guided by global PPI maps[109], and using individual protein structures from traditional structural biology methods or AlphaFold/RoseTTAFold, this system should in principle enable the modelling of interaction interface structures across the yeast proteome. In addition to facilitating increased scale, CRISPR–Cas9 genome editing can be used for the systematic generation of point mutations in mammalian cells[110–114]. At present, these approaches are not suitable for mammalian pE-MAP screening, owing to incomplete editing, off-target effects or other technical obstacles (BOX 2). However, these limitations are steadily diminishing[110], setting the stage for genetics-based structural modelling of protein complexes in human cells and providing a means of characterizing the effects of disease-causing mutations. By integration with recent efforts to generate multi-scale models of entire cells[115–119], genetic interaction mapping could thus inform on global function as well as the structures of protein complexes.

One of the most crucial, and currently tractable, applications to human systems relates to the rapidly growing field of host–pathogen interaction mapping[120–124]. This area of research is centred on the systematic identification of PPIs between pathogen and host proteins and the generation of interaction networks between the two organisms (FIG. 5a). These networks have proved highly effective for interrogating the mechanisms of infection, revealing important aspects of pathogen life cycles, host factor functions and host–pathogen interplay, as well as providing potential targets for drug discovery[120–124]. Host–pathogen PPI networks could be used as a blueprint for genetic interaction mapping between pathogen point mutants and human gene knockouts or knockdowns. To generate these maps, human cells would be infected by virus harbouring the relevant point mutations, and the human proteins from the PPI maps would be knocked down or knocked out (FIG. 5b), allowing for the construction of a host–pathogen genetic interaction map (FIG. 5c). The genetic interaction profiles of the viral point mutants would then be converted into spatial restraints for structural modelling of viral protein complexes (FIG. 5d), which would ultimately be re-integrated into the PPI map. The platforms required for such efforts have recently been developed. For example, a technology for generating viral E-MAPs (vE-MAPs), using infectivity as readout, was recently applied to HIV infection in human cells[125]. In an analogous fashion, DMS could be used for modelling individual viral proteins, by employing suitable selection assays[126]. For example, a DMS platform was developed to structurally map mutations in the SARS-CoV-2 Spike receptor-binding domain that alter ACE2 binding or escape antibody recognition[127,128]. Many pathogens adapt rapidly to circumvent immune and drug responses[128–130]. Genetic interaction-driven modelling of pathogen protein structures will provide an avenue to identify the mechanisms of these changes, laying the groundwork for therapeutic intervention.

## Conclusions

Structural modelling of proteins and protein complexes using genetically derived restraints lies at the intersection of network biology and structural biology. Until recently, these major areas of research were disparate and had little overlap. Network biology provided a large-scale systems view of interactions within and between cellular

Knockouts
Genes that have been inactivated (for example, deleted).

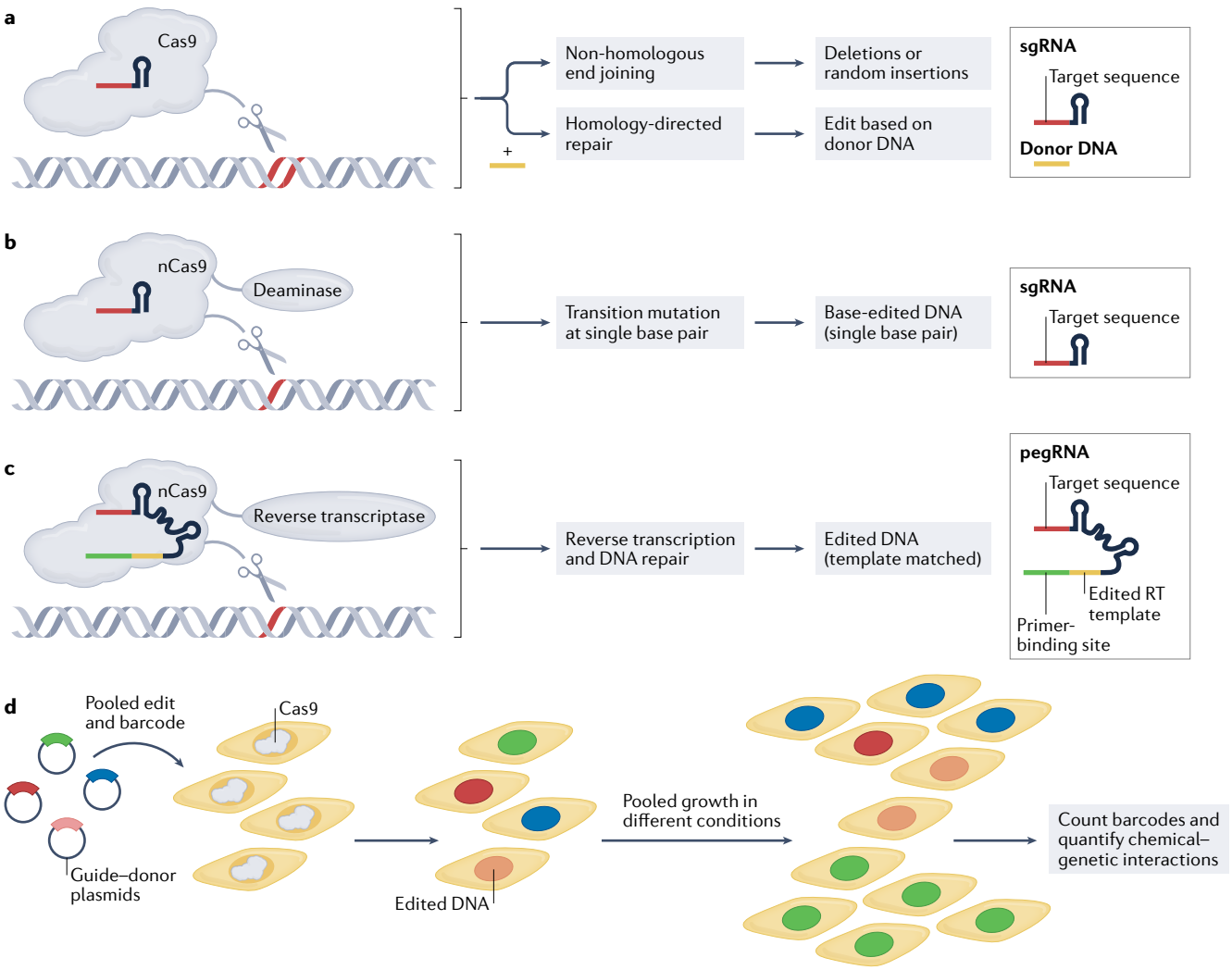
Box 2 | **CRISPR–Cas9 applications at residue-level resolution**

The CRISPR–Cas9 system sets up for genome editing by introducing a double-stranded break (DSB) in DNA (see the figure, panel **a**)[80]. The Cas9 enzyme is directed to the target DNA site by a single guide RNA (sgRNA), which contains the target sequence. Cas9 cuts the DNA at the target site, and the break is typically repaired via non-homologous end joining (NHEJ), resulting in insertions and deletions (indels) that lead to inactivation of the target gene. Alternatively, the DSB can be repaired via homology-directed repair (HDR), resulting in a specific edit based on the template of a stretch of donor DNA. However, HDR in mammalian cells is inefficient, and the natural preference of the cell for NHEJ would lead to the introduction of unwanted indels even in the presence of donor DNA.

Base editors offer a more fine-tuned alternative, by relying on catalytically impaired versions of Cas9 that do not introduce DSBs. Most base editors consist of a DNA deaminase enzyme fused to either nickase Cas9 (nCas9), which cuts a single strand of double-stranded DNA, or to catalytically dead Cas9 (dCas9). Base editors convert specific base pairs (as directed by the sgRNA) into different base pairs (see the figure, panel **b**). Base editing circumvents the need for donor DNA and avoids unintentional indels at target or off-target sites. However, the technique does not support all 12 possible DNA base-to-base conversions and suffers from other limitations, including unwanted bystander or off-target edits and sequence-specific requirements to allow for editing (for example, proximity of a protospacer adjacent motif (PAM) site)[112].

A recent development, termed prime editing, provides a flexible platform for DNA editing, allowing for all base-to-base conversions, insertions or deletions, without the need of a DSB or donor DNA, and with lower off-target activity than Cas9 (see the figure, panel **c**)[110]. The prime editor consists of nCas9 fused to a reverse transcriptase, which is guided to its target by a prime editing guide RNA (pegRNA). In addition to the target sequence, the pegRNA contains a reverse transcriptase template (RT template) for the desired edit, preceded by a primer-binding site. The primer-binding site hybridizes to the nicked target DNA, and the RT template dictates the sequence of the new edited DNA. Prime editing and base editing methods could both potentially be used for genetic interaction mapping in mammalian cells, but the editing efficiency is not yet high enough for robust application[112].

In budding yeast, which is more tractable for genome editing than mammalian cells, a CRISPR–Cas9-based method was recently developed for multiplexed genome editing in a pooled fashion, allowing for the rapid measurement of point mutant chemical–genetic interactions (see the figure, panel **d**)[108]. Here, guide–donor plasmids are first generated, which contain the desired sequence of donor DNA, combined with a barcode and guide sequences to direct the edit and barcode integration. The plasmids are transformed into Cas9-expressing yeast cells, resulting in genomically edited cells with the corresponding barcode integrated. Cells are grown in a pooled format and exposed to a large number of different conditions. Barcodes are counted via sequencing, and chemical–genetic interactions are quantified based on enrichment or depletion of each mutant in treated versus untreated conditions. This method would allow for proteome-wide measurement of chemical–genetic interactions for protein complex subunits, thereby providing the data required for global structural modelling of the budding yeast protein interactome.



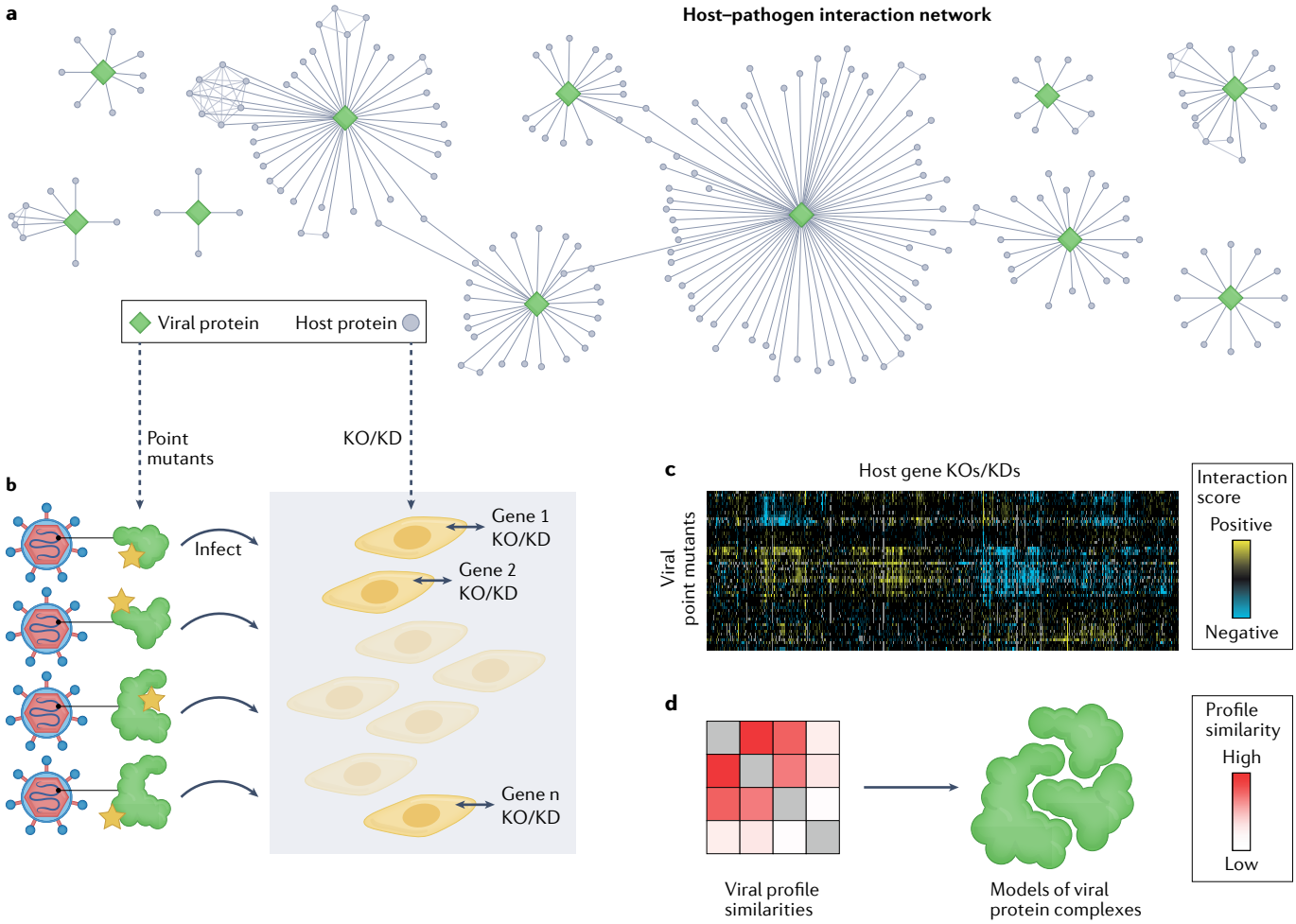



Fig. 5 | **Structural characterization of host–pathogen interaction networks. a** | A host–pathogen protein–protein interaction (PPI) network generated using affinity purification–mass spectrometry. The edges denote PPIs between pairs of proteins. **b** | To generate a host–pathogen point mutant epistatic miniarray profile (pE-MAP), host cells are infected with point mutant virus strains, in combination with CRISPR–Cas9 knockout (KO) or knockdown (KD) of the host genes identified in the host–pathogen PPI network (part **a**). **c** | The resulting pE-MAP comprises genetic interaction profiles for the viral point mutants, containing their genetic interactions with the library of host gene KOs and KDs. **d** | Viral genetic interaction profiles are compared across the subunits of viral protein complexes and the similarities are used for modelling their structures, which can then be integrated into the original network.

processes, whereas structural biology supplied structures of individual proteins and complexes, typically derived in vitro. Genetics-based structural modelling uses spatial restraints derived from functional data, such as coevolution or genetic interactions, to compute structural models. The methods are efficient and low cost, and enable structural characterization of protein interaction interfaces, with a potential to cover entire protein–protein interactomes, including those of host–pathogen systems. These techniques are not meant to replace traditional structural biology methods, which remain the gold standard in terms of resolution. Instead, the orthogonal datasets produced by genetics-based modelling are primed to complement traditional structural biology methods to provide a more accurate and complete description of the structures of proteins in vivo.

Published online 10 January 2022


1. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007).
2. Barabasi, A. L. Scale-free networks: a decade and beyond. *Science* **325**, 412–413 (2009).
3. Swaney, D. L. et al. A protein network map of head and neck cancer reveals PIK3CA mutant drug sensitivity. *Science* **374**, eabf2911 (2021).
4. Kim, M. et al. A protein interaction landscape of breast cancer. *Science* **374**, eabf3066 (2021).
5. Zheng, F. et al. Interpretation of cancer mutations using a multiscale map of protein systems. *Science* **374**, eabf3067 (2021).
6. Krogan, N. J. et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637–643 (2006).
7. Gavin, A. C. et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
8. Yu, H. et al. High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
9. Havugimana, P. C. et al. A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).
10. Shi, Y. A glimpse of structural biology through X-ray crystallography. *Cell* **159**, 995–1014 (2014).
11. Henderson, R. Realizing the potential of electron cryo-microscopy. *Q. Rev. Biophys.* **37**, 3–13 (2004).
12. Wuthrich, K. The way to NMR structures of proteins. *Nat. Struct. Biol.* **8**, 923–925 (2001).
13. Phillips, P. C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867 (2008).
14. Collins, S. R. et al. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806–810 (2007).
15. Tong, A. H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).

16. Dobson, C. M. Biophysical techniques in structural biology. *Annu. Rev. Biochem.* **88**, 25–33 (2019).
17. Murata, K. & Wolf, M. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochim. Biophys. Acta Gen. Subj.* **1862**, 324–334 (2018).
18. Huang, C. & Kalodimos, C. G. Structures of large protein complexes determined by nuclear magnetic resonance spectroscopy. *Annu. Rev. Biophys.* **46**, 317–336 (2017).
19. Wall, M. E., Wolff, A. M. & Fraser, J. S. Bringing diffuse X-ray scattering into focus. *Curr. Opin. Struct. Biol.* **50**, 109–116 (2018).
20. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193**, 693–707 (1987).
21. Gobel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309–317 (1994).
22. Neher, E. How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA* **91**, 98–102 (1994).
23. Taylor, W. R. & Hatrick, K. Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341–348 (1994).
24. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7**, 349–358 (1994).
25. Thomas, D. J., Casari, G. & Sander, C. The prediction of protein contacts from multiple sequence alignments. *Protein Eng.* **9**, 941–948 (1996).
26. Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340 (2008).
27. Fodor, A. A. & Aldrich, R. W. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56**, 211–221 (2004).
28. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–1080 (2012).
29. Thomas, J., Ramakrishnan, N. & Bailey-Kellogg, C. Graphical models of residue coupling in protein families. *IEEE/ACM Trans. Comput. Biol. Bioinform* **5**, 183–197 (2008).
30. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
31. Burger, L. & van Nimwegen, E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.* **6**, e1000633 (2010).
32. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72 (2009).
33. Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).
34. UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
35. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
**This study describes the first application of protein structure modelling using spatial restraints derived from coevolution data**.
36. Hopf, T. A. et al. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
37. Sulkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *Proc. Natl Acad. Sci. USA* **109**, 10340–10345 (2012).
38. Nugent, T. & Jones, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl Acad. Sci. USA* **109**, E1540–E1547 (2012).
39. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl Acad. Sci. USA* **110**, 15674–15679 (2013).
40. Hopf, T. A. et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430 (2014).
41. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).
42. Bitbol, A. F., Dwyer, R. S., Colwell, L. J. & Wingreen, N. S. Inferring interaction partners from protein sequences. *Proc. Natl Acad. Sci. USA* **113**, 12180–12185 (2016).
43. Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511–523 (1997).
44. Baldassi, C. et al. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS ONE* **9**, e92721 (2014).
45. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019).
**This study represents a major expansion of the utility of coevolution by applying it to predict PPIs on a proteome-wide scale in *E. coli* and *M. tuberculosis***.
46. Stiffler, M. A. et al. Protein structure from experimental evolution. *Cell Syst.* **10**, 15–24 e15 (2020).
47. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* **87**, 012707 (2013).
48. Ovchinnikov, S. et al. Protein structure determination using metagenome sequence data. *Science* **355**, 294–298 (2017).
49. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **13**, e1005324 (2017).
50. Zeng, H. et al. ComplexContact: a web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res.* **46**, W432–W437 (2018).
51. Jones, D. T. & Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* **34**, 3308–3315 (2018).
52. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
**This deep learning approach allows for efficient prediction of protein structures at near experimental accuracy**.
53. Burley, S. K. et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **49**, D437–D451 (2021).
54. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
55. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
56. Akdel, M. et al. A structural biology community assessment of AlphaFold 2 applications. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.26.461876 (2021).
57. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2 and extended multiple-sequence alignments. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.15.460468 (2021).
58. Ghani, U. et al. Improved docking of protein models by a combination of Alphafold2 and ClusPro. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.07.459290 (2021).
59. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.04.463034 (2021).
60. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
**This deep learning approach allows for efficient prediction of protein structures at near experimental accuracy**.
61. Humphreys, I. R. et al. Computed structures of core eukaryotic protein complexes. *Science* https://doi.org/10.1126/science.abm4805 (2021).
62. Gupta, M. et al. CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.10.443524 (2021).
63. Beltrao, P., Cagney, G. & Krogan, N. J. Quantitative genetic interactions reveal biological modularity. *Cell* **141**, 739–745 (2010).
64. Boone, C., Bussey, H. & Andrews, B. J. Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet.* **8**, 437–449 (2007).
65. Pan, X. et al. A robust toolkit for functional profiling of the yeast genome. *Mol. Cell* **16**, 487–496 (2004).
66. Collins, S. R., Schuldiner, M., Krogan, N. J. & Weissman, J. S. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol.* **7**, R63 (2006).
67. Schuldiner, M., Collins, S. R., Weissman, J. S. & Krogan, N. J. Quantitative genetic analysis in *Saccharomyces cerevisiae* using epistatic miniarray profiles (E-MAPs) and its application to chromatin functions. *Methods* **40**, 344–352 (2006).
68. Costanzo, M. et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).
69. Costanzo, M. et al. The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
70. Fiedler, D. et al. Functional organization of the *S. cerevisiae* phosphorylation network. *Cell* **136**, 952–963 (2009).
71. Kapitzky, L. et al. Cross-species chemogenomic profiling reveals evolutionarily conserved drug mode of action. *Mol. Syst. Biol.* **6**, 451 (2010).
72. Nichols, R. J. et al. Phenotypic landscape of a bacterial cell. *Cell* **144**, 143–156 (2011).
73. Chang, M., Bellaoui, M., Boone, C. & Brown, G. W. A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage. *Proc. Natl Acad. Sci. USA* **99**, 16934–16939 (2002).
74. Hillenmeyer, M. E. et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362–365 (2008).
75. Butland, G. et al. eSGA: *E. coli* synthetic genetic array analysis. *Nat. Methods* **5**, 789–795 (2008).
76. Typas, A. et al. High-throughput, quantitative analyses of genetic interactions in *E. coli*. *Nat. Methods* **5**, 781–787 (2008).
77. Lehner, B., Crombie, C., Tischler, J., Fortunato, A. & Fraser, A. G. Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat. Genet.* **38**, 896–903 (2006).
78. Roguev, A. et al. Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* **322**, 405–410 (2008).
79. Horn, T. et al. Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nat. Methods* **8**, 341–346 (2011).
80. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
81. Du, D. et al. Genetic interaction mapping in mammalian cells using CRISPR interference. *Nat. Methods* **14**, 577–580 (2017).
82. Shen, J. P. et al. Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods* **14**, 573–576 (2017).
83. Roguev, A. et al. Quantitative genetic-interaction mapping in mammalian cells. *Nat. Methods* **10**, 432–437 (2013).
84. Laufer, C., Fischer, B., Billmann, M., Huber, W. & Boutros, M. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nat. Methods* **10**, 427–431 (2013).
85. Bassik, M. C. et al. A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. *Cell* **152**, 909–922 (2013).
86. Haarer, B., Viggiano, S., Hibbs, M. A., Troyanskaya, O. G. & Amberg, D. C. Modeling complex genetic interactions in a simple eukaryotic genome: actin displays a rich spectrum of complex haploinsufficiencies. *Genes Dev.* **21**, 148–159 (2007).
87. Ryan, C. J. et al. High-resolution network biology: connecting sequence with function. *Nat. Rev. Genet.* **14**, 865–879 (2013).
88. Zhang, Z., Shibahara, K. & Stillman, B. PCNA connects DNA replication to epigenetic inheritance in yeast. *Nature* **408**, 221–225 (2000).
89. Braberg, H. et al. From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. *Cell* **154**, 775–788 (2013).
90. Braberg, H., Moehle, E. A., Shales, M., Guthrie, C. & Krogan, N. J. Genetic interaction analysis of point mutations enables interrogation of gene function at a residue-level resolution: exploring the applications

of high-resolution genetic interaction mapping of point mutations. *Bioessays* **36**, 706–713 (2014).
91. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
92. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* **19**, 1537–1551 (2013).
93. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).
94. Sahoo, A., Khare, S., Devanarayanan, S., Jain, P. C. & Varadarajan, R. Residue proximity information and protein model discrimination using saturation-suppressor mutagenesis. *eLife* **4**, e09532 (2015).
95. Perica, T. et al. Systems-level effects of allosteric perturbations to a model molecular switch. *Nature* **599**, 152–157 (2021).
96. Rollins, N. J. et al. Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176 (2019).
**This study describes the use of deep mutational scanning to generate restraints for determining the structures of small proteins or domains**.
97. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).
**This study describes the use of deep mutational scanning to generate restraints for determining the structures of small proteins or domains**.
98. Eccleston, R. C., Pollock, D. D. & Goldstein, R. A. Selection for cooperativity causes epistasis predominately between native contacts and enables epistasis-based structure reconstruction. *Proc. Natl Acad. Sci. USA* **118**, e2010057 (2021).
99. Araya, C. L. et al. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl Acad. Sci. USA* **109**, 16858–16863 (2012).
100. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, e32472 (2018).
101. Kobori, S. & Yokobayashi, Y. High-throughput mutational analysis of a twister ribozyme. *Angew. Chem. Int. Ed. Engl.* **55**, 10354–10357 (2016).
102. Newberry, R. W., Leong, J. T., Chow, E. D., Kampmann, M. & DeGrado, W. F. Deep mutational scanning reveals the structural basis for alpha-synuclein activity. *Nat. Chem. Biol.* **16**, 653–659 (2020).
103. Bolognesi, B. et al. The mutational landscape of a prion-like domain. *Nat. Commun.* **10**, 4162 (2019).
104. Braberg, H. et al. Genetic interaction mapping informs integrative structure determination of protein complexes. *Science* **370**, eaaz4910 (2020).
**This study describes the modelling of protein complex structures, using restraints derived from genome-scale genetic interaction data and chemical–genetic interaction data**.
105. Rout, M. P. & Sali, A. Principles for integrative structural biology studies. *Cell* **177**, 1384–1403 (2019).
**This publication describes integrative structural biology, which serves as a crucial tool for integrating different types of dataset for the structural modelling of protein complexes**.
106. Shiver, A. L. et al. Chemical-genetic interrogation of RNA polymerase mutants reveals structure-function relationships and physiological tradeoffs. *Mol. Cell* **81**, 2201–2215 e2209 (2021).
107. Hockenberry, A. J. & Wilke, C. O. Evolutionary couplings detect side-chain interactions. *PeerJ* **7**, e7280 (2019).
108. Roy, K. R. et al. Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nat. Biotechnol.* **36**, 512–520 (2018).
109. Collins, S. R. et al. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell Proteom.* **6**, 439–450 (2007).
110. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
**This CRISPR–Cas9-based genome editing approach allows for all base-to-base conversions, insertions or deletions, without the need of a double-stranded break or donor DNA, and with lower off-target activity than Cas9 nuclease**.
111. Ma, L. et al. CRISPR-Cas9-mediated saturated mutagenesis screen predicts clinical drug resistance with improved accuracy. *Proc. Natl Acad. Sci. USA* **114**, 11751–11756 (2017).
112. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
113. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
114. Erwood, S. et al. Saturation variant interpretation using CRISPR prime editing. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.11.443710 (2021).
115. McGuffee, S. R. & Elcock, A. H. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.* **6**, e1000694 (2010).
116. Singla, J. et al. Opportunities and challenges in building a spatiotemporal multi-scale model of the human pancreatic β cell. *Cell* **173**, 11–19 (2018).
117. Takamori, S. et al. Molecular anatomy of a trafficking organelle. *Cell* **127**, 831–846 (2006).
118. Thul, P. J. et al. A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).
119. Wilhelm, B. G. et al. Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins. *Science* **344**, 1023–1028 (2014).
120. Eckhardt, M., Hultquist, J. F., Kaake, R. M., Huttenhain, R. & Krogan, N. J. A systems approach to infectious disease. *Nat. Rev. Genet.* **21**, 339–354 (2020).
121. Gordon, D. E. et al. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* **370**, eabe9403 (2020).
122. Gordon, D. E. et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
123. Ramage, H. R. et al. A combined proteomics/genomics approach links hepatitis C virus infection with nonsense-mediated mRNA decay. *Mol. Cell* **57**, 329–340 (2015).
124. Jager, S. et al. Global landscape of HIV-human protein complexes. *Nature* **481**, 365–370 (2011).
125. Gordon, D. E. et al. A quantitative genetic interaction map of HIV infection. *Mol. Cell* **78**, 197–209.e197 (2020).
126. Tenthorey, J. L., Young, C., Sodeinde, A., Emerman, M. & Malik, H. S. Mutational resilience of antiviral restriction favors primate TRIM5alpha in host-virus evolutionary arms races. *eLife* **9**, e59988 (2020).
127. Starr, T. N. et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310 e1220 (2020).
128. Greaney, A. J. et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* **29**, 44–57 e49 (2021).
129. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2**, e00631 (2013).
130. Wong, A. H. M. et al. Receptor-binding loops in alphacoronavirus adaptation and evolution. *Nat. Commun.* **8**, 1735 (2017).
131. Sali, A. From integrative structural biology to cell biology. *J. Biol. Chem.* **296**, 100743 (2021).
132. Kim, S. J. et al. Integrative structure and functional anatomy of a nuclear pore complex. *Nature* **555**, 475–482 (2018).
133. Lasker, K. et al. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc. Natl Acad. Sci. USA* **109**, 1380–1387 (2012).
134. Gutierrez, C. et al. Structural dynamics of the human COP9 signalosome revealed by cross-linking mass spectrometry and integrative modeling. *Proc. Natl Acad. Sci. USA* **117**, 4088–4098 (2020).
135. Kwon, Y. et al. Structural basis of CD4 downregulation by HIV-1 Nef. *Nat. Struct. Mol. Biol.* **27**, 822–828 (2020).
136. Luo, J. et al. Architecture of the human and yeast general transcription and DNA repair factor TFIIH. *Mol. Cell* **59**, 794–806 (2015).
137. Wang, S., Li, W., Liu, S. & Xu, J. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res.* **44**, W430–W435 (2016).
138. Fernandez-de-Cossio-Diaz, J., Uguzzoni, G. & Pagnani, A. Unsupervised inference of protein fitness landscape from deep mutational scan. *Mol. Biol. Evol.* **38**, 318–328 (2021).
139. Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A. & Bonvin, A. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins* **86** (Suppl. 1), 51–66 (2018).
140. Viswanath, S. & Sali, A. Optimizing model representation for integrative structure determination of macromolecular assemblies. *Proc. Natl Acad. Sci. USA* **116**, 540–545 (2019).
141. Saltzberg, D. J. et al. Using Integrative Modeling Platform to compute, validate, and archive a model of a protein complex structure. *Protein Sci.* **30**, 250–261 (2021).
142. Viswanath, S., Chemmama, I. E., Cimermancic, P. & Sali, A. Assessing exhaustiveness of stochastic sampling for integrative modeling of macromolecular structures. *Biophys. J.* **113**, 2344–2353 (2017).
143. Russel, D. et al. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10**, e1001244 (2012).

**Acknowledgements**
The authors thank P. Beltrao and R. B. Babu for helpful discussion and comments on the manuscript. This research was funded by grants from the National Institutes of Health (NIH) (U54CA209891, U54NS100717, 1U01MH115747, U19 AI135990, U19AI135972, and P50AI150476 to N.J.K; R01GM083960 and P41GM109824 to A.S.). This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreements HR00111920020 and HR00112020029 to N.J.K. The views, opinions and/or findings contained in this material are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the US Government.

**Author contributions**
The authors contributed equally to all aspects of the article.

**Competing interests**
The Krogan Laboratory has received research support from Vir Biotechnology and F. Hoffmann-La Roche. N.J.K. has consulting agreements with the Icahn School of Medicine at Mount Sinai, New York, Maze Therapeutics and Interline Therapeutics. N.J.K. is a shareholder in Tenaya Therapeutics, Maze Therapeutics and Interline Therapeutics, and a financially compensated Scientific Advisory Board Member for GEn1E Lifesciences, Inc. The other authors declare no competing interests.

**Peer review information**
*Nature Reviews Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Publisher's note**
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.